ENCYCLOPEDIA OF
MODERN OPTICS
SECOND EDITION

EDITORS IN CHIEF

Bob D. Guenther

Duke University, Durham, NC, United States

Duncan G. Steel

University of Michigan, Ann Arbor, MI, United States

VOLUME 4

Communications ■ Optical Modulation ■ Holography
■ Optical Processing ■ Transformation Optics ■ Optical Interconnects
■ Optics of Semiconductors and 2D Materials ■ Nonlinear Optics ■ Laser Radar

ELSEVIER

Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

ISBN 978-0-12-809283-5

Publisher: Oliver Walter
Acquisition Editor: Ruth Ireland
Content Project Manager: Sean Simms
Associate Content Project Manager: Marise Willis
Designer: Greg Harris

Printed and bound in the United Kingdom

EDITORIAL BOARD

CONTENTS OF VOLUME 4

VOLUME 4

EDITORS IN CHIEF

Bob D. Guenther received his undergraduate degree from Baylor University and his graduate degrees in Physics from University of Missouri. He has had research experience in condensed matter and optical physics. For 9 years he was active in research management as a Senior Executive in the Army, responsible for the physics research sponsored by the Army. After retiring from the government he held the position of Interim Director of the Free Electron Laser Laboratory and helped establish Duke's Fitzpatrick Center for Photonics and Communication Systems and served as Executive Director of the Center until his retirement. In a continuation of the retirement process, he moved to Applied Quantum Technology and has just retired from that company. He is author of the textbook, *Modern Optics*, second edition. He is now composing an elementary book in optics.

Duncan G. Steel is the Robert J. Hiller Professor of Electrical and Computer Engineering, as well as Professor of Physics and Biophysics. Prior to joining the faculty of the University of Michigan in 1985, he was a senior research scientist at Hughes Aircraft Company at the Hughes Research Laboratories. At the University of Michigan, he was Area Chair and Director of the Optical Sciences Laboratory for 20 years until 2007 when he took over the position of Chair of the Biophysics Research Division during its transition to an academic unit. As an educator and teacher, he has chaired or cochaired over 60 doctoral committees. His research includes coherent optical studies of semiconductors and their application to quantum information. His work also included 30 years of studies on age-related modifications in proteins where he exploited numerous optical techniques including single molecule studies in neurons in his studies on Alzheimer's Disease. He was a Guggenheim Scholar and received the 2010 Isakson Prize from the American Physical Society.

INTRODUCTION

This is the second, updated edition of the *Encyclopedia of Modern Optics*. There are 197 entries, many of them new or updated, reflecting the enormous progress in the optical sciences and technology and the ever-expanding impact since the publication of the first edition. Some of the new topics are:

- Nano-photonics and Plasmonics
- Quantum Optics
- Quantum Information
- Optical Interconnects
- Photonic Crystals and Their Applications
- High Efficiency LED's
- Displays
- Transformation Optics
- Fiber Lasers
- Terahertz
- Multidimensional Spectroscopy
- Organic Optoelectronics
- Gravitational Wave Detectors
- Meta Materials and Plasmonics

Selection of article topics and recruiting authors for those topics in this edition has been the work of the topical editors listed in the prologue. They were able to solicit contributions from internationally recognized leaders in their field.

The entries of the encyclopedia are arranged by subject as best as possible, a task made difficult because the field is now highly interdisciplinary and there are many subjects that have an impact in many different areas. We have added references to other articles so that readers can obtain a deeper understanding of the material or understand how a specific discussion on basic science may impact an application or technology.

Basic Concepts of Optical Communication Systems

S Lee and AE Willner, University of Southern California, Los Angeles, CA, USA

Introduction

Since modern lightwave communication started in the 1970s, extensive efforts have been made to increase data rate and transmission distance. Recently, the bit rates for single-channel and multichannel systems have exceeded 40 and 2 Tb/s on single fibers respectively, for transoceanic distances. One revolutionary development in lightwave systems has been the deployment of erbium doped fiber amplifiers (EDFAs). These amplifiers facilitate transmission of an optical signal over long distances, by providing periodic analog-like amplification rather than digital-like regeneration. The wide bandwidth provided by EDFAs has made possible the increased use of wavelength division multiplexing (WDM) in which multiple lightwave signals, each having a different wavelength, are co-propagated on a single fiber. EDFAs and WDM techniques can enhance the lightwave system capacity, both in terms of obtainable transmission distance, and total number of data rates (see **Fig. 1**).

Optical Fiber

Most telecommunication fibers are made of silica, therefore following the attenuation of silica material. **Fig. 2** shows the silica fiber attenuation as a function of wavelength. The downward slope towards 1.6 μm and upward slope away from 1.6 μm, are the theoretical limits due to Rayleigh scattering and absorption of silica material, respectively. The absorption peak near 1.4 μm is due to water molecule absorption. There are two low attenuation regions: ~1.3 μm and 1.55 μm, both of which are used for communications in general, and some other wavelengths are also used for shorter distance communications. **Fig. 2** also shows the comparison between gain bandwidth of EDFA (~3 THz) and the low attenuation window around 1.55 μm (~25 THz). The typical values of attenuation at 1.3 μm and 1.55 μm for single-mode fiber (SMF) are ~0.35 dB/km, and 0.2 dB/km, respectively.

When an electromagnetic wave interacts with the bound electrons of a silica fiber, the medium response depends upon optical frequency (ω). This property, referred to as material dispersion, manifests itself through the frequency dependence of the refractive

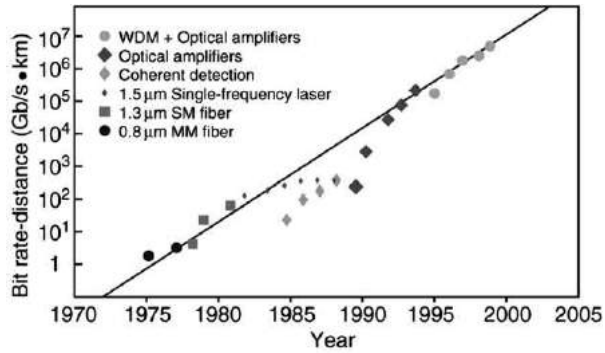


Fig. 1 Bit rate product. Data from T. Li (private communication).

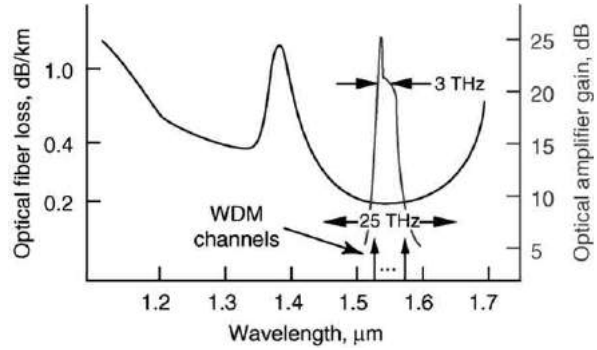


Fig. 2 Silica fiber attenuation versus wavelength.

index $n(\omega)$. Fiber material dispersion plays a critical role in propagation of short optical pulses since different spectral components associated with the pulse travel at different speeds, given by $c/n(\omega)$. Consequently, the optical pulse at the output of the fiber is broadened in time. The commonly used system parameter is called the dispersion parameter D (the negative D value is called normal dispersion, and the positive D value is called anomalous dispersion), and is measured in ps/(nm · km); D for SMF is $\sim +17$ ps/(nm · km) at 1.5 μm. However, because of dielectric waveguiding, the effective mode index is slightly lower than the material index $n(\omega)$, with the reduction itself being ω-dependent (waveguide dispersion). This results in a waveguide contribution that must be added to the material contribution to obtain the total dispersion (see **Fig. 3**).

Generally, the waveguide contribution to dispersion D is to shift zero dispersion wavelength (λ_0) to longer wavelengths. Furthermore, the waveguiding can be tailored to shift the λ_0 from 1.3 μm to 1.5 μm, the wavelength at which the fiber has the minimum attenuation loss. The fiber having λ_0 in the neighborhood of 1.5 μm, is called dispersion-shifted fiber (DSF); D for DSF is usually between -2.5 and $+2.5$ ps/nm · km, at a wavelength of 1.5 μm. The typical dispersion parameter, D, as a function of a wavelength for both SMF and DSF, is shown in **Fig. 4**.

In addition to chromatic dispersion, optical fiber has another dispersion properties related to polarization. An optical wave of arbitrary polarization can be represented as the superposition of two orthogonally polarized modes. In an ideal fiber, these two modes are indistinguishable, and have the same propagation constants owing to the cylindrical symmetry of the waveguide. However, in real fibers there is some residual anisotropy due to unintentional circular asymmetry, usually caused by noncircular waveguide geometry or asymmetrical stress around the core, as shown in **Fig. 5**.

In either case, the loss of circular symmetry gives rise to two distinct orthogonally-polarized modes with different propagation constants (differential phase velocity) responsible for polarization mode dispersion (PMD) in the fiber, and can be related to the difference in refractive indices (birefringence) between the two orthogonal polarization axes. The differential phase velocity indicated is accompanied by a difference in the group velocities for the two polarization modes, therefore the pulse at the end of

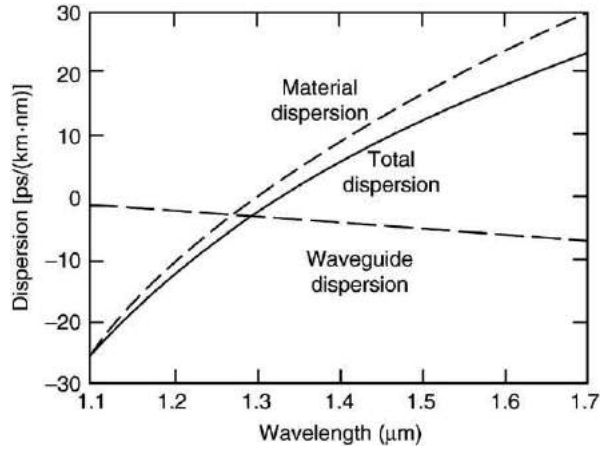


Fig. 3 Total dispersion (D) and contribution of material dispersion and waveguide dispersion in conventional SMF.

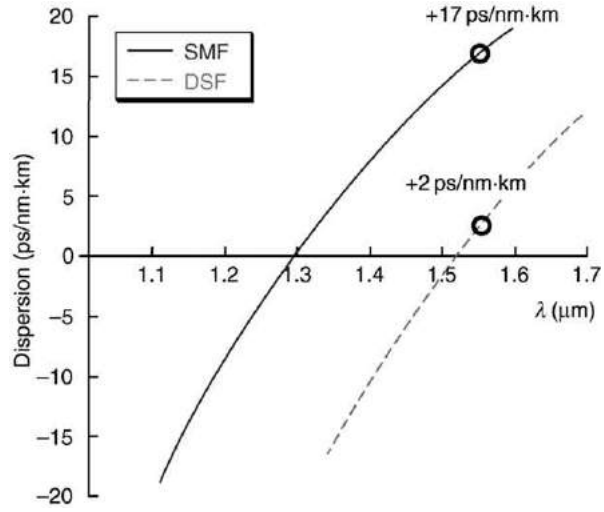


Fig. 4 Dispersion versus wavelength for conventional SMF and DSF.

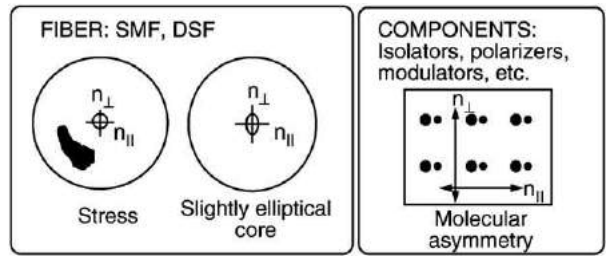


Fig. 5 Origin of PMD.

the transmission fiber is broadened by this differential group delay (DGD). DGD is usually expressed in units of ps/km for a short length (1 m to 1 km) of birefringent fiber. However, DGD does not accumulate along a long fiber link in a linear fashion. Instead, because of random variations in the perturbations along a fiber span, the DGD in one section may either add to or subtract from another section of the fiber. As a result, average DGD in long fiber spans accumulates in a random-walk-like process that leads to a square root of transmission-length dependence. Therefore, average DGD is expressed in $ps/km^{1/2}$ in long fiber spans, referred to as the PMD of the fiber, and the typical PMD parameter has a value of 0.01 to 10 $ps/km^{1/2}$. Because of the many perturbations that act on a real-world fiber (e.g., temperature, vibration, etc.), transmission properties typically vary with time. Therefore, the PMD of the fiber link fluctuates randomly, thus causing random fluctuations in system performance.

It is well known that optical fibers show nonlinear behavior under conditions of high power and long interaction length. The power-times-distance products for amplified transmission systems can be large enough to make fiber nonlinear effects the dominant factor in the design of long-distance systems.

There are two categories of fundamental optical nonlinear effects: stimulated scattering effects and refractive index effects. Stimulated scattering effects arise from parametric interactions between light and acoustic or optical phonons in the fiber. Two nonlinear effects fall into this category: stimulated Brillouin scattering (SBS) and stimulated Raman scattering (SRS). The main difference between the two is that optical phonons participate in SRS, while acoustic phonons participate in SBS. In a simple quantum-mechanical picture, a photon of the incident field is annihilated to create a photon at a longer wavelength. The new photon is co-propagated and counter-propagated along the original signal in SRS, while counter-propagated in SBS. Refractive index effects are caused by modulation due to changes in light intensity. There are three types of refractive index effects: self-phase modulation (SPM), cross-phase modulation (XPM), and four-wave mixing (FWM). SPM introduces the change of optical phase by its own intensity, and, this leads to frequency chirp of the pulse, depending on the relative position from the peak. When this nonlinear frequency chirp interacts with the fiber dispersion, the pulse either broadens or compresses, depending on the sign and the amount of dispersion along the fiber. In XPM, the phase of the signal in one wavelength channel is modulated by the intensity fluctuation of the other wavelength channels. In a WDM system, XPM imposes far more damaging effects than SPM because XPM is stronger by a factor of two than SPM when the channel power is the same, and large number of WDM channels can contribute to XPM. FWM is the generation of modulation sideband at new frequencies, due to the phase modulation of channels between lights at different frequencies in multichannel system. FWM causes penalties in a WDM system if the newly generated frequency is either equal to or close to the frequency of existing WDM channels.

Fiber nonlinearities impose significant degradation in optical transmission system, and limit the system on allowable number of channels, channel power, and channel spacing (see **Fig. 6**). But, the effects can be well suppressed by carefully designing the fiber dispersion profile (i.e., dispersion map) in the transmission system, and controlling the optical launch power for each fiber span.

Communication Components

Distributed feedback laser (DFB) is widely used as a light source for metro, long-haul, and undersea applications, due to its narrow spectral width, and wavelength stability. Fabry–Perot (FP) lasers, and vertical cavity surface emitting lasers (VCSEL) are used for local area networks (LAN), and for access applications, because these are cheaper than DFB lasers. For an application, where the cost is the most important parameter with low data modulation speed, light-emitting diode (LED) is used as a light source.

In optical data modulation, the simplest and easiest technique is turning ON or OFF the laser, depending on the binary logic '1' or '0' (direct modulation). This type of modulation is simple and cheap compared to external modulation technique. But, its application is limited by the bandwidth of modulation, and the induced frequency chirp (i.e., difference in optical frequency at the turn ON state and just before the turn OFF state of laser) that limits the transmission distance. Two different types of external modulators are widely used for digital optical data transmission. Electro absorption modulator (EAM) uses the Franz–Keldysh effect, where it is observed that the wavelength of the optical absorption edge in a semiconductor is lengthened by applying an electric field. Electro-optic modulator (EOM) utilizes the linear electro-optic effect, called the Pokels effect, which changes the refractive index of the material caused by and proportional to an applied electric field, so therefore changes the phase of the optical signal. This phase change is used to modulate the intensity of a lightwave through a Mach–Zehnder (MZ) interferometer (see **Fig. 7**). Because the Pockels effect exists only in crystal, lithium niobate crystal ($LiNbO_2$) or electro-optic polymers are used for EOM.

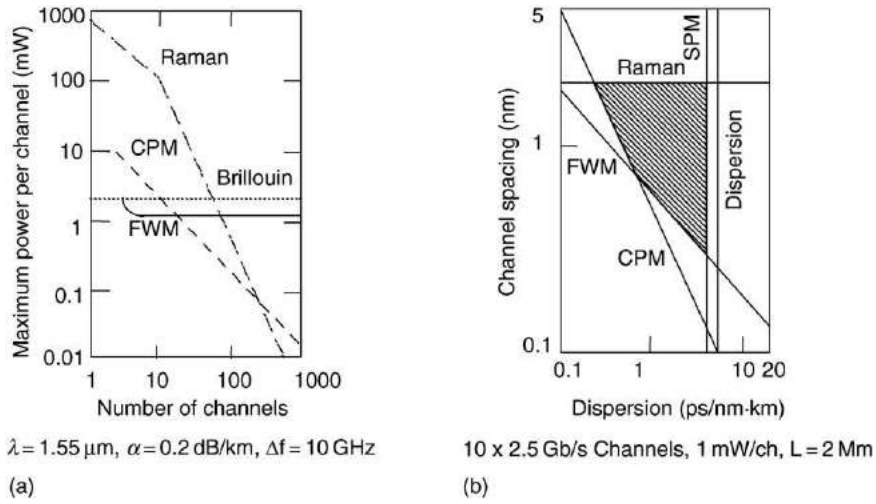


Fig. 6 Limitation due to fiber nonlinearities in (a) maximum channel power and number of channels, and (b) channel spacing and dispersion. Reproduced with permission from Chraplyvy AR (1990) Limitation on lightwave communications imposed by optical-fiber nonlinearities. *IEEE Journal of Lightwave Technology* 8:1548–1557. Copyright © 1990 IEEE.

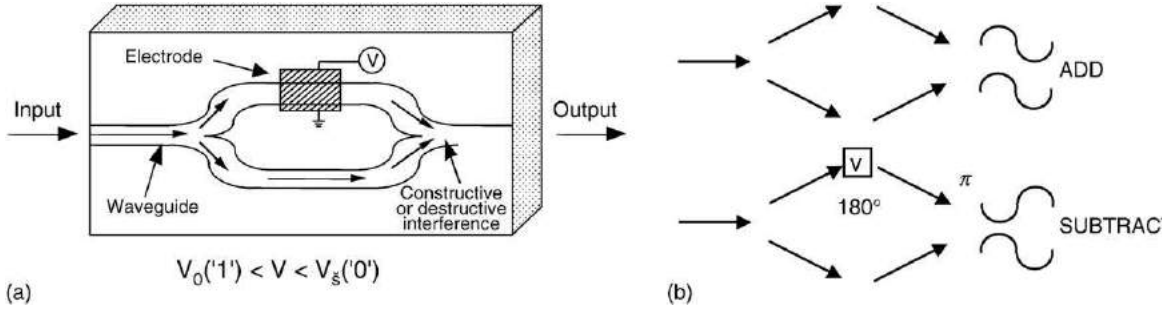


Fig. 7 (a) EOM: MZ interferometer on LiNbO2, (b) operation of EOM; constructive and destructive interference resulting in intensity modulation.

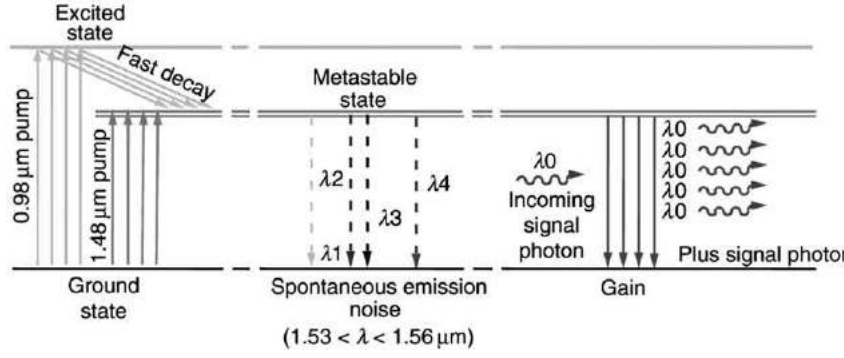


Fig. 8 Energy level and transition diagram of EDFA. Reproduced with permission from Willner AE (1997) Mining the optical bandwidth for a terabit per second. *IEEE Spectrum* Apr.: 32–41. Copyright © 1997 IEEE.

EDFA is a wideband optical amplifier that has merits in that: (i) erbium ions (Er^{3+}) emit light in the 1.55 μm loss-minimum band of optical fiber, (ii) a circular fiber-based amplifier is inherently compatible with a fiber optics system; (iii) it provides amplification independent of bit-rate, modulation format, power, and wavelength; and (iv) it has low distortion and low noise during amplification. EDFA contains a gain medium (i.e., erbium-doped fiber) that must be inverted by a pump source. A signal initiates stimulated emission, resulting in gain, and spontaneous emission occurs naturally, which results in noise.

Fig. 8 shows the energy level and transition diagram of EDFA. A 0.98 μm pump photon is absorbed and excites a carrier (Er^{3+} ions) into higher excited states, and the excited carrier decays rapidly to the first excited state that has a very long lifetime of ~10 ms (metastable state). In contrast, a 1.48 μm pump photon is absorbed and excites a carrier into a metastable state.

This long metastable state is one of the key advantages of EDFA, and intersymbol distortion and interchannel crosstalk are negligible as a result of these slow dynamics. Depending on the external optical excitation signal, this carrier will decay in a stimulated or spontaneous fashion to the ground state and emit a photon. Both absorption and emission spectra have an associated bandwidth depending on the spread in wavelengths that can be absorbed or emitted from a given energy level. This is highly desirable because (i) a pump laser does not need to be an exact wavelength; and (ii) the signal may be at one of several wavelengths, especially in a WDM system. **Fig. 9** shows the bandwidth in the 1.48 μm absorption and 1.55 μm fluorescence spectrum of a typical erbium-doped amplifier.

When the noise characteristics of EDFA are considered, noise figure (NF) is an important parameter, which is defined by $NF \equiv SNR_{in}/SNR_{out}$, where SNR_{in} (SNR_{out}) is the electrical equivalent signal-to-noise ratio (SNR) of the optical wave going into (coming out) of the amplifier, if it were be detected. The typical value of NF in commercially-available EDFA is 4.5–6 dB. The importance of NF in optical communication systems is presented later in this chapter, when the required optical SNR (OSNR) for error-free optical data transmission is considered. **Fig. 10** shows an example of EDFA diagram.

Recent development of EDFA in longer wavelength regimes or L-band (1570–1620 nm) has doubled the useful transmission bandwidth over the conventional band or C-band (1525–1265 nm). L-band gain of EDFA is generally achieved by proper arrangement of pump lasers through longer lengths of erbium-doped fiber than conventional C-band EDFA. **Fig. 11(a)** shows the concept of wideband EDFA. The incoming WDM signal is split into two bands (C- and L-bands), amplified separately, and combined at the output. Gain and NF characteristics of this wideband EDFA is shown at **Fig. 11(b)**.

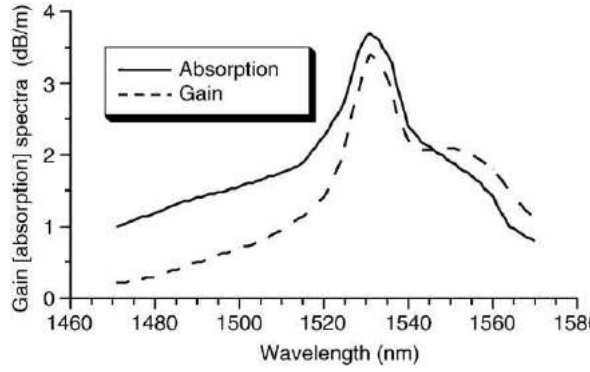


Fig. 9 The absorption and fluorescence spectra for erbium near 1.5 μm. Reproduced with permission from Miniscalco WJ (1991) Erbium-doped glasses for fiber amplifiers at 1550nm. *IEEE Journal of Lightwave Technology* 9: 234–250. Copyright © 1991 IEEE.

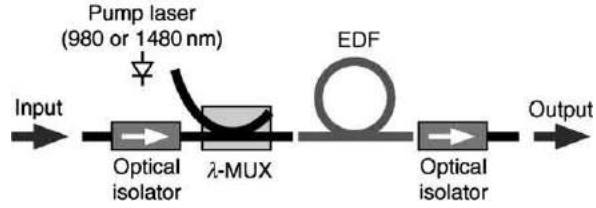


Fig. 10 Typical EDFA diagram.

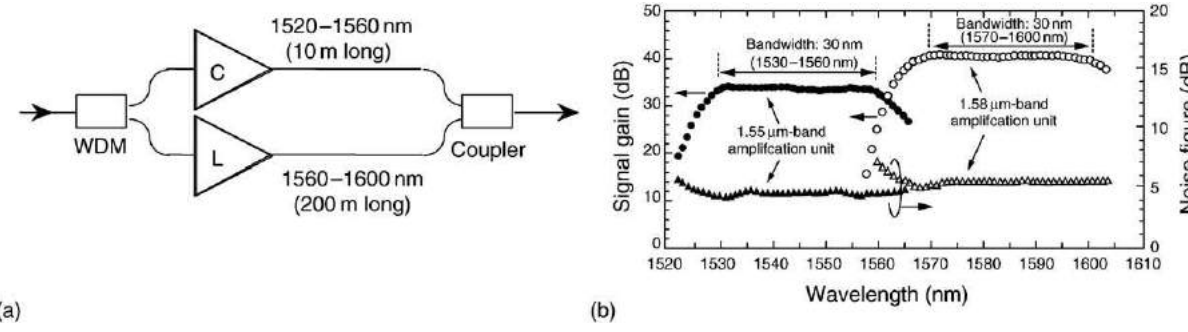


Fig. 11 Wide band EDFA. Reproduced with permission from Yanada M (1997) *Electronics Letters*. Copyright © 1997 IEEE.

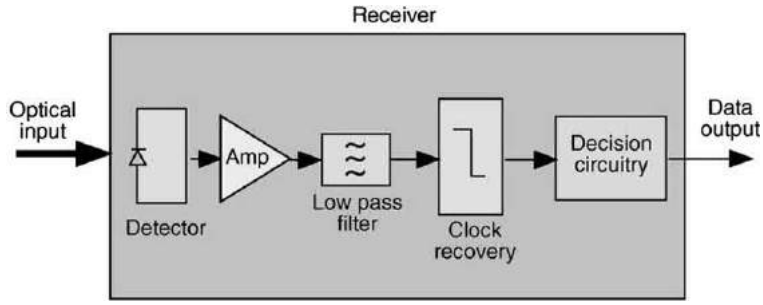


Fig. 12 Block diagram of functions performed in advanced receiver packages.

Because SRS transfers the energy of a lower wavelength channel into the higher wavelength channel in the optical fiber, the transmission fiber itself can be used as a gain medium on specific channel (or channels) when the proper pump signal (or signals) are provided in the fiber. Even though Raman gain is much lower than the gain of EDFA (i.e., <0.1 dB gain/mW pump in Raman amplifier (RA) compared to a few dB gain/mW pump in EDFA), RA has several advantages over EDFA (i) RA has the capability to provide gain at any signal wavelengths, (ii) the amplification window is expandable by combining multipump wavelengths; and (iii) RA offers improved noise performance because the signal is amplified over transmission fiber (i.e., distributed amplifier). Especially improved noise performance increases system OSNR, which can be used to extend system reach or create more wavelength channels on the system. The Raman pumps are typically located ~ 100 nm shorter than the wavelength of the signal to be amplified, and a few 100s mW pump lasers are used to get a 10–15 dB gain.

In general, telecom receiver (RCVR) consists of photo diode (PD) with transimpedence amplifier (TIA), followed by limiting amplifier (LA) or automatic gain control amplifier (AGC), low pass filter, and clock data recovery circuit (CDR) (see **Fig. 12**).

Vertically illuminated positive-intrinsic-negative (PIN) photodiode and avalanche photo diode (APD) are the most common PD that is used in RCVR. In PIN, incoming light is absorbed in the n^- (low doped or intrinsic) region and generates an electron-hole pair, which drifts to a depletion region and collected n^+ and p^+ regions, therefore generating electric current. APD is used for high-sensitivity detection up to 10 Gb/s. It is structured as absorption, grading, and multiplication layers. The multiplication layer of APD provides additional transit time delay compared to PIN, therefore the bandwidth of APD is smaller than PIN in general. TIA is a broadband low noise amplifier that is used to convert and amplify the weak current from PD to the desired output voltage. A low pass filter is used to suppress the noise and data distortion from high frequency components with a typical cut-off frequency ~ 0.7 bitrate. LA and AGC are amplifiers that limit the output voltage as a constant. LA is a nonlinear amplifier that makes a decision either to make logic '1' or '0' based on input voltage level, and limits output voltage as fixed values. AGC is a linear amplifier that limits the output voltage by adjusting the gain of the amplifier based on input voltage level. CDR consists of a phase lock loop (PLL) with a voltage controlled oscillator (VCO) and data flip-flop circuit. The PLL is used to synchronize the VCO to the incoming high-speed bit stream, therefore recovering the system clock. The recovered clock is used in D-FF to retime and regenerate the data.

Data Modulation Formats

In digital optical communications, numerous modulation formats have been developed and utilized, in order to achieve higher spectral efficiency (i.e., ratio of individual channel bit rate to dense WDM (DWDM) separation), and to improve system performance over fiber impairments. A simple amplitude modulation of lightwave is still the most widely used method for optical communication. The easiest amplitude modulation technique is to using a nonreturn-to-zero (NRZ) format. **Fig. 13(a)** shows an example of NRZ binary sequence in time domain, and the corresponding frequency domain spectrum. In NRZ, data '1' and '0' correspond to the ON and OFF state of a lightwave transmitter, respectively, occupying entire bit time (i.e., binary symbol time). When the ON or OFF state of a lightwave transmitter does not occupy the entire bit time (T_B) (i.e., when the data pulse duration (T_d) is less then (T_B) it is called the return-to-zero (RZ). **Fig. 13(b)** shows a binary sequence of RZ, and its frequency spectrum when $T_d \approx 0.5T_B$. The Fourier transform of the square wave is the Sinc function. Therefore, the frequency spectrum of an ideal NRZ and RZ follows the pattern of a Sinc square function. Note that the first null of frequency spectrum of RZ is about twice as high as the null of NRZ.

Even though NRZ is better in spectral efficiency than RZ, RZ has an advantage over NRZ in tolerance of fiber impairments. **Fig. 14** shows an example on eye closure penalty comparison between NRZ and RZ data formats at a 16-channel WDM system at 40 Gb/s data rate. It is clearly seen that the penalty of RZ is much less severe than the penalty of NRZ, as the transmission distance increases. (Note, nonlinearities and dispersion increases as the transmission distance increases.)

There has been much effort to increase the spectral efficiency, to improve the system performance over fiber impairments, or to achieve both simultaneously. **Fig. 15** shows some of examples of these. Chirped RZ (CRZ) is the modulation format inducing frequency chirp on conventional RZ. CRZ is robust on fiber impairment at the cost of losing spectral efficiency. In carrier-suppressed RZ (CS-RZ), a strong un-modulated optical carrier is removed from conventional RZ, therefore suppressing the nonlinear effects. Duo-binary is a three-level signal modulation format generated by a series of delay and added circuits in the

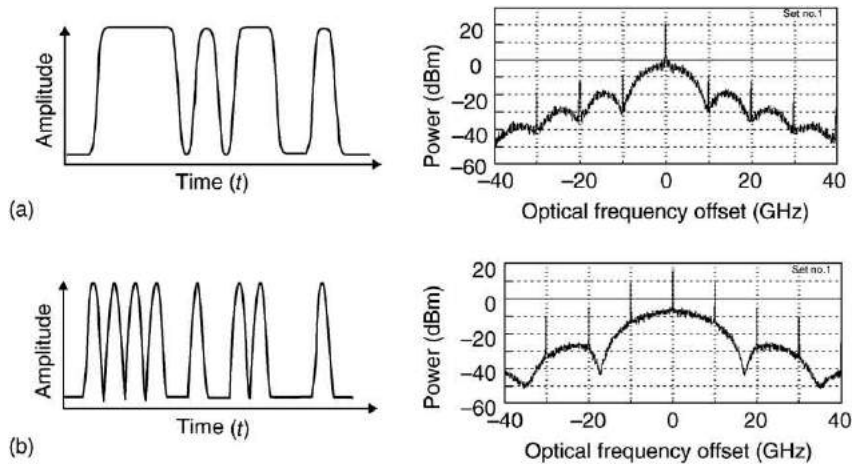


Fig. 13 Amplitude modulation (a) nonreturn-to-zero (NRZ), and (b) return-to-zero (RZ).

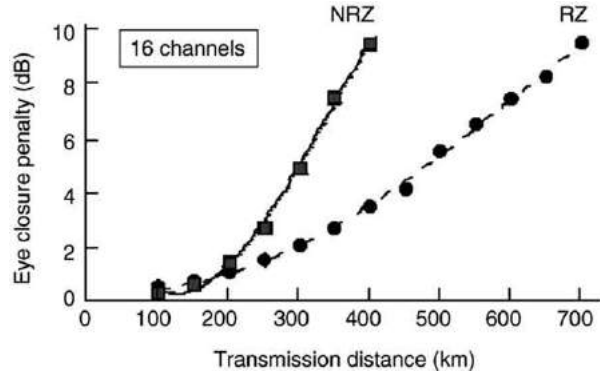


Fig. 14 System performance of different modulation format (NRZ vs. RZ).

transmitter. The advantage of duo-binary is that its bandwidth is reduced to one half of conventional NRZ, therefore having enhanced spectral efficiency, and has a high tolerance in chromatic dispersion, and suppression of fiber nonlinearites.

Single-sideband (SSB) or vestigial singleside band (VSB) can be generated into either NRZ or RZ, by either optically utilizing optical filtering with a sharp cut-off filter, or electrically applying tapped-delay-line filter approximation on both amplitude and phase modulators. SSB (or VSB) has about a two times higher spectral efficiency and higher tolerance on chromatic dispersion than conventional modulation formats.

Data Multiplexing

At their most basic, optical networks can imitate electrical networks in which time division multiplexing (TDM) is overwhelmingly used for digital data transmission. A fiber can carry many time-multiplexed channels, in which each channel can transmit its data in an assigned time slot. A typical TDM link is shown in **Fig. 16**, in which N transmitters are sequentially polled by a fast multiplexer to transmit their data. The time-multiplexed data are sequentially and rapidly demultiplexed at the receiving node.

Time multiplexing and demultiplexing functions can be performed either electrically with an electrical time (DE)MUX switch and an optical transmitter/receiver, or optically with multiple optical transmitters/receivers and an optical time (DE)MUX switch. The major advantage of TDM is that there is no output–port contention problem (each data bit occupies its own time slot and there is only a single high-speed signal present at any given instant). The disadvantage of TDM is that the scheme requires a ultra-high-speed switching component if the individual signals are themselves high-speed and if there are many users.

In WDM, multiple wavelength channels are transmitted through a single fiber, therefore enabling the fiber to carry more throughput. By using wavelength-selective devices for the ON and OFF ramps, independent signal routing also can be accomplished. **Fig. 17(a)** shows a simple point-to-point WDM system in which several channels are multiplexed at one node, the combined signals transmitted across a distance of fiber, and the channels demultiplexed at a destination node. As shown in

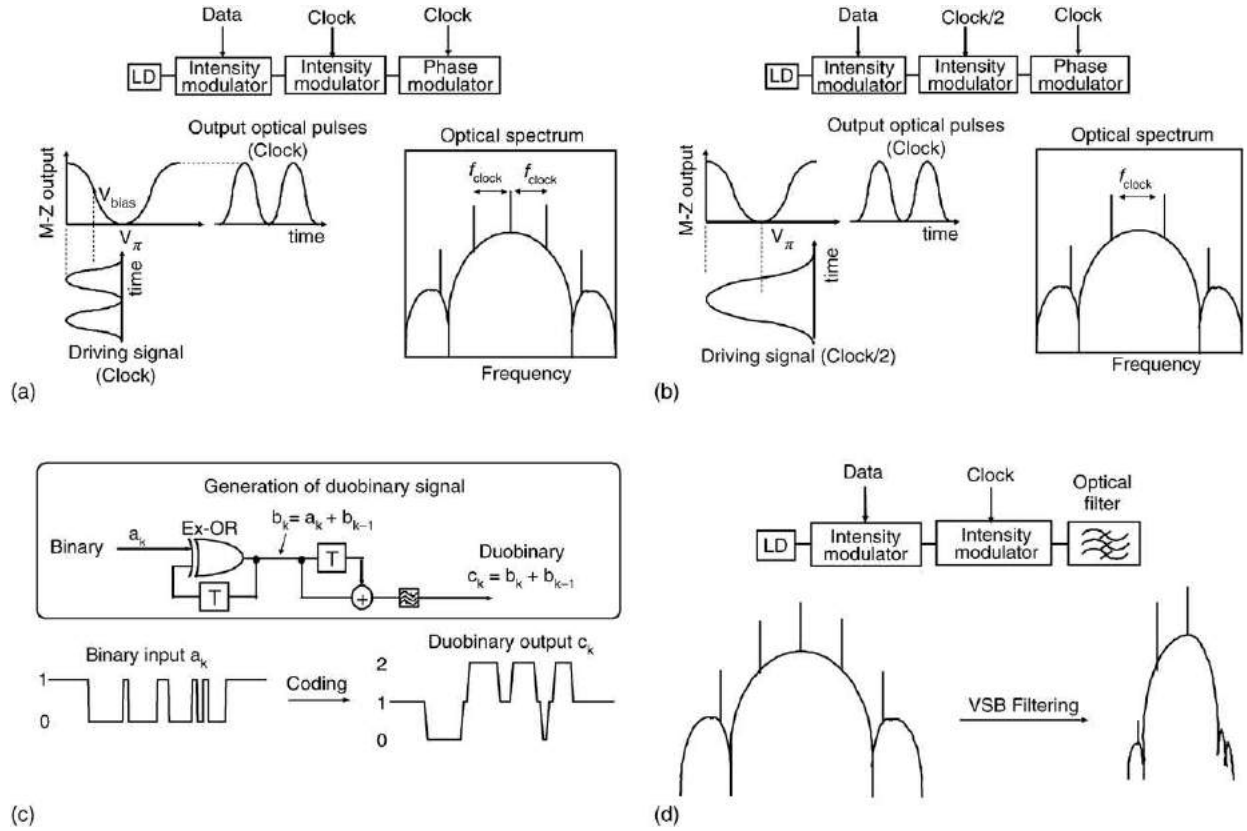


Fig. 15 Examples of bandwidth efficient modulation formats. (a) Chirped RZ (CRZ) generation, (b) carrier suppressed-RZ (CS-RZ) generation, (c) duo-binary generation, and (d) vestigial single side band (VSB) RZ generation.

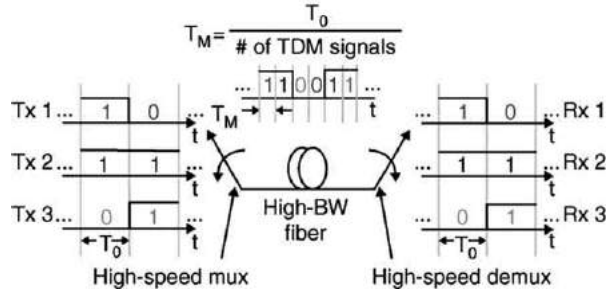


Fig. 16 Concept of bit-interleaving TDM.

Fig. 17(b), the wavelength becomes the signature address for either the transmitter or the receivers, and the wavelength will determine the routing path through an optical network.

Many interesting challenges face the eventual implementation of WDM systems and networks. Several of these challenges involve the control and management of the data through this novel high-speed network. Fig. 18 shows a small subset of critical component technologies typically required for a WDM system, including multiple-wavelength transmitters, multiport star couplers, passive and active wavelength routers, EDFAs, and tunable optical filters.

Some generic goals that a WDM-device technologist aims to achieve include; large wavelength tuning range; multi-user capability; wavelength selectivity and repeatability; low cross-talk; high extinction ratio; minimum excess losses; fast wavelength tunability; high-speed modulation bandwidth; low residual chirp; high finesse; low noise; robustness; high yield; potential low cost. Depending on system requirements, device availability and cost, WDM technologies divide into two; course-WDM (CWDM) and dense-WDM DWDM. Fig. 19 shows the wavelength allocation for CWDM and DWDM. CWDM technology uses an International Telecommunication Union (ITU) standard 20 nm spacing between the wavelengths from 1310 nm to 1610 nm. Also DWDM technology uses an ITU standard 100 GHz or 200 GHz between wavelengths, arranged in several bands at around 1500–1600 nm.

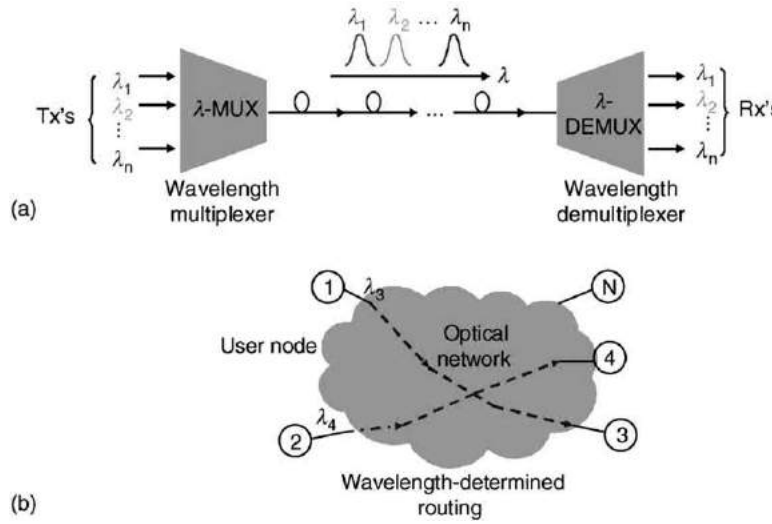


Fig. 17 (a) A simple point-to-point WDM transmission system; (b) a generic multiuser network in which the communication links and routing paths are determined by the wavelengths used within the optical switching fabric.

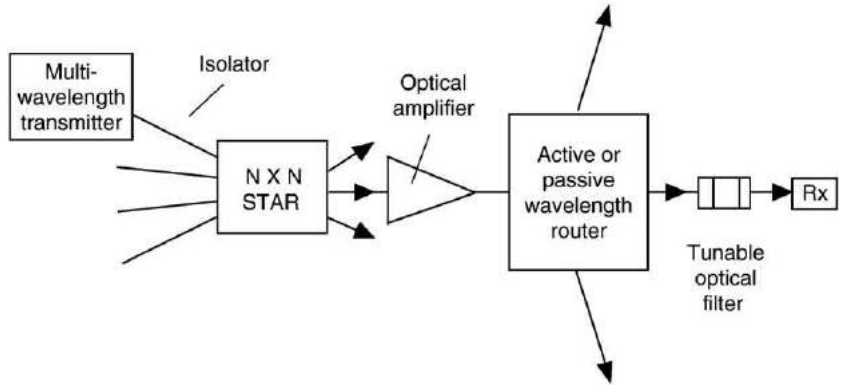


Fig. 18 Schematic of a small subset of enabling device technologies for a WDM system.

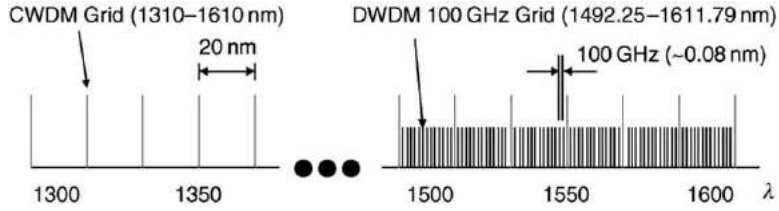


Fig. 19 WDM wavelength allocation.

Dispersion Management and Compensation

As mentioned previously, chromatic dispersion is one of the most basic characteristics of fiber, although it is possible to manufacture fiber that induces zero chromatic dispersion. But such fiber is incompatible with the deployment of a WDM system since harmful nonlinear effects are generated, therefore chromatic dispersion must exist, and must be compensated for. The effect of chromatic dispersion is cumulative and increases quadratically with the data rate (see **Fig. 20**).

The quadratic dependence of dispersion with the data rate is a result of two effects. First, a doubling of the data rate will double the Fourier-transformed frequency spectrum of the signal, thereby doubling the effect of dispersion. Second, the same doubling of the data rate makes the data pulse only half as long in time and therefore twice as sensitive to temporal spreading due to dispersion. The conventional wisdom for the maximum distance over which data can be transmitted is to consider a broadening of

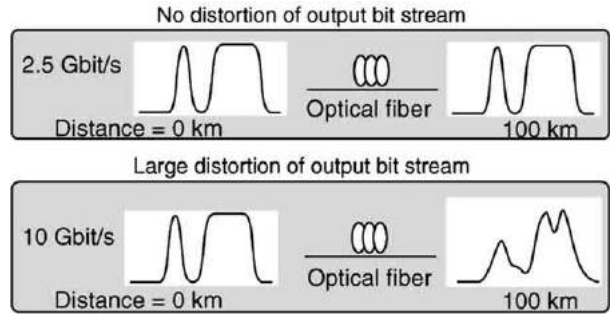


Fig. 20 Pulse broadening at two different data rates (2.5 Gbit/s, and 10 Gbit/s) as a result of quadratic nature of chromatic dispersion.

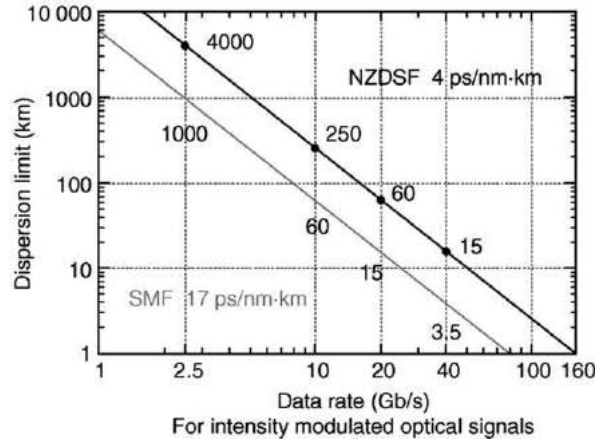


Fig. 21 Dispersion-limited (uncompensated) transmission distance in single-mode fiber (SMF) as a function of data rate for intensity-modulated optical signals. Courtesy of L.D. Garrett.

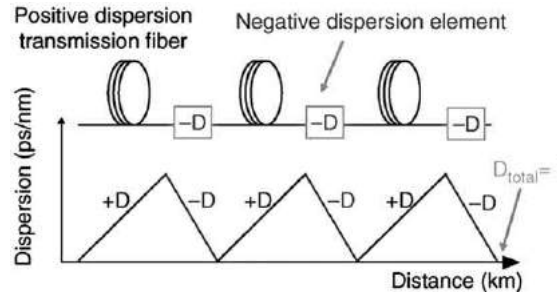


Fig. 22 Typical static management of chromatic dispersion.

the pulse equal to the bit time period. For a bit period B a dispersion value D and a spectral width $\Delta\lambda$ the dispersion-limited distance is given by $L_D = 1/(D \cdot B \cdot \Delta\lambda)$ (see Fig. 21).

In theory, compensation of chromatic dispersion for high-speed or long-distance systems can be fixed in value if each link's dispersion value is known. A simple yet elegant solution is to create a dispersion 'map', in which the designer of a transmission link alternates elements that produce positive and then negative dispersion (see Fig. 22). This is a very powerful concept: at each point along the fiber the dispersion has some nonzero value, effectively eliminating FWM and XPM, but the total accumulated dispersion at the end of the fiber link is zero, so that minimal pulse broadening is induced. The specific system design, as to the periodicity of management, depends on several variables, but a typical number for SMF as the embedded base is a compensation at every 80 km in a 10 Gb/s system. Dispersion-compensating fiber (DCF) has been generally used as a dispersion compensating element.

However, there are several important aspects of optical systems and networks that make tunable dispersion compensation solutions attractive, including: (i) it significantly reduces the inventory of different required types of compensation modules; (ii) it tunes to adapt to routing path changes in a reconfigurable network, (iii) it tracks dynamic changes in dispersion due to environment, and (iv) it achieves a high degree of accuracy necessary for 40 Gb/s channels (see Fig. 23).

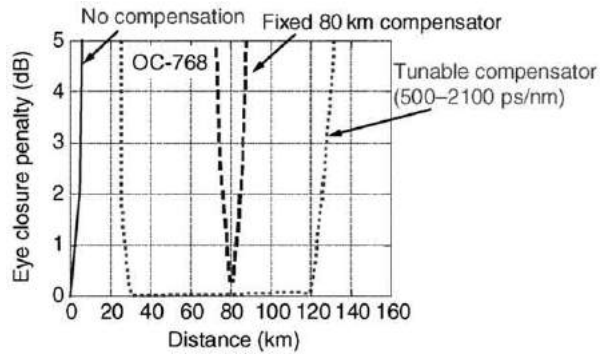


Fig. 23 Effect of tunable compensation at 40 Gb/s (OC-768); it allows a wide range of transmission distance. On the contrary, the 80 km fixed compensator allows only small range of transmission around 80 km.

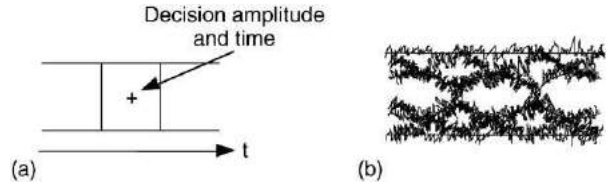


Fig. 24 (a) A perfect square eye diagram, (b) a closed eye diagram due to bandwidth, fiber impairments, noise, and timing jitter.

One brute-force method to achieve tunable dispersion compensation is to build a module with optical switches used to add or remove sections of fixed DCF to achieve a discrete set of dispersion compensation. Many other elegant, yet viable, approaches have been developed for tunable dispersion compensation, and these approaches include linearly chirped FBG with nonuniform heating, nonlinearly chirped Fiber Bragg grating (FBG) with a simple mechanical stretcher, virtually imaged phase array, electronic tap delay filter with weights, multiple stage all pass filter, etc.

System Performance Parameters

An eye diagram provides a simple and useful tool to visualize intersymbol interference between data bits. Fig. 24(a) shows a perfect eye diagram. A square bit stream (i.e., series of symbol '1's and '0's) is sliced into sub-bit stream with predetermined eye intervals (i.e., several bit periods), and displayed through bit analyzing equipment (e.g., digital channel analyzer), overlapping the sliced sub-bit stream in order to obtain the eye diagram. When a perfect transmitter and receiver (i.e., infinite receiver bandwidth with zero noise characteristics and jitter), and a perfect transmission media (i.e., no dispersion and nonlinearites) are used, the received eye diagram is shaped as a perfect rectangular. In reality, the transmitter and receiver have a limited bandwidth with noise and jitter, and the transmission media (i.e., optical fiber) has dispersion and nonlinearites. Therefore, the eye diagram deviates from the perfect rectangular shape. Fig. 24(b) shows the eye diagram close to a real situation. The shape of the eye is generally broadened and distorted (i.e., eye is closed) due to limited bandwidth and fiber impairments, and noise and timing jitter are added onto this broadened and distorted eye shape.

The Q-factor (q) is also an important system parameter widely used in long-distance optical transmission system design. It is defined as the electrical signal-to-noise ratio before the decision circuit at receiver. The Q-factor is directly related to bit-error-rate (BER: the percentage of bits that has an error relative to total number of bits received in a specific time) by: $\mathrm{BER} = 0.5 \cdot \mathrm{erfc}(q/2^{0.5})$, where $\mathrm{erfc}(x)$ is the complementary error function. The Q-factor is an unitless linear ratio, and is express in dB by $20 \cdot \log(q)$. As a conventional relationship, Q-factor of 15.6 dB (i.e., linear ratio 6) is required to achieve $\mathrm{BER} = 10^{-9}$. Fig. 25 shows the relationship of eye diagram, Q-factor, and BER.

Power penalty (PP) is one of the most important system parameters, and is defined as the received optical power difference in dB with and without signal impairments at a specified BER (conventionally 10^{-9}), from measured BER versus optical power curve. Therefore, 1 dB PP means that a system with signal impairments requires 1 dB more optical power at the receiver in order to achieve the same BER performance compared to the system without signal impairments. Fig. 26 shows the typical BER curve as a function of received optical power and PP. Note that PP reduces from $\sim$3.5 dB to $\sim$1 dB with dispersion compensation.

In long-haul and undersea transmission system with many EDFA chains, OSNR is an important system parameter to design and characterize the optical transport system. OSNR is defined as the ratio of the power of signal channel to the power of ASE in a specified optical bandwidth (conventionally 0.1 nm). The OSNR (in dB) of a signal channel at the end of the system is approximated by: $\mathrm{OSNR\ (dB)} = 58 + P_{\mathrm{out}} - L_{\mathrm{span}} - \mathrm{NF} - 10\ \log(N_{\mathrm{amp}})$, when the system has N_{amp} fiber spans, span loss L_{span}

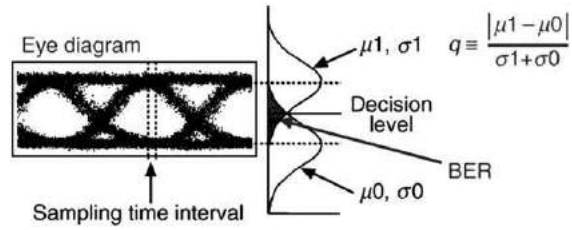


Fig. 25 The relationship of eye diagram, Q-factor, and BER ($\mu 1$, $\mu 2$: mean levels of logic '1' and '0', $\sigma 1$, $\sigma 2$: standard deviation of logic '1' and '0').

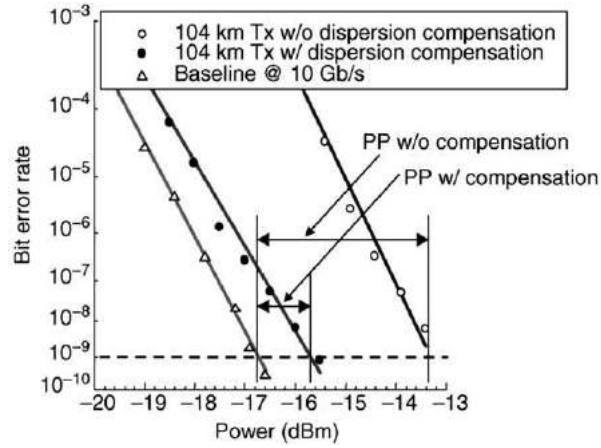


Fig. 26 BER curve and power penalty.

(in dB) followed by an optical amplifier with output power P_{out} (in dBm) per channel launched into the span and noise figure (NF) (in dB). As an example, a typical OSNR requirement for BER $= 10^{-9}$ is about 17 dB at 10 Gb/s data rate. The required OSNR should be increased by 6 dB when the data rate is increased by a factor of 4. Improving the amplifier's NF can increase OSNR, and the improved OSNR can be used to increase the system reach, reducing the channel power in order to suppress the nonlinearities, etc.

See also: Historical Development of Optical Communication Systems. Measuring Fiber Characteristics

Further Reading

Emmanuel, D., 1994. Erbium-Doped Fiber Amplifiers. New York: Wiley.
Govind, P.A., 1995. Nonlinear Fiber Optics. San Diego, CA: Academic Press.
Ivan, K., Tingye, L., 2002. Optical Fiber Telecommunications IV A Components. San Diego, CA: Academic Press.
Ivan, K., Tingye, L., 2002. Optical Fiber Telecommunications IV B Systems and Impairments. San Diego, CA: Academic Press.
Leonid, K., Sergio, B., Alan, W., 1996. Optical Fiber Communication Systems. Norwood, MA: Artech House Inc.

Historical Development of Optical Communication Systems

G Keiser, PhotonicsComm Solutions, Inc., Newton Center, MA, USA

Introduction

This article describes the development of optical fiber communication systems. After describing some of the motivations for using optical fiber communications and the advantages of this technology, the key milestones and the principal people involved in developing optical fibers and compatible light sources are presented. Following this, the article looks at the evolution of fielded systems and the use of optical fiber links in undersea applications.

One of the principal needs of people since antiquity has been to communicate. This need created interests in devising communication systems for sending messages from one distant place to another. Many forms of such systems have appeared over the years. The basic motivations behind each new one were either to improve the transmission fidelity, to increase the data rate so that more information could be sent, or to increase the transmission distance between relay stations. Prior to the nineteenth century, all communication systems operated at a very low information rate and involved only optical or acoustical means, such as signal lamps or horns. One of the earliest known optical transmission links, for example, was the use of a fire signal by the Greeks in the eighth century BC for sending alarms, calls for help, or announcements of special events.

The invention of the telegraph by Samuel FB Morse in 1838 ushered in the era of electrical communications. In the ensuing years an increasingly larger portion of the electromagnetic spectrum, shown in Fig. 1, was utilized for the conveying of information from one place to another. The reason for this trend is that, in electrical systems, the data usually are transferred over the communication channel by superimposing the information onto a sinusoidally varying electromagnetic wave, which is known as the carrier. When reaching its destination the information is removed from the carrier wave and processed as desired. Since the amount of information that can be transmitted is related directly to the frequency range over which the carrier operates, increasing the carrier frequency theoretically increases the available transmission bandwidth and, consequently, provides a larger information capacity. Thus the trend in electrical communication system developments was to employ progressively higher frequencies (shorter wavelengths), which offer corresponding increases in bandwidth or information capacity. This activity led to the birth of communication mechanisms such as radio, television, microwave, and satellite links.

Another important portion of the electromagnetic spectrum encompasses the optical region shown in Fig. 1. In contrast to electrical communications, transmission of information in an optical format is carried out, not by frequency modulation of the carrier but by varying the intensity of the optical power. Similar to the radio-frequency spectrum, two classes of transmission medium can be used: an atmospheric channel or a guided-wave channel. Whereas transmission of optical signals through the atmosphere was done thousands of years ago, the use of a guided-wave optical channel such as an optical fiber is a fairly recent application.

Fiber optic communication systems have a number of inherent advantages over their copper-based and radio-transmission counterparts. Fiber optic cable can transmit at a higher capacity, thereby reducing the number of physical lines and the amount of equipment needed for a given transmission span. The lower weight and smaller size of optical fibers offer a distinct advantage over heavy, bulky copper cables in crowded underground cable ducts, in ceiling-mounted cable trays, and in mobile platforms such as aircraft and ships. Their dielectric composition is an especially important feature of optical fibers, since this ensures freedom from electromagnetic interference between adjacent fibers, eliminates ground loops, and results in extremely low fiber-to-fiber crosstalk.

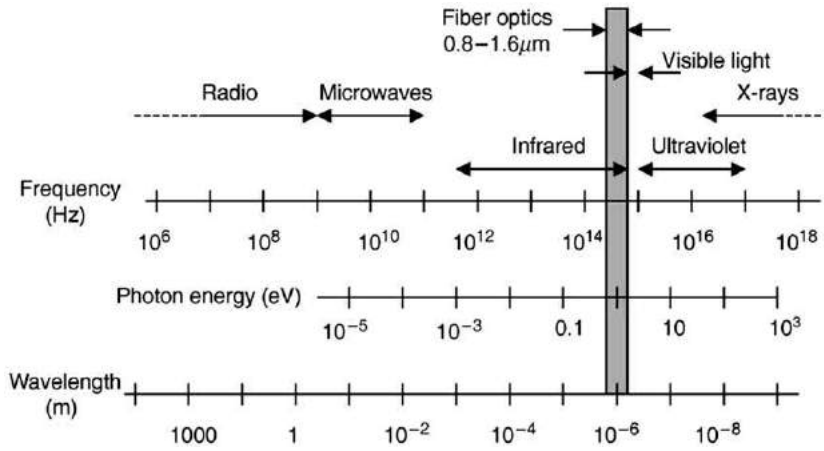


Fig. 1 Electromagnetic spectrum for electrical and optical communications.

In addition, an optical fiber affords a high degree of data security since the optical signal is well confined within the optical waveguide.

Optical Fiber Development

Since an atmospheric channel requires a line-of-sight link and because it can be adversely affected by weather conditions, a guided-wave channel is the preferred approach in most cases for communication system applications. A challenge in using an optical fiber channel is to have a flexible, low-loss medium that transfers the optical signal over long distances without significant attenuation and distortion. Glass is an obvious material for such applications. The earliest known glass was made around 2500 BC and glass already was drawn into fibers during the time of the Roman Empire. However, such glasses have very high attenuations and thus are not suitable for communication applications. One of the first known attempts of using optical fibers for communication purposes was a demonstration in 1930 by Heinrich Lamm of image transmission through a short bundle of optical fibers for potential medical imaging. However, no further work was done beyond the demonstration phase since the technology for producing reasonably low-loss fibers with good light confinement was not yet mature.

Further work and experiments continued on using optical fibers for image transmission and by 1960 glass-clad fibers had attenuations of about one decibel per meter. This attenuation allowed fibers to be used for medical imaging, but it was still much too high for communications. Optical fibers had attracted the attention of researchers at that time, because they were analogous in theory to plastic dielectric waveguides used in certain microwave applications. In 1961, Elias Snitzer demonstrated this similarity by drawing fibers with cores so small they carried light in only one waveguide mode. He published a classic theoretical description of single-mode fibers in May 1961. However, to be useful for communication systems, optical fibers would need to have a loss of no more than 10 or 20 decibels per kilometer.

As a reminder, decibels measure the ratio of the output power to the input power on a logarithmic scale. The abbreviation for a decibel is 'dB.' The power ratio in decibels is given by the expression '10 log (power out/power in).' As an example, a power loss of 20 dB over a 1-km distance in an optical fiber means that only 1% of the light injected into the fiber comes out of the other end. **Table 1** gives some typical examples of various power ratios and their decibel equivalents.

In the early 1960s Charles Kao pursued the idea of using a clad glass fiber for an optical waveguide, building on optical waveguide research being done at the Standard Telecommunication Laboratories in England. After he and George Hockman painstakingly examined the transparency properties of various types of glass, Kao made a prediction, in 1966, that losses of no more than 20 dB/km were possible in optical fibers. In July 1966, Kao and Hockman presented a detailed analysis for achieving such a loss level. Kao then went on to actively advocate and promote the prospects of fiber communications, which generated interest in laboratories around the world to reduce fiber loss. It took four years to reach Kao's predicted goal of 20 dB/km, and the final solution was different from what many had expected.

To understand the process of making a fiber, consider the schematic of a typical fiber structure shown in **Fig. 2**. A fiber consists of a solid dielectric (glass or plastic) cylinder of refractive index n_1 called the core. This is surrounded by a dielectric cladding which has a refractive index n_2, that is less than n_1, in order to achieve light guiding in the fiber. This structure is then encapsulated by a buffer coating to protect the fiber from abrasion and outside contaminants. The first step in making a fiber is to form the clear glass

Table 1 Examples of some optical power ratios and their decibel equivalents

Power ratio (P_{out}/P_{in})	*Decibel equivalent (dB)*
0.01	−20
0.10	−10
0.50	−3
1	0
2	3
10	10
100	20

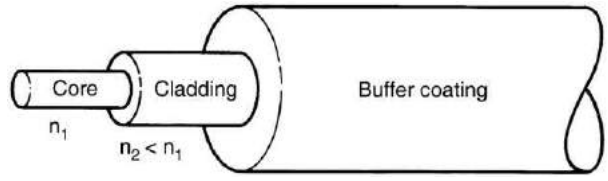


Fig. 2 Typical optical fiber structure.

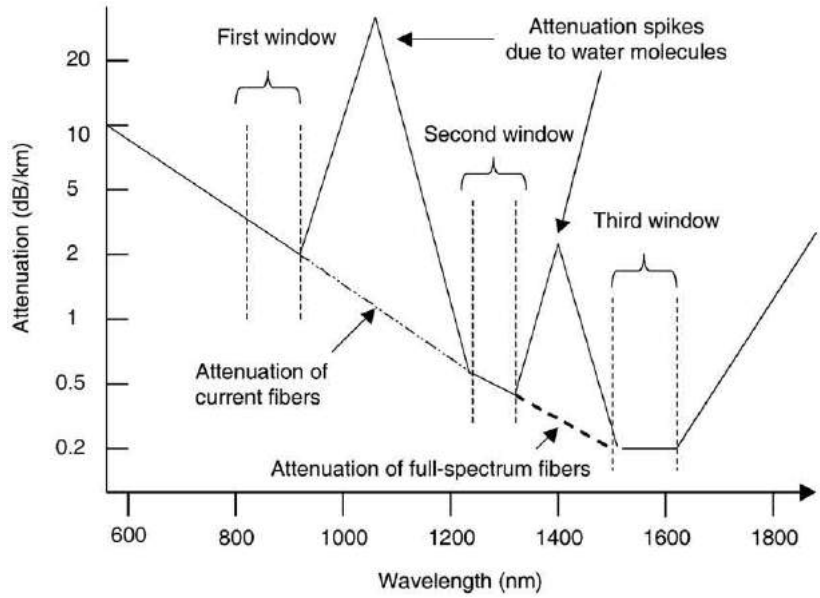


Fig. 3 Attenuation versus wavelength approximation for silica fibers.

rod or tube called a preform. This normally is done by a vapor-phase oxidation process. The preform has two distinct regions that correspond to the core and cladding of the eventual fiber. Fibers are made from the preform by precision feeding it into a circular furnace that softens the end of the preform to the point where it can be drawn into a very thin filament which becomes the optical fiber.

Most researchers had tried to purify the compound glasses used for standard optics, which are easy to melt and draw into fibers. A different approach was taken at the Corning Glass Works where Robert Maurer, Donald Keck, and Peter Schultz started with fused silica. This material can be made extremely pure, but has a high melting point and a low refractive index. Silica has the approximate attenuation versus wavelength characteristic shown in Fig. 3. Note that for early silica fibers there are regions of low attenuation around 850, 1310, and 1550 nm, which the literature refers to as the first, second, and third windows, respectively. The large attenuation spikes in the 1000- and 1400-nm spectral regions are due to absorption by residual water molecules in the glass. Although ultra-pure material processing techniques can eliminate these spikes to produce what are known as full-spectrum or low-water-peak fibers, most installed fibers still have a relatively large attenuation between 1350 and 1500 nm.

Note that since the attenuation spikes no longer are present in various types of fibers, the idea of operating windows has been replaced by the concept of spectral bands. The International Telecommunications Union (ITU) has designated six spectral bands for use in intermediate-range and long-distance optical fiber communications within the 1260 to 1675-nm region. The regions, which are known by the letters O, E, S, C, L, and U, are defined as follows:

- Original Band (O-Band): 1260 to 1360 nm
- Extended Band (E-Band): 1360 to 1460 nm
- Short Band (S-Band): 1460 to 1530 nm
- Conventional Band (C-Band): 1530 to 1565 nm
- Long Band (L-Band): 1565 to 1625 nm
- Ultra-long Band (U-Band): 1625 to 1675 nm

The Corning team made cylindrical preforms by depositing purified materials from the vapor phase. To produce a fiber that has light guiding properties they carefully added controlled trace levels of titanium to the core to make its refractive index slightly higher than that of the cladding without raising the attenuation significantly. In September 1970, they announced the fabrication of single-mode fibers with an attenuation of 17 dB/km at the 633-nm helium-neon line. The fibers were fragile, but independent tests at the British Post Office Research Laboratories facility in Martlesham Heath, England confirmed the low loss.

This dramatic breakthrough was the first among the many developments that opened the door to fiber optic communications. The ensuing years saw further reductions in optical fiber attenuations. By mid-1972, Maurer, Keck, and Schultz at Corning had made multimode germania-doped fibers with a 4-dB/km loss and much greater strength than the earlier brittle titania-doped fibers. In order to couple a sufficient amount of optical power into a fiber, early applications used multimode fibers with a refractive-index gradient between core and cladding of around 2% and core diameters of 50 or 62.5 micrometers.

Single-mode fibers have much smaller core diameters on the order of 9 micrometers in order to allow only one propagation mode. This type of fiber has a much higher transmission capacity since the effect of modal dispersion is eliminated. The first single-mode fibers were optimized to have a zero dispersion value at 1310 nm, since the silica material used at that time exhibited a low

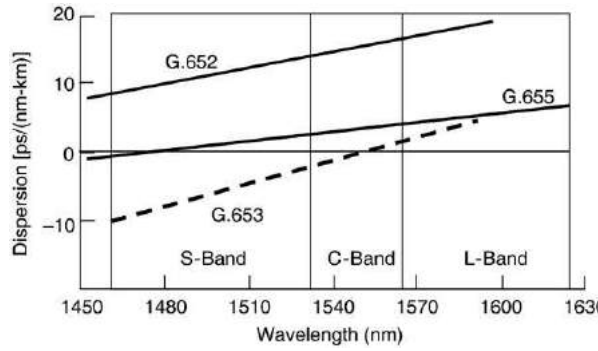


Fig. 4 Dispersion characteristics of three types of standard optical fibers.

loss within a spectral band around this wavelength. Fig. 4 shows the dispersion characteristic of this type of fiber as a function of wavelength in the S-, C-, and L-bands. As a result of its widespread use in early single-mode transmission systems, this fiber design has been standardized by the International Telecommunication Union (ITU-T) under the designation Recommendation G.652.

Standard G.652 silica fibers provide the lowest attenuation at 1550 nm, but have a much larger signal dispersion at this wavelength than at 1310 nm. Since system designers wanted to use fibers at a 1550-nm wavelength, in order to transmit high-speed data over longer distances, fiber manufacturers overcame the larger signal dispersion limitation by creating the so-called dispersion-shifted fibers. This was done through a clever manipulation of the core and cladding designs that allowed the zero-dispersion point to shift to longer wavelengths. In particular, the ITU-T created a specification for operation at 1550 nm which is designated as recommendation G.653. Fig. 4 also shows the dispersion characteristic of this type of fiber as a function of wavelength.

Although the G.652 and G.653 fibers work well for single-wavelength operation, a different type of fiber, having non-zero dispersion within a broad spectral range, is needed when implementing systems that use many independent light sources simultaneously within a particular wavelength band. This led to the specification of the non-zero dispersion shifted fiber (NZDSF) in the ITU-T Recommendation G.655 that are designated to operate around 1550 nm. The main purpose of having a positive dispersion value over the entire operating spectrum is to mitigate a nonlinear optical effect called four-wave mixing (FWM), which is analogous to intermodulation distortion in electrical systems. Fig. 4 shows the dispersion characteristic of the G.655 fiber in the S-, C-, and L-bands.

Light Sources

The fiber by itself is not practical unless there is a compatible optical source for launching light signals into it. The most suitable device for this is a semiconductor light-emitting diode (LED) or a laser diode. In the 1960s, a great deal of effort took place to achieve laser action in pn-junction diodes. The early devices were GaAs and GaAsP lasers that operated at a temperature of 77 K, emitted at a wavelength around 850 nm, and had high lasing threshold current densities. To make devices that were more application-friendly by operating at room temperature, structures consisting of sandwiched layers of AlGaAs and GaAs were investigated. Finally, in 1970, researchers at Bell Laboratories and a team at the Ioffe Physical Institute in Leningrad made the first semiconductor diode lasers based on a layered AlGaAs/GaAs/AlGaAs structure that were able to emit continuous-wave light in the 850-nm region at room temperature.

Major improvements in laser diode performance and reliability followed this achievement during the next decade. In addition, around 1976, room-temperature laser diodes operating at longer wavelengths, in the 1100 to 1600-nm region, started to appear. Of particular interest were GaInAsP/InP-based laser diodes emitting in the 1310-nm and 1550-nm low-loss windows of optical fibers. The progressive development of ever-improving devices during the 1980s and 1990s, included single-frequency emission with narrow linewidths under continuous operation, low levels of chirp under direct modulation, high output power, and the ability to tune specially constructed laser diodes over a wavelength range of up to 30 nm.

Fielded Systems

The bit rate-distance product $B \times L$ where B is the transmission bit rate and L is the repeater spacing, measures the transmission capacity of optical fiber links. Since the inception of optical fiber communications in the mid-1970s, the link transmission capacity has experienced a tenfold increase every four years. Several major technology advances spurred this growth. Among the technology developments are laser diodes emitting over an extremely narrow spectral band, optical amplifiers, fibers with low losses and low dispersions, and the concepts of wavelength division multiplexing.

Some of the initial telephone-system field trials in the USA were carried out in 1977 by GTE in Los Angeles and by AT&T in Chicago. These transmission links were largely for the trunking of telephone lines, which are digital links consisting of

Table 2 Commonly used SONET and SDH transmission rates

SONET level	*Electrical level*	*Line rate (Mb/s)*	*SDH equivalent*	*Common rate name*
OC-1	STS-1	51.84	–	
OC-3	STS-3	155.52	STM-1	155 Mb/s
OC-12	STS-12	622.08	STM-4	622 Mb/s
OC-24	STS-24	1244.16	STM-8	
OC-48	STS-48	2488.32	STM-16	2.5 Gb/s
OC-96	STS-96	4976.64	STM-32	
OC-192	STS-192	9953.28	STM-64	10 Gb/s
OC-768	STS-768	9,813.12	STM-256	40 Gb/s

time-division-multiplexed 64-kb/s voice channels. Similar demonstrations were carried out in Europe and Japan. Applications ranged from 45 to 140 Mb/s with repeater spacings around 10 km.

With the advent of high-capacity fiber optic transmission lines in the 1980s, service providers established a standard signal format called synchronous optical network (SONET) in North America and synchronous digital hierarchy (SDH) in other parts of the world. These standards define a synchronous frame structure for sending multiplexed digital traffic over optical fiber trunk lines. The basic building block and first level of the SONET signal hierarchy is called the synchronous transport signal–level 1 (STS-1), which has a bit rate of 51.84 Mb/s. Higher-rate SONET signals are obtained by byte-interleaving N STS-1 frames, which are then scrambled and converted to an optical carrier–level N (OC-N) signal. Thus the OC-N signal will have a line rate exactly N times that of an OC-1 signal. For SDH systems, the fundamental building block is the 155.52-Mb/s synchronous transport module–level 1 (STM-1). Again, higher-rate information streams are generated by synchronously multiplexing N different STM-1 signals to form the STM-N signal. **Table 2** shows commonly used SDH and SONET signal levels.

The development of optical sources and photodetectors capable of operating at 1310 nm allowed a shift in the transmission wavelength from 850 nm to 1310 nm. This resulted in a substantial increase in the repeaterless transmission distance for long-haul telephone trunks, since optical fibers exhibit lower power loss and less signal dispersion at 1310 nm. Intercity applications first used multimode fibers, but in 1984 switched exclusively to single-mode fibers, which have a significantly larger bandwidth. Bit rates for long-haul links typically range between 155 and 622 Mb/s (OC-3 and OC-12), and in some cases up to 2.5 Gb/s (OC-48), over repeater spacings of 40 km. Both multimode and single-mode 1310-nm fibers are used in local area networks, where bit rates range from 10 Mb/s to 100 Mb/s over distances ranging from 500 m to tens of kilometers.

In the next step of system evolution, links operating in the low-loss window around 1550 nm attracted much attention for high-capacity, long-span terrestrial and undersea transmission links. These links routinely carry traffic at 2.5 Gb/s over 90-km repeaterless distances. By 1996, advances in high-quality lasers and receivers allowed single-wavelength transmission rates of 10 Gb/s (OC-192).

In 1989, the introduction of optical amplifiers gave a major boost to fiber transmission capacity. Although there are GaAlAs-based solid-state optical amplifiers for the first window and InGaAsP amplifiers for the second window, the most successful and widely used devices are erbium-doped fiber amplifiers (commonly called EDFAs) operating in the 1530 to 1560-nm range and Raman fiber amplifiers that are used for operation in the 1560 to 1600-nm region.

During the same time period, impressive demonstrations of long-distance high-capacity systems were made using optical soliton signals. A soliton is a nondispersive pulse that makes use of nonlinear dispersion properties in a fiber to cancel out chromatic dispersion effects. As an example, solitons at rates of 10 Gb/s have been sent over a 12200-km experimental link using optical amplifiers and special modulation techniques.

Entrance of Wavelength Division Multiplexing

The use of wavelength division multiplexing (WDM) offers a further boost in fiber transmission capacity. The basis of WDM is to use multiple sources operating at slightly different wavelengths to transmit several independent information streams over the same fiber. Although researchers started looking at WDM in the 1970s, during the ensuing years it generally turned out to be easier to implement higher-speed electronic and optical devices than to invoke the greater system complexity called for in WDM. However, a dramatic surge in WDM popularity started in the early 1990s, as electronic devices neared their modulation limit and high-speed equipment became increasingly complex.

Fig. 5 shows the concept of implementing many closely spaced wavelengths within a spectral band centered around 1552.524 nm. This scheme is referred to as dense WDM or DWDM. Conceptually, the DWDM scheme is the same as frequency division multiplexing (FDM) used in microwave radio and satellite systems. Just as in FDM, the wavelengths (or optical frequencies) in a DWDM link must be properly spaced to avoid interference between channels. In an optical system this interference may arise from the fact that the center wavelength of laser diode sources and the spectral operating characteristics of other optical components in the link may drift with temperature and time, thereby giving rise to the need for a guard band between wavelength channels.

Since WDM is essentially frequency division multiplexing at optical carrier frequencies, the ITU developed DWDM standards that specify channel spacings in terms of frequency. The ITU-T Recommendation G.694.1, which is entitled 'Dense Wavelength

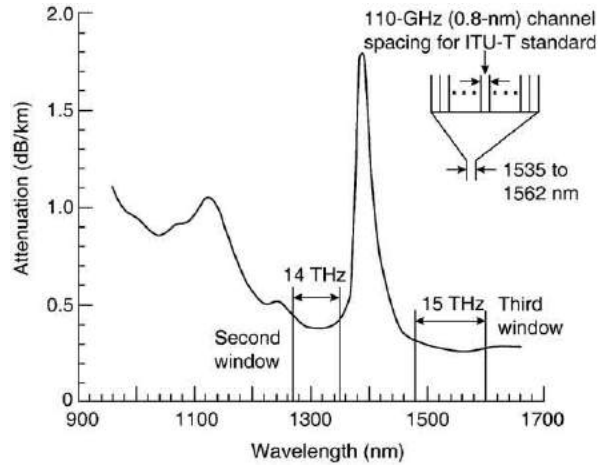


Fig. 5 Wavelength division multiplexing (WDM) concept.

Division Multiplexing (DWDM),' specifies WDM operation in the S-, C-, and L-bands for high-quality, high-rate metro area network (MAN) and wide area network (WAN) services. It calls out for narrow frequency spacings of 100 to 12.5 GHz (or, equivalently, 0.8 to 0.1 nm at 1550 nm). This implementation requires the use of stable, high-quality, temperature-controlled and wavelength-controlled (frequency-locked) laser diode light sources.

With the production of full-spectrum (low-water-content) fibers, the development of relatively inexpensive optical sources, and the desire to have low-cost optical links operating in metro and local area networks, came the concept of coarse WDM (CWDM). In 2002, the ITU-T released a standard aimed specifically at CWDM. This is Recommendation G.694.2, which is entitled 'Coarse Wavelength Division Multiplexing (CWDM).' The CWDM grid is made up of 18 wavelengths defined within the range 1270 nm to 1610 nm (O-through L-bands) spaced by 20 nm with wavelength-drift tolerances of ± 2 nm. This can be achieved with inexpensive light sources that are not temperature-controlled. The targeted transmission distance for CWDM is 50 km on single-mode fibers, such as those specified in ITU-T Recommendations G.652, G.653, and G.655.

Wavelength tunability of a source is an important property of WDM systems. Obviously it is not desirable or practical to maintain an inventory of dozens of lasers that emit at different wavelengths for WDM applications. The ideal tunable laser should be adjustable to emit at a specific wavelength across a broad spectral range. One such device is a distributed Bragg reflector (DBR) laser diode that can be tuned over a 10 to 20 nm spectral range. Work on perfecting such devices are still underway.

Starting in the mid-1990s, a combination of EDFAs and WDM was used to boost fiber information capacity to even higher levels and to increase the transmission distance. A major system consideration in these super-high capacity links is to ensure that there is appropriate link and equipment redundancy, so that alternate paths are available in case of disruptions in communications resulting from cable ruptures (for example, caused by errant digging from a backhoe) or equipment failures at an intermediate node. Such disruptions otherwise could have a devastating effect on a large group of users.

Undersea Optical Cable Systems

The first transoceanic fiber optic cable systems were installed in the Atlantic and Pacific Oceans in 1988 and 1989. Initially these systems operated at 280 Mb/s per fiber pair using 1310-nm lasers and single-mode fibers. The links consisted of a series of point-to-point optical fiber segments between electronic-based undersea regeneration points that were located nominally 60 km apart. Later the transmission capacity of these links was upgraded to 2.5 Gb/s and the regenerator spacing was increased to 100 km by converting the 1310-nm multiple-frequency light sources to 1550-nm single-frequency laser diodes. Later, the regenerator spacing was increased to 140 km.

Although these cable systems significantly improved the quality of the international telephone service, the optical-to-electrical conversion process at each regeneration point remained a capacity bottleneck. The introduction of erbium-doped optical fiber amplifiers (EDFA) eliminated this bottleneck from undersea lightwave systems by amplifying signals directly in the optical domain. Since EDFAs operate over a 30-nm wavelength band, they are well suited for use with undersea WDM links, which have provided a further capacity increase. These undersea optical amplifiers are typically spaced about 45 km apart.

One example of the many worldwide installations of optically amplified WDM networks is the SEA-ME-WE-3 Cable System. This undersea network runs from Germany to Singapore, connecting more than a dozen countries in between. Hence the name SEA-ME-WE, which refers to Southeast Asia (SEA), the Middle East (ME), and Western Europe (WE). The network has two pairs of undersea fibers with a capacity of eight STM-16 wavelengths per fiber.

See also: Basic Concepts of Optical Communication Systems

Further Reading

Bergano, N.S., 1997. Undersea amplified lightwave system design. In: Kaminow, I.P., Koch, T.L. (Eds.), Optical Fiber Telecommunications, vol. IIIA. New York: Academic, pp. 302–335.
Goralski, W.J., 2002. SONET/SDH, 3rd edn. New York: McGraw-Hill.
Hecht, J., 1999. City of Light. New York: Oxford University Press.
Joyner, C.H., 1997. Semiconductor laser growth and fabrication technology. In: Kaminow, I.P., Koch, T.L. (Eds.), Optical Fiber Telecommunications, vol. IIIB. New York: Academic, pp. 163–199.
Kaiser, P., Keck, D.B., 1988. Fiber types and their status. In: Miller, S.E., Kaminow, I.P. (Eds.), Optical Fiber Telecommunications II. New York: Academic, pp. 29–54.
Kao, K.C., Hockman, G.A., 1966. Dielectric-fibre surface waveguides for optical frequencies. In: Proceedings IEE, vol. 113. . pp. 1151–1158.
Keiser, G., 2000. Optical Fiber Communications, 3rd edn. Burr Ridge, IL: McGraw-Hill.
Keiser, G., 2003. Optical Communications Essentials. New York: McGraw-Hill.
Mollenauer, L.F., Gordon, J.P., Mamyshev, P.V., 1997. Solitons in high bit rate long-distance transmission. In: Kaminow, I.P., Koch, T.L. (Eds.), Optical Fiber Telecommunications, vol. IIIA. New York: Academic, pp. 373–460.
Nyman, B., Farries, M., Si, C., 2001. Technology trends in dense WDM demultiplexers. Optical Fiber Technology 7 (4), 255–274.
Ramaswami, R., Sivarajan, K.N., 2002. Optical Networks, 2nd edn. San Francisco, CA: Morgan Kaufmann.
Snitzer, E., 1961. Cylindrical dielectric waveguide modes. Journal Optical Society of America 51, 491–498.
Trischitta, P.R., Marra, W.C., 1998. Applying WDM technology to undersea cable networks. IEEE Communication Magazine 36 (2), 62–66.

Dispersion Management

AE Willner, Y-W Song, J McGeehan, and Z Pan, University of Southern California, Los Angeles, CA, USA
B Hoanca, University of Alaska Anchorage, Anchorage, AK, USA

Glossary

Chromatic dispersion The time-domain pulse broadening in an optical fiber caused by the frequency dependence of the refractive index. This results in photons at different frequencies traveling at different speeds [ps/nm/km].

Chromatic dispersion slope Different wavelengths have different amounts of dispersion values in an optical fiber [ps/nm^2/km].

Conventional single mode fiber (SMF) Fiber that transmits only a single optical mode by virtue of a very small core diameter relative to the cladding. Provides lowest loss in the 1,550 nm region and has zero dispersion at 1,300 nm.

Cross-phase modulation (XPM) A nonlinear Kerr effect in which a signal undergoes a nonlinear phase shift induced by a copropagating signal at a different wavelength in a WDM system.

Dispersion compensating fiber (DCF) Optical fiber that has both a large negative dispersion and dispersion slope around 1,550 nm.

Fiber Bragg gratings (FBGs) A small section of optical fiber in which there is a periodic change of refractive index along the core of the fiber. An FBG acts as a wavelength-selective mirror, reflecting only specific wavelengths (Bragg wavelengths) and passing others.

Four-wave mixing (FWM) A nonlinear Kerr effect in which two or more signal wavelengths interact to generate a new wavelength.

Kerr effect Nonlinear effect in which the refractive index of optical fiber varies as a function of the intensity of light within the fiber core.

Modulation The process of encoding digital data onto an optical signal so it can be transmitted through an optical network.

Nonlinear effects Describes the nonlinear responses of a dielectric to intense electromagnetic fields. One such effect is the intensity dependence of the refractive index of an optical fiber (Kerr effect).

Nonzero dispersion-shifted fiber (NZDSF) Optical fiber that has a small amount of dispersion in the 1,550 nm region in order to reduce deleterious nonlinear effects.

Polarization mode dispersion (PMD) Dispersion resulting from the fact that different light polarizations within the fiber core will travel at different speeds through the optical fiber [ps/km$^{0.5}$].

Self-phase modulation (SPM) A nonlinear Kerr effect in which a signal undergoes a self-induced phase shift during propagation in optical fiber.

Wavelength division multiplexing (WDM) Transmitting many different wavelengths down the same optical fiber at the same time in order to increase the amount of information that can be carried.

Introduction

Optical communications systems have grown explosively in terms of the capacity that can be transmitted over a single optical fiber. This trend has been fueled by two complementary techniques, those being the increase in data-rate-per-channel coupled with the increase in the total number of parallel wavelength channels. However, there are many considerations as to the total number of wavelength-division-multiplexed (WDM) channels that can be accommodated in a system, including cost, information spectral efficiency, nonlinear effects, and component wavelength selectivity.

Dispersion is one of the critical roadblocks to increasing the transmission capacity of optical fiber. The dispersive effect in an optical fiber has several ingredients, including intermodal dispersion in a multimode fiber, waveguide dispersion, material dispersion, and chromatic dispersion. In particular, chromatic dispersion is one of the critical effects in a single mode fiber (SMF), resulting in a temporal spreading of an optical bit as it propagates along the fiber. At data rates ≤ 2.5 Gbit/s, the effects of chromatic dispersion are not particularly troublesome. For data rates ≥ 10 Gbit/s, however, transmission can be tricky and the chromatic dispersion-induced degrading effects must be dealt with in some way, perhaps by compensation. Furthermore, the effects of chromatic dispersion rise quite rapidly as the bit rate increases – when the bit rate increases by a factor of four, the effects of chromatic dispersion increase by a factor of 16! This article will only deal with the management of chromatic dispersion in single mode fiber.

One of the critical limitations of optical fiber communications comes from chromatic dispersion, which results in a pulse broadening as it propagates along the fiber. This occurs as photons of different frequencies (created by the spreading effect of data modulation) travel at different speeds, due to the frequency-dependent refractive index of the fiber core. Compounding the problems cause by chromatic dispersion is the fact that as bit rates rise, chromatic dispersion effects rise quadratically with respect to the increase in the bit rate. One can eliminate these effects using fiber with zero chromatic dispersion, known as dispersion shifted fiber (DSF). However, with zero dispersion, all channels in a wavelength-division-multiplexed (WDM) system travel at the

 doi:10.1016/B978-0-12-809283-5.00675-3

same speed, in-phase, and a number of deleterious nonlinear effects such as cross-phase modulation (XPM) and four-wave mixing (FWM) result. Thus, in WDM systems, some amount of chromatic dispersion is necessary to keep channels out-of-phase, and as such chromatic dispersion compensation is required. Any real fiber link may also suffer from 'dispersion slope' effects, in which a slightly different dispersion value is produced in each WDM channel. This means that while one may be able to compensate one channel exactly, other channels may progressively accumulate increasing amounts of dispersion, which can severely limit the ultimate length of the optical link and the wavelength range that can be used in a WDM system. This article will address the concepts of chromatic dispersion and dispersion slope management followed by some examples highlighting the need for tunability to enable robust optical WDM systems in dynamic environments. Some dispersion monitoring techniques are then discussed and examples given.

Chromatic Dispersion in Optical Fiber Communication Systems

In any medium (other than vacuum) and in any waveguide structure (other than ideal infinite free space), different electromagnetic frequencies travel at different speeds. This is the essence of chromatic dispersion. As the real fiber-optic world is rather distant from the ideal concepts of both vacuum and infinite free space, dispersion will always be a concern when one is dealing with the propagation of electromagnetic radiation through fiber. The velocity in fiber of a single monochromatic wavelength is constant. However, data modulation causes a broadening of the spectrum of even the most monochromatic laser pulse. Thus, all modulated data have a nonzero spectral width which spans several wavelengths, and the different spectral components of modulated data travel at different speeds. In particular, for digital data intensity modulated on an optical carrier, chromatic dispersion leads to pulse broadening – which in turn leads to chromatic dispersion limiting the maximum data rate that can be transmitted through optical fiber (see Fig. 1).

Considering that the chromatic dispersion in optical fibers is due to the frequency-dependent nature of the propagation characteristics, for both the material (the refractive index of glass) and the waveguide structure, the speed of light of a particular wavelength λ will be expressed as follows, using a Taylor series expansion of the value of the refractive index as a function of the wavelength:

$$v(\lambda) = \frac{c_0}{n(\lambda)} = \frac{c_0}{n_0(\lambda_0) + \frac{\partial n}{\partial \lambda}\delta\lambda + \frac{\partial^2 n}{\partial \lambda^2}(\delta\lambda)^2} \tag{1}$$

Here, c_0 is the speed of light in vacuum, λ_0 is a reference wavelength, and the terms in $\partial n/\partial\lambda$ and $\partial^2 n/\partial\lambda^2$ are associated with the chromatic dispersion and the dispersion slope (i.e., the variation of the chromatic dispersion with wavelength), respectively. Transmission fiber has positive dispersion, i.e. longer wavelengths result in longer propagation delays. The units of chromatic dispersion are picoseconds per nanometer per kilometer, meaning that shorter time pulses, wider frequency spread due to data modulation, and longer fiber lengths will each contribute linearly to temporal dispersion. Higher data rates inherently have both shorter pulses and wider frequency spreads. Therefore, as network speed increases, the impact of chromatic dispersion rises precipitously as the square of the increase in data rate. The quadratic increase with the data rate is a result of two effects, each with a linear contribution. On one hand, a doubling of the data rate makes the spectrum twice as wide, doubling the effect of dispersion. On the other hand, the same doubling of the data rate makes the data pulses only half as long (hence twice as sensitive to dispersion). The combination of a wider signal spectrum and a shorter pulse width is what leads to the overall quadratic impact. Moreover, the data modulation format used can significantly affect the sensitivity of a system to chromatic dispersion. For example, the common nonreturn-to-zero (NRZ) data

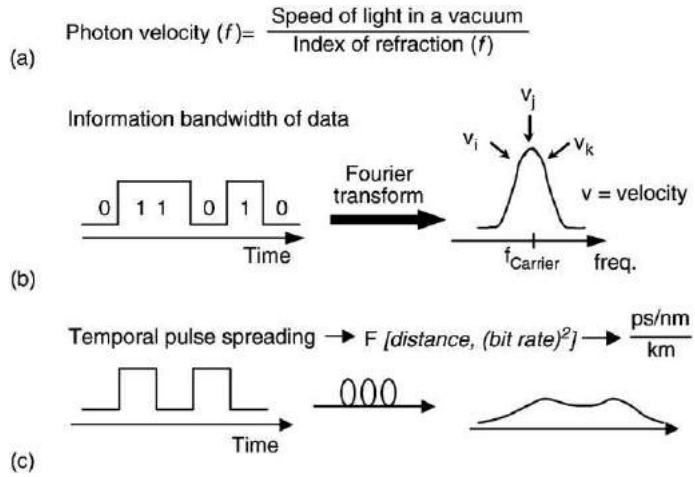


Fig. 1 The origin of chromatic dispersion in data transmission. (a) Chromatic dispersion is caused by the frequency-dependent refractive index in fiber. (b) The nonzero spectral width due to data modulation. (c) Dispersion leads to pulse broadening, proportional to the transmission distance and the data rate.

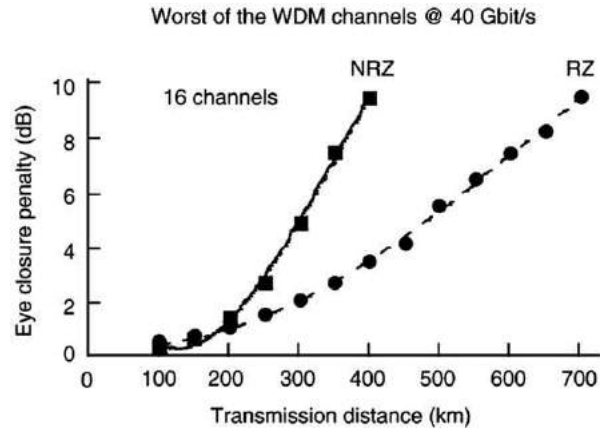


Fig. 2 Performances of RZ and NRZ formats in a real fiber transmission link. Reproduced with permission from Hayee I and Willner AE (1999) NRZ versus RZ in 10–40-Gb/s dispersion-managed WDM transmission systems. *IEEE Photon. Tech. Lett.* 11(8): 991–993.

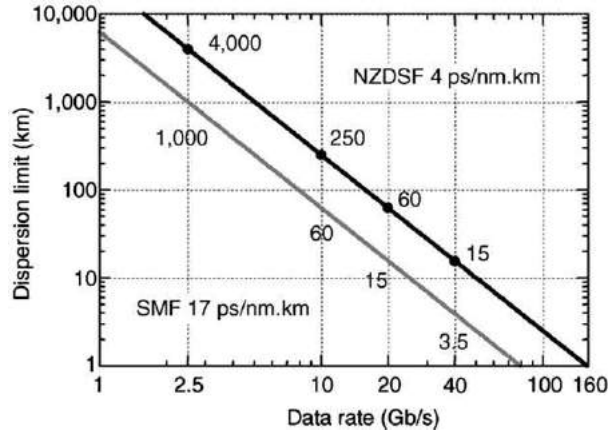


Fig. 3 Transmission distance limitations due to uncompensated dispersion in SMF as a function of data rate for intensity modulated optical signals. Reproduced with permission from Garrett LD (2001) Invited Short Course, *Optical Fiber Communication Conference*.

format, in which the optical power stays high throughout the entire time slot of a '1' bit, is more robust to chromatic dispersion than is the return-to-zero (RZ) format, in which the optical power stays high in only part of the time slot of a '1' bit. This difference is due to the fact that RZ data have a much wider channel frequency spectrum compared to NRZ data, thus incurring more chromatic dispersion. However, in a real WDM system, the RZ format increases the maximum allowable transmission distance by virtue of its reduced duty cycle (compared to the NRZ format), making it less susceptible to fiber nonlinearities as can be seen in **Fig. 2**.

A rule for the maximum distance over which data can be transmitted is to consider a broadening of the pulse equal to the bit period. For a bit period B, a dispersion value D and a spectral width $\Delta\lambda$, the dispersion-limited distance is given by

$$L_{\mathrm{D}} = \frac{1}{D \cdot B \cdot \Delta\lambda} = \frac{1}{D \cdot B \cdot (cB)} \propto \frac{1}{B^2} \tag{2}$$

(see **Fig. 3**). For example, for single mode fiber, $D = 17$ ps/nm/km, so for 10 Gbit/s data the distance is $L_{\mathrm{D}} = 52$ km. In fact, a more exact calculation shows that for 60 km, the dispersion induced power penalty is less than 1 dB (see **Fig. 4**). The power penalty for uncompensated dispersion rises exponentially with transmission distance, and thus to maintain good signal quality, dispersion compensation is required.

Chromatic Dispersion Management

Optical Nonlinearities as Factors to be Considered in Dispersion Compensation

Even though it is possible to manufacture fiber with zero dispersion, it is not practical to use such fiber for WDM transmission, due to large penalties induced by fiber nonlinearities. Most nonlinear effects originate from the nonlinear refractive index of fiber, which is not only dependent on the frequency of light but also on the intensity (optical power), and is related to the optical

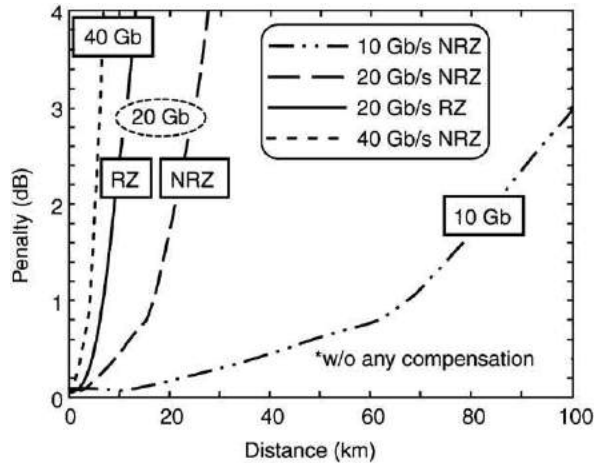


Fig. 4 Power penalties due to uncompensated dispersion in single mode fiber (SMF) as a function of distance and data rate. Reproduced with permission from Garrett LD (2001) Invited Short Course, *Optical Fiber Communication Conference*.

power as:

$$\bar{n}(f,P) = n(f) + n_2 \frac{P}{A_{\text{eff}}} \tag{3}$$

where $n(f)$ is the linear part of the refractive index, P is the optical power inside the fiber, and n_2 is the nonlinear-index coefficient for silica fibers. The typical value of n_2 is 2.6×10^{-20} m^2/W. This number takes into account the averaging of the polarization states of the light as it travels in the fiber. The intensity dependence of the refractive index gives rise to three major nonlinear effects.

Self-phase modulation (SPM)

A million photons 'see' a different glass than does a single photon, and a photon traveling along with many other photons will slow down. SPM occurs because of the varying intensity profile of an optical pulse on a single WDM channel. This intensity profile causes a refractive index profile and, thus, a photon speed differential. The resulting phase change for light propagating in an optical fiber is expressed as:

$$\Phi_{\text{NL}} = \gamma P L_{\text{eff}} \tag{4}$$

where the quantities γ and L_{eff} are defined as:

$$\gamma = \frac{2\pi n_2}{\lambda A_{\text{eff}}} \quad \text{and} \quad L_{\text{eff}} = \frac{1 - e^{-\alpha L}}{\alpha} \tag{5}$$

where A_{eff} is the effective mode area of the fiber and α is the fiber attenuation loss. L_{eff} is the effective nonlinear length of the fiber that accounts for fiber loss, and γ is the nonlinear coefficient measured in rad/km/W. A typical range of values for γ is between 10–30 rad/km/W. Although the nonlinear coefficient is small, the long transmission lengths and high optical powers, that have been made possible by the use of optical amplifiers, can cause a large enough nonlinear phase change to play a significant role in state-of-the-art lightwave systems.

Cross-phase modulation (XPM)

When considering many WDM channels co-propagating in a fiber, photons from channels 2 through N can distort the index profile that is experienced by channel 1. The photons from the other channels 'chirp' the signal frequencies on channel 1, which will interact with fiber chromatic dispersion and cause temporal distortion. This effect is called cross-phase modulation. In a two-channel system, the frequency chirp in channel 1, due to power fluctuation within both channels, is given by

$$\Delta B = \frac{\mathrm{d}\Phi_{\text{NL}}}{\mathrm{d}t} = \gamma L_{\text{eff}} \frac{\mathrm{d}P_1}{\mathrm{d}t} + 2\gamma L_{\text{eff}} \frac{\mathrm{d}P_2}{\mathrm{d}t} \tag{6}$$

where, $\mathrm{d}P_1/\mathrm{d}t$ and $\mathrm{d}P_2/\mathrm{d}t$ are the time derivatives of the pulse powers of channels 1 and 2, respectively. The first term on the right-hand side of the above equation is due to SPM, and the second term is due to XPM. Note that the XPM-induced chirp term is double that of the SPM-induced chirp term. As such, XPM can impose a much greater limitation on WDM systems than can SPM, especially in systems with many WDM channels.

Four-wave-mixing (FWM)

The optical intensity propagating through the fiber is related to the electric field intensity squared. In a WDM system, the total electric field is the sum of the electric fields of each individual channel. When squaring the sum of different fields, products emerge that are beat terms at various sum and difference frequencies to the original signals. Fig. 5 depicts that if a WDM channel exists at

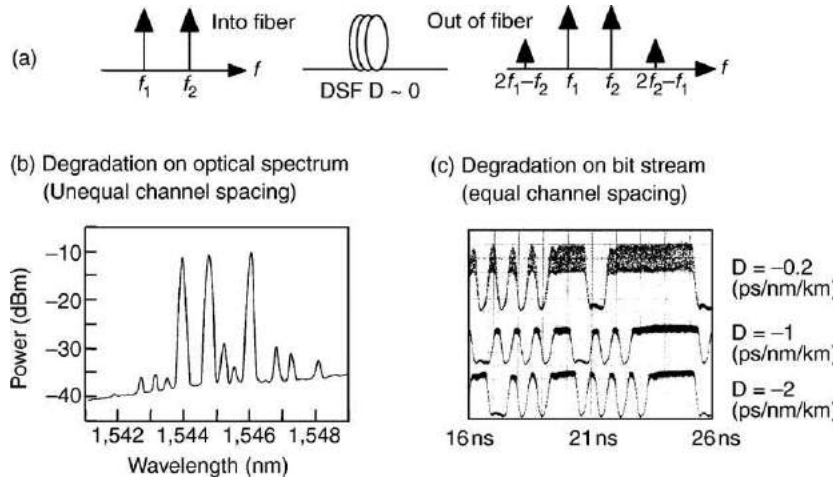


Fig. 5 (a) and (b) FWM induces new spectral components via nonlinear mixing of two wavelength signals. (c) The signal degradation due to FWM products falling on a third data channel can be reduced by even small amounts of dispersion. Reproduced with permission from Tkach RW, Chraplyvy AR, Forghieri F, Gnauck AH and Derosier RM (1995) Four-photon mixing and high-speed WDM systems. *Journal of Photon Technology* 13(5): 841–849.

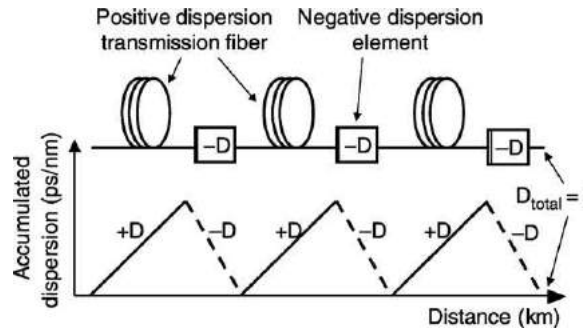


Fig. 6 Dispersion map of a basic dispersion managed system. Positive dispersion transmission fiber alternates with negative dispersion compensation elements such that the total dispersion is zero end-to-end.

Table 1 Commercially available fibers and their characteristics

	Dispersion @ 1,550 nm [ps/nm/km]	Dispersion slope [ps/nm^2/km]	Attenuation [dB]	Mode field diameter [μm]	PMD [ps/km$^{0.5}$]
SMF	18	0.08	≤0.25	11	≤0.5
Tera Light	8.0	0.058	≤0.2	9.2	≤0.04
TW-RS	4.4	0.043	≤0.25	8.4	≤0.03
LEAF	4.0	0.085	≤0.25	9.6	≤0.08
Standard DCF	−90	−0.22	≤0.5	5.2	≤0.08

one of the four-wave-mixing beat-term frequencies, then the beat term will interfere coherently with this other WDM channel and potentially destroy the data.

Dispersion Maps

While zero-dispersion fiber is not a good idea, a large value of the accumulated dispersion at the end of a fiber link is also undesirable. An ideal solution is to have a 'dispersion map,' alternating sections of positive and negative dispersion as can be seen in Fig. 6. This is a very powerful concept: at each point along the fiber the dispersion has some nonzero value, eliminating FWM and XPM, but the total dispersion at the end of the fiber link is zero, so that no pulse broadening is induced (Table 1). The most advanced systems require periodic dispersion compensation, as well as pre- and post-compensation (before and after the transmission fiber).

The addition of negative dispersion to a standard fiber link has been traditionally known as 'dispersion compensation,' however, the term 'dispersion management' is more appropriate. SMF has positive dispersion, but some new varieties of nonzero

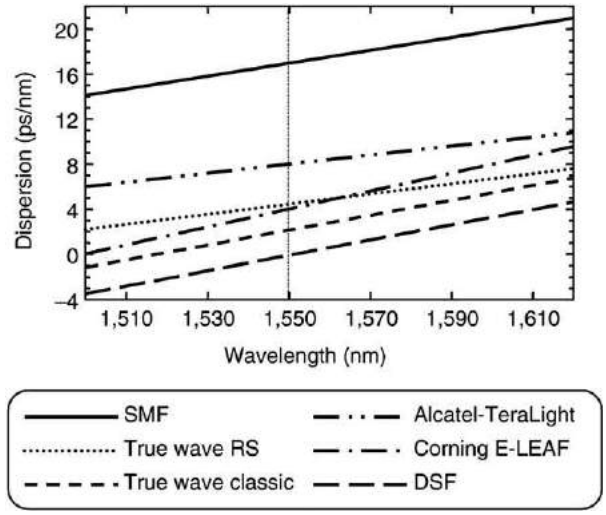


Fig. 7 Chromatic dispersion characteristics of various commercially available types of transmission fiber.

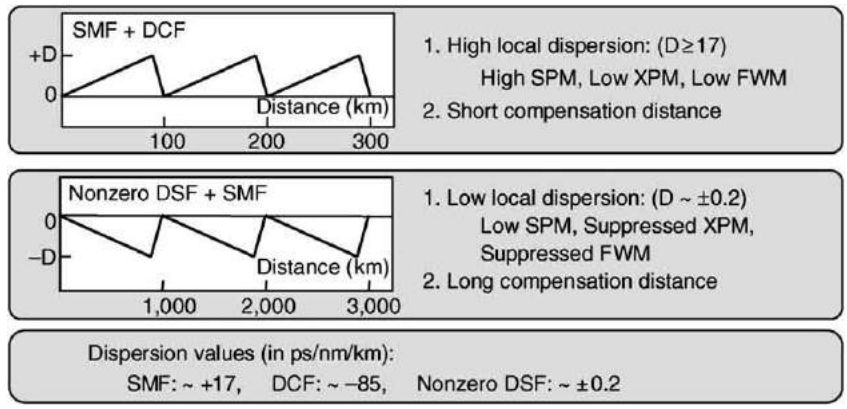


Fig. 8 Various dispersion maps for SMF-DCF and NZDSF-SMF.

dispersion-shifted fiber (NZDSF) come in both positive and negative dispersion varieties. Some examples are shown in Fig. 7. Reverse dispersion fiber is also now available, with a large dispersion comparable to that of SMF, but with the opposite sign. When such flexibility is available in choosing both the magnitude and sign of the dispersion of the fiber in a link, dispersion-managed systems can be fully optimized to the desired dispersion map using a combination of fiber and dispersion compensation devices (see Fig. 8).

Dispersion is a linear process, so first-order dispersion maps can be understood as linear systems. However, the effects of nonlinearities cannot be ignored, especially in WDM systems, with many tens of channels, where the launch power may be very high. In particular, in systems deploying dispersion compensating fiber (DCF), the large nonlinear coefficient of the DCF can dramatically affect the dispersion map.

Corrections to Linear Dispersion Maps

Chromatic dispersion is a necessity in WDM systems, to minimize the effects of fiber nonlinearities. A chromatic dispersion value as small as a few ps/nm/km is usually sufficient to make XPM and FWM negligible. To mitigate the effects of nonlinearities but maintain small amounts of chromatic dispersion, NZDSF is commercially available. Due to these nonlinear effects, chromatic dispersion must be managed, rather than eliminated.

If a dispersion-management system was perfectly linear, it would be irrelevant whether the dispersion along a path is small or large, as long as the overall dispersion is compensated to zero (end to end). Thus, in a linear system the performance should be similar, regardless of whether the transmission fiber is SMF, and dispersion compensation modules are deployed every 60 km, or the transmission fiber is NZDSF (with approximately a quarter of the dispersion value of SMF) and dispersion compensation modules are deployed every 240 km. In real life, optical nonlinearities are very important, and recent results seem to favor the use of large, SMF-like, dispersion values in the transmission path and correspondingly high dispersion compensation devices. A recent

study of performance versus channel spacing showed that the capacity of SMF could be more than four times that of NZDSF. This is because the nonlinear coefficients are much higher in NZDSF than in SMF, and for dense WDM the channel interactions become a limiting factor. A critical conclusion is that not all dispersion compensation maps are created equal: a simple calculation of the dispersion compensation, to cancel the overall dispersion value, does not lead to optimal dispersion map designs.

Additionally, several solutions have been shown to be either resistant to dispersion, or have been shown to rely on dispersion itself for transmission. Such solutions include chirped pulses (where prechirping emphasizes the spectrum of the pulses so that dispersion does not broaden them too much), dispersion assisted transmission (where an initial phase modulation tailored to the transmission distance leads to full-scale amplitude modulation at the receiver end due to the dispersion), and various modulation formats robust to chromatic dispersion and nonlinearities.

Dispersion Management Solutions

Fixed Dispersion Compensation

From a systems point of view, there are several requirements for a dispersion compensating module: low loss, low optical nonlinearity, broadband (or multichannel) operation, small footprint, low weight, low power consumption, and clearly low cost. It is unfortunate that the first dispersion compensation modules, based on DCF only, met two of these requirements: broadband operation and low power consumption. On the other hand, several solutions have emerged that can complement or even replace these first-generation compensators.

Dispersion compensating fiber (DCF)

One of the first dispersion compensation techniques was to deploy specially designed sections of fiber with negative chromatic dispersion. The technology for DCF emerged in the 1980s and has developed dramatically since the advent of optical amplifiers in 1990. DCF is the most widely deployed dispersion compensator, providing broadband operation and stable dispersion characteristics, and the lack of a dynamic, tunable DCF solution has not reduced its popularity.

As can be seen in Fig. 9, the core of the average dispersion compensating fiber is much smaller than that of standard SMF, and beams with longer wavelengths experience relatively large changes in mode size (due to the waveguide structure) leading to greater propagation through the cladding of the fiber, where the speed of light is greater than that of the core. This leads to a large negative dispersion value. Additional cladding layers can lead to improved DCF designs that can include negative dispersion slope to counteract the positive dispersion slope of standard SMF.

In spite of its many advantages, DCF has a number of drawbacks. First, it is limited to a fixed compensation value. In addition, DCF has a weakly guiding structure and has a much smaller core cross-section, 19 μm^2, compared to the 85 μm^2 of SMF. This leads to higher nonlinearity, higher splice losses, as well as higher bending losses. Last, the length of DCF required to compensate for SMF dispersion is rather long, about one-fifth of the length of the transmission fiber for which it is compensating. Thus DCF modules induce loss, and are relatively bulky and heavy. The bulk is partly due to the mass of fiber, but also due to the resin used to hold the fiber securely in place. One other contribution to the size of the module is the higher bend loss associated with the refractive index profile of DCF; this limits the radius of the DCF loop to 6–8 inches, compared to the minimum bend radius of 2 inches for SMF.

Traditionally, DCF-based dispersion compensation modules are usually located at amplifier sites. This serves several purposes. First, amplifier sites offer relatively easy access to the fiber, without requiring any digging or unbraiding of the cable. Second, DCF

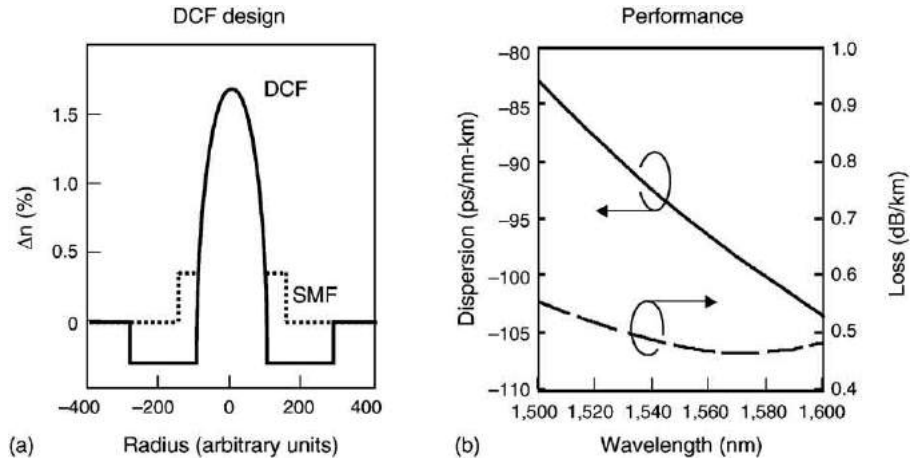


Fig. 9 Typical DCF (a) refractive index profile and (b) dispersion and loss as a function of wavelength. Δn is defined as refractive index variation relative to the cladding.

has high loss (usually at least double that of standard SMF), so a gain stage is required before the DCF module to avoid excessively low signal levels. DCF has a cross-section four times smaller then SMF, hence a higher nonlinearity, which limits the maximum launch power into a DCF module. The compromise is to place the DCF in the mid-section of a two-section EDFA. This way, the first stage provides pre-DCF gain, but not to a power level that would generate excessive nonlinear effects in the DCF. The second stage amplifies the dispersion compensated signal to a power level suitable for transmission though the fiber link. This launch power level is typically much higher than could be transmitted through DCF without generating large nonlinear effects. Many newer dispersion compensation devices have better performance than DCF, in particular lower loss and lower nonlinearities. For this reason, they may not have to be deployed at the mid-section of an amplifier. **Fig. 10** shows the real demonstration results using the DCF.

Chirped fiber Bragg gratings

Fiber Bragg gratings have emerged as major components for dispersion compensation because of their low loss, small footprint, and low optical nonlinearity. Bragg gratings are sections of single-mode fiber in which the refractive index of the core is modulated in a periodic fashion, as a function of the spatial coordinate along the length of the fiber. When the spatial periodicity of the modulation matches what is known as a Bragg condition with respect to the wavelength of light propagating through the grating, the periodic structure acts like a mirror, reflecting the optical radiation that is traveling through the core of the fiber. An optical circulator is traditionally used to separate the reflected output beam from the input beam.

When the periodicity of the grating is varied along its length, the result is a chirped grating which can be used to compensate for chromatic dispersion. The chirp is understood as the rate of change of the spatial frequency as a function of position along the grating. In chirped gratings the Bragg matching condition for different wavelengths occurs at different positions along the grating length. Thus, the roundtrip delay of each wavelength can be tailored by designing the chirp profile appropriately. **Fig. 11** compares the chirped FBG with uniform FBG. In a data pulse that has been distorted by dispersion, different frequency components arrive

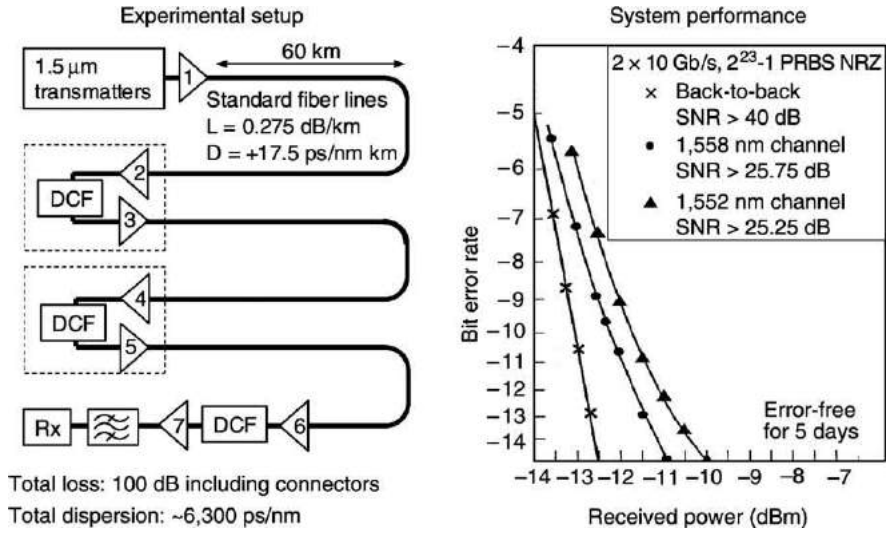


Fig. 10 System demonstration of dispersion compensation using DCF. Reproduced with permission from Park YK, Yeates PD, Delavaux J-MP, *et al.* (1995) A field demonstration of 20-Gb/s capacity transmission over 360 km of installed standard (non-DSF) fiber. *Photon. Technol. Lett.* 7 (7): 816–818.

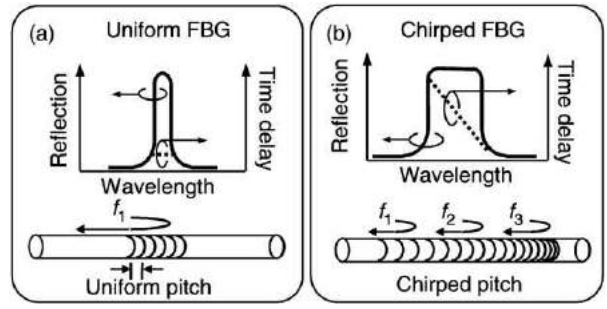


Fig. 11 Uniform and chirped FBGs. (a) A grating with uniform pitch has a narrow reflection spectrum and a flat time delay as a function of wavelength. (b) A chirped FBG has a wider bandwidth, a varying time delay and a longer grating length.

with different amounts of relative delay. By tailoring the chirp profile such that the frequency components see a relative delay which is the inverse of the delay of the transmission fiber, the pulse can be compressed back. The dispersion of the grating is the slope of the time delay as a function of wavelength, which is related to the chirp.

The main drawback of Bragg gratings is that the amplitude profile and the phase profile as a function of wavelength have some amount of ripple. Ideally, the amplitude profile of the grating should have a flat (or rounded) top in the passband, and the phase profile should be linear (for linearly chirped gratings) or polynomial (for nonlinearly chirped gratings). The grating ripple is the deviation from the ideal profile shape. Considerable effort has been expended on reducing the ripple. While early gratings were plagued by more than 100 ps of ripple, published results have shown vast improvement to values close to ± 3 ps.

Higher order mode dispersion compensation fiber

One of the challenges of designing standard DCF is that high negative dispersion is hard to achieve unless the cross-section of the fiber is small (which leads to high nonlinearity and high loss). One way to reduce both the loss and the nonlinearity is to use a higher-order mode (HOM) fiber (LP_{11} or LP_{02} near cutoff instead of the LP_{01} mode in the transmission fiber).

Such a device requires a good-quality mode converter between LP_{01} and LP_{02} to interface between the SMF and HOM fiber. HOM fiber has a dispersion per unit length greater than six times that of DCF. Thus, to compensate for a given transmission length in SMF, the length of HOM fiber required is only one sixth the length of DCF. Thus, even though losses and nonlinearity per unit length are larger for HOM fiber than for DCF, they are smaller overall, because of the shorter HOM fiber length. As an added bonus, the dispersion can be tuned slightly by changing the cutoff wavelength of LP_{02} (via temperature tuning). A soon to be released HOM fiber-based commercial dispersion compensation module is not tunable, but can fully compensate for dispersion slope.

Tunable Dispersion Compensation

The need for tunability

In a perfect world, all fiber links would have a known, discrete, and unchanging value of chromatic dispersion. Network operators would then deploy fixed dispersion compensators periodically along every fiber link to exactly match the fiber dispersion. Unfortunately, several vexing issues may necessitate that dispersion compensators are tunability, that they have the ability to adjust the amount of dispersion to match system requirements.

First, there is the most basic business issue of inventory management. Network operators typically do not know the exact length of a deployed fiber link nor its chromatic dispersion value. Moreover, fiber plants periodically undergo upgrades and maintenance, leaving new and nonexact lengths of fiber behind. Therefore, operators would need to keep in stock a large number of different compensator models, and even then the compensation would only be approximate. Second, we must consider the sheer difficulty of 40 Gbit/s signals. The tolerable threshold for accumulated dispersion for a 40 Gbit/s data channel is 16 times smaller than at 10 Gbit/s. If the compensation value does not exactly match the fiber to within a few percent of the required dispersion value, then the communication link will not work. Tunability is considered a key enabler for this bit rate (see **Figs. 12** and **13**). Third, the accumulated dispersion changes slightly with temperature, which begins to be an issue for 40 Gbit/s systems and 10 Gbit/s ultra long-haul systems. In fiber, the zero-dispersion wavelength changes with temperature at a typical rate of 0.03 nm/°C. It can been shown that a not-uncommon 50 °C variation along a 1,000 km 40 Gbit/s link can produce significant degradation (see **Fig. 14**). Fourth, we are experiencing the dawn of reconfigurable optical networking. In such systems, the network path, and therefore the accumulated fiber dispersion, can change. It is important to note that even if the fiber spans are compensated span-by-span, the pervasive use of compensation at the transmitter and receiver suggests that optimization and tunability based on path will still be needed.

Other issues that increase the need for tunability include: (i) laser and (de)mux wavelength drifts for which a data channel no longer resides on the flat-top portion of a filter, thereby producing a chirp on the signal that interacts with the fiber's chromatic dispersion; (ii) changes in signal power that change both the link's nonlinearity and the optimal system dispersion map; and (iii) small differences that exist in transmitter-induced signal chirp.

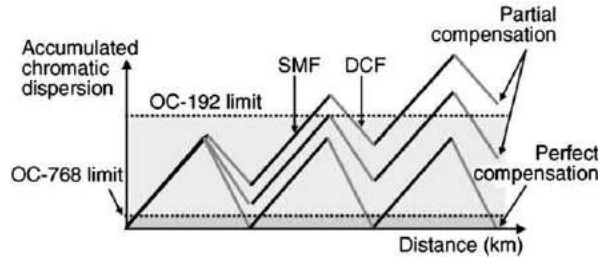


Fig. 12 The need for tunability. The tolerance of OC-768 systems to chromatic dispersion is 16 times lower than that of OC-192 systems. Approximate compensation by fixed in-line dispersion compensators for a single channel may lead to rapid accumulation of unacceptable levels of residual chromatic dispersion.

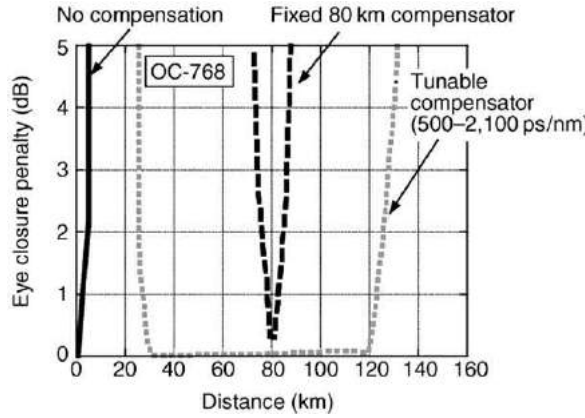


Fig. 13 Tunable dispersion compensation at OC-768 (40 Gb/s) is essential for achieving a comfortable range of acceptable transmission distances (80 km for tunable, only ~4 km for fixed compensation).

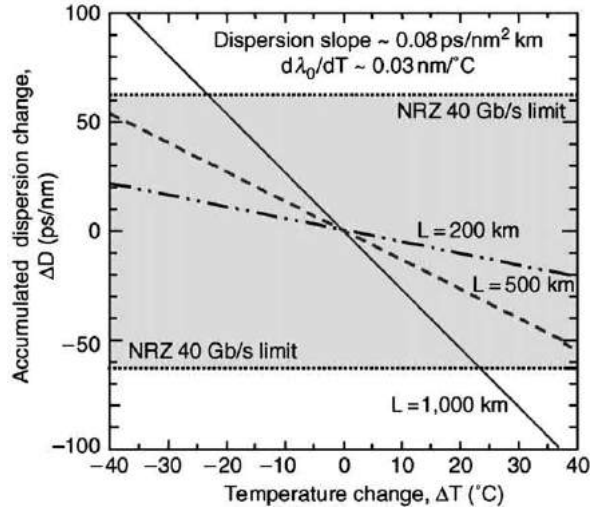


Fig. 14 Accumulated dispersion changes as a function of the link length and temperature fluctuation along the fiber link.

Approaches to tunable dispersion compensation

A host of techniques for tunable dispersion compensation have been proposed in recent years. Some of these ideas are just interesting research ideas, but several have strong potential to become viable technologies.

Fiber gratings offer the inherent advantages of fiber compatibility, low loss, and low cost. If a FBG has a refractive-index periodicity that varies nonlinearly along the length of the fiber, it will produce a time delay that also varies nonlinearly with wavelength (see **Fig. 15**). Herein lies the key to tunability. When a linearly chirped grating is stretched uniformly by a single mechanical element, the time delay curve is shifted towards longer wavelengths, but the slope of the ps-vs.-nm curve remains constant at all wavelengths within the passband. When a nonlinearly-chirped grating is stretched, the time delay curve is shifted toward longer wavelengths, but the slope of the ps-vs.-nm curve at a specific channel wavelength changes continuously. Ultimately, tunable dispersion compensators should accommodate multichannel operation. Several WDM channels can be accommodated by a single chirped FBG in one of two ways: fabricating a much longer (i.e., meters-length) grating, or using a sampling function when writing the grating, thereby creating many replicas of transfer function of the FBG in the wavelength domain (see **Fig. 16**).

One free-space-based tunable dispersion compensation device is the virtually imaged phased array (VIPA), based on the dispersion of a Fabry–Perot interferometer. The design requires several lenses, a movable mirror (for tunability), and a glass plate with a thin film layer of tapered reflectivity for good mode matching. Light incident on the glass plate undergoes several reflections inside the plate. As a result, the beam is imaged at several virtual locations, with a spatial distribution that is wavelength-dependent.

Several devices used for dispersion compensation can be integrated on a chip, using either an optical chip media (semiconductor-based laser or amplifier medium) or an electronic chip. One such technology is the micro-ring resonator, a device that, when used in a structure similar to that of an all-pass filter (see **Fig. 17**), can be used for dispersion compensation on a chip-scale. Although these technologies are not yet ready for deployment as dispersion compensators, they have been used in other applications and have the potential to offer very high performance at low cost.

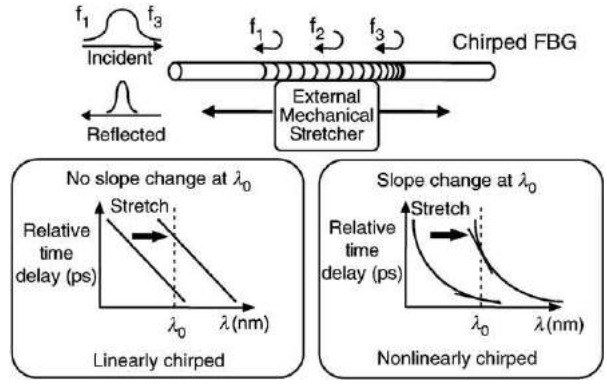


Fig. 15 Tuning results for both linearly and nonlinearly chirped FBGs using uniform stretching elements. The slope of the dispersion curve at a given wavelength λ_0 is constant when the linearly chirped grating is stretched, but changes as the nonlinearly chirped grating is stretched.

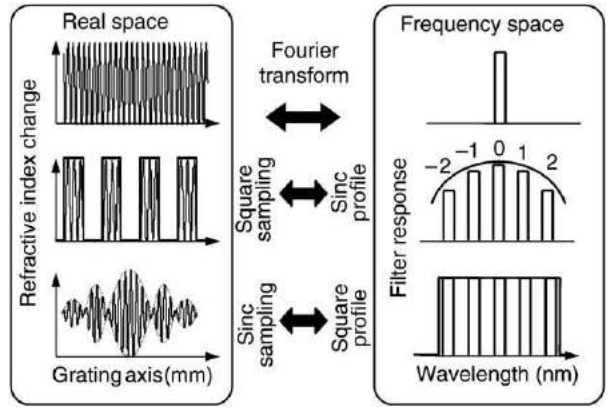


Fig. 16 The concept of 'sampled' FBGs, where a superstructure is written on top of the grating that produces a Fourier transform in the frequency domain, leading to multiple grating passbands.

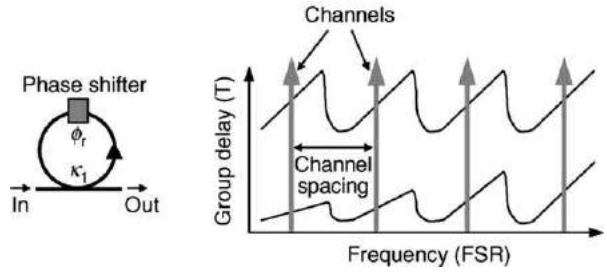


Fig. 17 Architecture of an all-pass filter structure for chromatic dispersion and slope compensation. Reproduced with permission from Madsen CK, Lenz G, Bruce AJ, *et al.* (1999) Integrated all-pass filters for tunable dispersion and dispersion slope compensation. *Photon. Technol. Lett.* 11 (12): 1623–1625.

As the ultimate optical dispersion compensation devices, photonic bandgap fibers (holey fibers) are an interesting class in themselves (see **Fig. 18**). These are fibers with a hollow structure, with holes engineered to achieve a particular functionality. Instead of being drawn from a solid preform, holey fibers are drawn from a group of capillary tubes fused together. This way, the dispersion, dispersion slope, the nonlinear coefficients could in principle all be precisely designed and controlled, up to very small or very large values, well outside the range of those of the solid fiber.

Dispersion Slope Mismatch

Transmission fiber, especially fiber with dispersion compensation built in, may suffer from a dispersion slope in which a slightly different dispersion value is produced for each WDM channel (see **Fig. 19**). Even though the compensator would be able to cancel

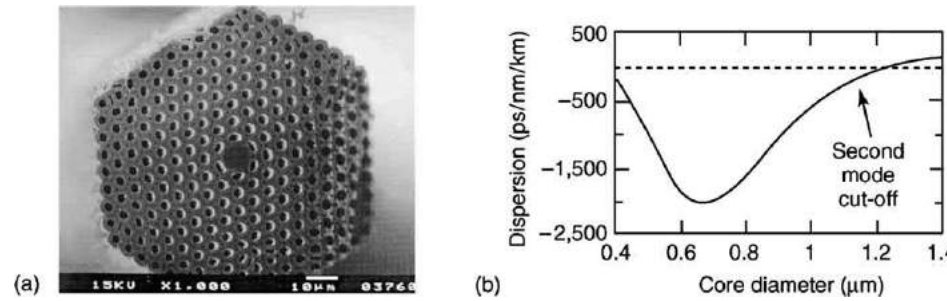


Fig. 18 (a) SEM image of a photonic crystal fiber (holey fiber), and (b) net dispersion of the fiber at 1550 nm as a function of the core diameter. Reproduced with permission from Birk TA, Mogihetsev D, Knight JC and Russell PSt.J (1999) Integrated all-pass filters for tunable dispersion and dispersion slope compensation. *Photon. Technol. Lett.* 11(6): 674–676.

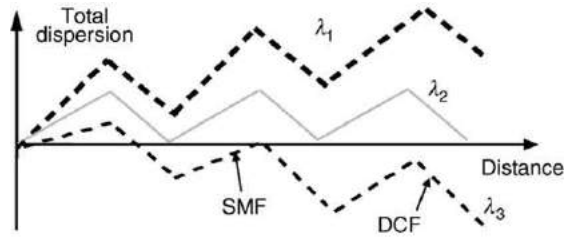


Fig. 19 The chromatic dispersion slope mismatch caused by the different slopes of transmission fiber (SMF or NZDSF) and DCF.

the dispersion of the fiber at the design wavelength, there will be residual dispersion left at the other wavelength channels unless the compensator can match the slope of the dispersion curve of the fiber as well. Some solutions for dispersion slope compensation are described in this section.

First, DCF, with negative dispersion slope, is a prime candidate for deployment as a dispersion slope compensator even though it cannot easily be made tunable. By designing the DCF with the same ratio of dispersion to dispersion slope as that of a real fiber link, new types of DCF can be used to compensate for both dispersion and dispersion slope, much like DCF is used for dispersion compensation today. DCF's popularity, wideband functionality, and stable dispersion characteristics make this a particularly attractive solution. Designing the DCF to match the dispersion characteristics of the transmission fiber is the critical engineering challenge in this slope compensation scheme. Second, third-order nonlinearly chirped FBG can act as a tunable dispersion slope compensator. A simple modification of the nonlinearly chirped FBG allows tuning of the compensated dispersion slope value via stretching the grating. The grating is prepared such that the time delay as a function of wavelength has a cubic profile that covers several WDM channels over a continuous bandwidth of many nanometers. Since the resulting dispersion curve is quadratic over the grating bandwidth, the dispersion slope experienced by the WDM channels can be tuned by stretching the grating using a single mechanical element (see **Fig. 20**). Third, combining the VIPA, with either a 3D mirror or diffraction grating, can also provide tunable free-space dispersion slope compensation. Slope tuning is achieved by dynamically controlling the MEMS-based 3D mirror or the diffraction grating. Fourthly, using an FBG with many spaced thin-film heater sections can also enable tunable dispersion slope. Each heater can be individually electrically controlled, allowing the time delay profile of the grating to be dynamically tuned via changing the temperature along the length of the grating. Advanced applications of this technique can allow alterations to the entire time delay profile of the grating, providing a truly flexible dispersion and dispersion slope compensation mechanism.

Chromatic Dispersion Monitoring

Another important issue related to dispersion management is dispersion monitoring techniques. In a reconfigurable system, it is necessary to reconfigure any tunable chromatic dispersion compensation modules on the fly as the network changes. An in-line chromatic dispersion monitor can quickly measure the required dispersion compensation value while data are still being transmitted through the optical link. This is very different from the more traditional chromatic dispersion measurement techniques where dark fiber is used and the measurement is done off-line over many hours (or days).

Chromatic dispersion monitoring is most often done at the receiving end, where the Q-factor of the received data or some other means is employed to assess the accumulated dispersion. Existing techniques that can monitor dispersion in-line, fast, and with relatively low cost include: (i) general performance monitoring using the bit-error rate (BER) or eye opening, but this approach cannot differentiate among different degrading effects; (ii) detecting the intensity modulation induced by phase modulation at the transmitter; (iii) extracting the bit-rate frequency component (clock) from photo-detected data and monitoring its RF power (see **Fig. 21**); (iv) inserting a subcarrier at the transmitter and subsequently monitoring the subcarrier power degradation;

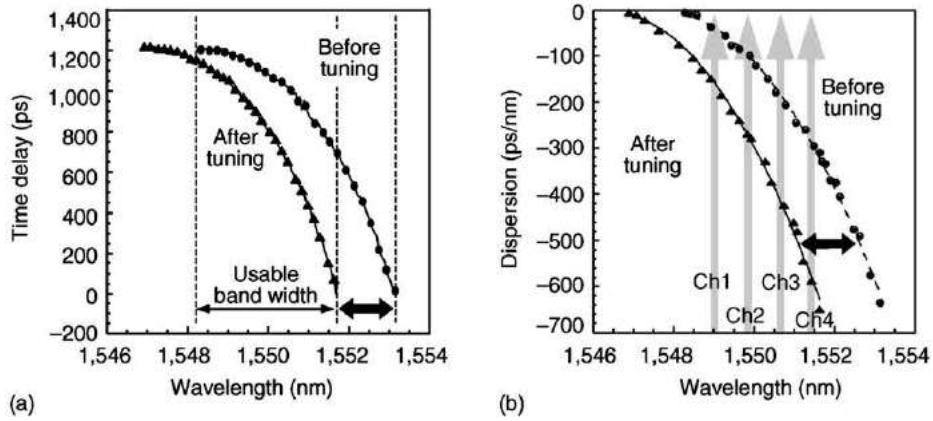


Fig. 20 Tunable dispersion slope compensation using a third-order nonlinearly chirped FBG. (a) Cubic time delay curves of the grating, and (b) quadratic dispersion curves showing the change in dispersion in each channel before and after tuning. Reproduced with permission from Song YW, Motoghian SMR, Starodubov D, *et al.* (2002) Tunable dispersion slope compensation for WDM systems using a non-channelized third-order-chirped FBG. *Optical Fiber Communication Conference* 2002. Paper ThAA4.

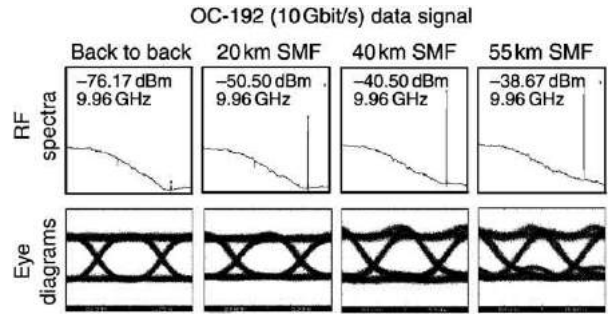


Fig. 21 Clock regenerating effect due to chromatic dispersion for NRZ data. As the amount of residual dispersion increases, so does the amount of power at the clock frequency. This power can be used to monitor the amount of uncompensated chromatic dispersion. Reproduced with permission from Pan Z, Yu Q, Xie Y, *et al.* (2001) Chromatic dispersion monitoring and automated compensation for NRZ and RZ data using clock regeneration and fading without adding signalling. *Optical Fiber Communication Conference 2001*. Paper WH5.

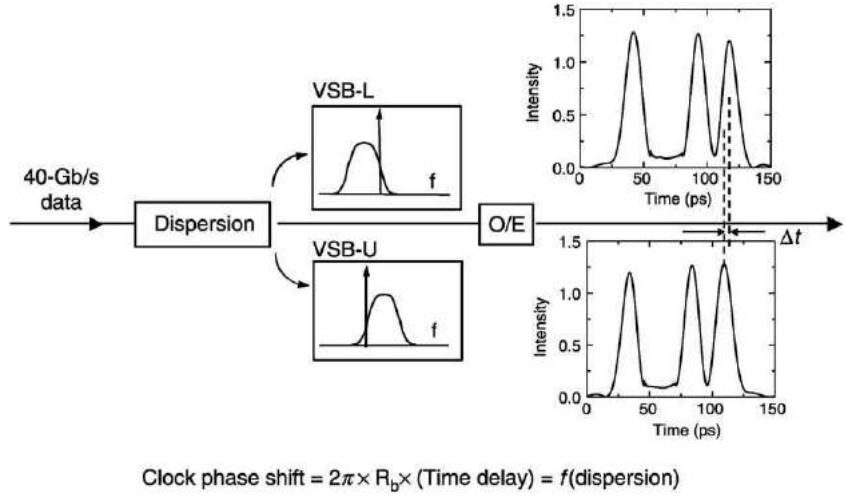


Fig. 22 Chromatic dispersion monitoring using the time delay (Δt) between two VSB signals, which is a function of chromatic dispersion. Reproduced with permission from Yu Q, Yan L-S, Pan Z and Willner AE (2002) Chromatic dispersion monitor for WDM systems using vestigial-sideband optical filtering. *Optical Fiber Communication Conference 2002*. Paper WE3.

(v) extracting several frequency components from the optical data using very narrow-band optical filters and detecting the optical phase; (vi) dithering the optical-carrier frequency at the transmitter and measuring the resultant phase modulation of the clock extracted at receiver with an additional phase-locked loop; and (vii) using an optical filter to select the upper and lower vestigial sideband (VSB) signals in transmitted optical data and determine the relative group delay caused by dispersion (see **Fig. 22**).

Conclusion

Chromatic dispersion is a phenomenon with profound implications for optical fiber communications systems. It has negative effects, broadening data pulses, but it also helps reduce the effects of fiber nonlinearities. For this reason, managing dispersion, rather than trying to eliminate it altogether, is the key. Fixed dispersion components are suitable for point-to-point OC-192 systems. Tunable dispersion components are essential for dispersion management in reconfigurable and OC-768 systems. These dispersion compensation elements must have low loss, low nonlinearity, and must be cost effective.

Although several technologies have emerged that meet some or all of the above requirements, no technology is a clear winner. The trend is towards tunable devices, or even actively self-tunable compensators, and such devices will allow system designers to cope with the shrinking system margins and with the emerging rapidly reconfigurable optical networks.

See also: Dispersion

Further Reading

Agrawal, G.P., 1997. Fiber-Optic Communication Systems. Rochester, NY: John Wiley & Sons.
Agrawal, G.P., 1997. Nonlinear Fiber Optics, 2nd edn. Academic Press.
Gowar, J., 1993. Optical Communication Systems, 2nd edn. Prentice Hall.
Kaminow, I.P., Koch, T.L., 1997. Optical Fiber Telecommunications IIIA & IIIB. Academic Press.
Kashyap, R., 1999. Fiber Bragg Gratings. Academic Press.
Kazovsky, L., Benedetto, S., Willner, A., 1996. Optical Fiber Communication Systems. Artech House.
Willner, A.E., Hoanca, B., 2002. Fixed and tunable management of fiber chromatic dispersion. In: Kaminow, I.P., Li, T. (Eds.), Optical Fiber Telecommunications IV. Academic Press.
Yariv, A., Yeh, P., 1984. Optical Waves in Crystals. John Wiley & Sons.

All-Optical Multiplexing/Demultiplexing

Z Ghassemlooy, Northumbria University, Newcastle upon Tyne, UK
G Swift, Sheffield Hallam University, Sheffield, UK

Introduction

A_i	ith data bit in a packet
c	Speed of light in a vacuum
c_f	Speed of light within the fiber
E_c	Electric fields of control signals
E_{ds}	Electric fields of data signals
f_{FWM}	Frequency due to four-wave mixing
$G_{SLA}(t)$	Gain profiles of SLA
I_c	Intensity of the control signal
I_D	Intensity of the data signal
$\bar{I}_m$	Average photo current for a mark
$\bar{I}_s$	Average photo current for a space
k	Packet length in bits
L	Fiber length
L_d	Distance between two SLAs
L_E	Fiber effective length
n	Number of stages
n_2	Kerr coefficient
N_{SLA}	SLA index
P	Polarization
P_c	Control signal launch power
P_{ds}	Data signal launch power
t_{asy}	The window width is given by $t_{asy}=2\Delta x/c_f$
T	Incoming bit interval
T_w	Walk off time per unit length between control and data pulses
T_x	Transmittance of the NOLM
$\delta(\cdot)$	Optical carrier pulse shape
$\Delta\phi(t)$	Time-dependent phase difference
Δx	SLA offset from the fiber loop center
ε_0	Vacuum permittivity of the medium
λ_s	Data signal wavelength
$\sigma_{amp,m}$	Variances for optical pre-amplifier for a mark
$\sigma_{amp,s}$	Variances for optical pre-amplifier for a space
$\sigma_{rec,m}$	Variances for receiver for a mark
$\sigma_{rec,s}$	Variances for receiver for a space
$\sigma_{RIN,m}$	Variances of RIN for a mark
$\sigma_{RIN,s}$	Variances of RIN for a space
τ	Outgoing bit interval
$\phi(t)$	Phase profile of SLA
$\chi^{(3)}$	Third-order non-linear susceptibility of an optical fiber

Introduction

The ever-increasing aggregate demand of electrically based time division multiplexing systems should have coped with the steady growth rate of voice traffic. However, since 1990, the explosive growth of the Internet and other bandwidth-demanding multimedia applications, has meant that long-haul telecommunication traffic has been increasingly dominated by data, not voice traffic. Such systems suffer from a bandwidth bottleneck due to speed limitations of the electronics. This limits the maximum data rate to considerably less than the THz bandwidth offered by an optical fiber. Optical technology is proposed as the only viable option and is expected to play an ever increasing role in future ultrahigh-speed links/networks. There are a number of multiplexing techniques, such as space division multiplexing (SDM), wavelength division multiplexing (WDM), and optical time division multiplexing (OTDM), that are currently being applied to effectively utilize the bandwidth of optical fiber as a means to overcome the bandwidth bottleneck imposed by electrical time division multiplexing (TDM). In SDM a separate optical fiber is allocated to each channel, but this is the least preferred option for increasing channel numbers. In WDM, a number of different data channels are allocated to discrete optical wavelengths for transmission over a single fiber. Dense WDM technology has been improving at a steady rate in recent years, with the latest systems capable of operating at a data rate of >1 T bps, using a large number of wavelengths over a single fiber link. However, there are a number of problems associated with the WDM systems such as:

- Performance of WDM is highly dependent on the nonlinearities associated with fiber, i.e.:
  - *Stimulated Raman scattering*: degrades the signal-to-noise (SNR) as the number of channels increases;
  - *Four-wave mixing*: limits the channel spacing;
  - *Cross-phase modulation*: limits the number of channels.
- Relatively static optical paths, thus offering no fast switching with high performance within the network;
- Switching is normally carried out by separating each wavelength of each fiber onto different physical outputs. Space switches are then used to spatially switch the separated wavelengths, an extremely inefficient way of utilizing network resources;
- The need for amplifiers with high gain and flat spectra.

In order to overcome these problems, OTDM was introduced that offers the following:

- Flexible high bandwidth on demand (>1 Tbit/s compared to the bit rates of 2.5–40 Gbit/s per wavelength channel in WDM systems);
- The total bandwidth offered by a single channel network is equal to DWDM;

 doi:10.1016/B978-0-12-809283-5.00813-2

- In a network environment, OTDM provides potential improvements in:
  - Network user access time, delay and throughput, depending on the user rates and statistics.
  - Less complex end node equipment (single-channel versus multichannels).
- Self-routing and self-clocking characteristics;
- Can operate at second- and third-transmission windows:
  - 1500 nm (like WDM) due to Erbium doped fiber amplifier (EDFA);
  - 1300 nm wavelengths.
- Offers both broadcast and switched based networks.

Principle of OTDM

Fig. 1 show the generic block diagram of a point-to-point OTDM transmission link, where N optical data channels, each of capacity M Gbps, are multiplexed to give an aggregate rate of $N \times M$ Gbps. The fundamental components are a pulsed light source, an optical modulator, a multiplexer, channel, add/drop unit, and a demultiplexer. The light source needs to have good stabilities and be capable of generating ultrashort pulses (< 1 ps). Direct modulation of the laser source is possible but the preferred method is based on external modulation where the optical signal is gated by the electronic data. The combination of these techniques allows the time division multiplexed data to be encoded inside a subnanosecond time slot, which is subsequently interleaved into a frame format. Add/drop units provide added versatility (see Fig. 2) allowing the 'adding' and 'dropping' of selected OTDM channels to interchange data at chosen points on the link. At the receiving end, the OTDM pulse stream is demultiplexed down to the individual channels at the initial M Gbps data rate. Data retrieval is then within the realm of electronic devices and the distinction between electronic and optical methods is no longer relevant. Demultiplexing requires high-speed all optical switching and can be achieved using a number of methods, which will be discussed in more detail below.

The optical multiplexing (or interleaving) can be carried out at the bit level (known as bit interleaving) or at the packet level (known as packet interleaving), where blocks of bits are interleaved sequentially. This is in accord with the popular conception of packet switched networks. The high data rates required and consequently narrow time slots necessitate the need for strict tolerances at processing nodes, e.g., switches. As such, it is important that the duration of the optical pulses is chosen to be significantly shorter than the bit period of the highest multiplexed line rate, in order to reduce the crosstalk between channels.

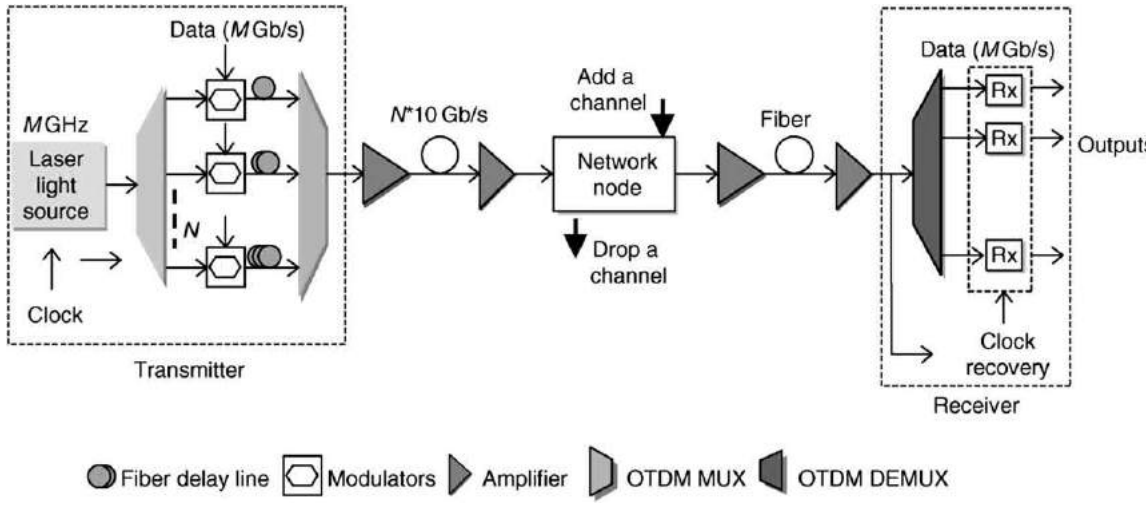


Fig. 1 Block diagram of a typical OTDM transmission system.

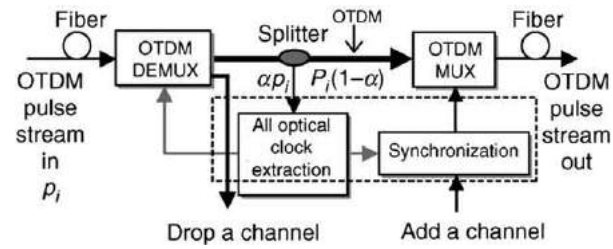


Fig. 2 Add/drop unit in an OTDM network node.

Bit Interleaved OTDM

A simple conceptual description of a bit interleaved multiplexer is shown in Fig. 3. It uses a number of different length optical fiber delay lines (FDL) to interleave the channels. The propagation delay of each FDL is chosen to position the optical channel in its corresponding time slot in relation to the aggregate OTDM signal. Prior to this, each optical pulse train is modulated by the data stream. The output of the modulators and an undelayed pulse train, labeled the framing signal, are combined, using a star coupler or combiner, to produce the high bit rate OTDM signal (see Fig. 3(b)). As shown in Fig. 3(b), the framing pulse has a higher intensity for clock recovery purpose. At the demultiplexer, the incoming OTDM pulse train is split into two paths (see Fig. 4(a)). The lower path is used to recover the framing pulse by means of thresholding, this is then delayed by an amount corresponding to the position of the *i*th (wanted) channel (see Fig. 4(b)). The delayed framing pulse and the OTDM pulse stream are then passed through an AND gate to recover the *i*th channel. The AND operation can be carried out all optically using, for example, a nonlinear loop mirror, a terahertz optical asymmetric demultiplexer, or a soliton-trapping gate.

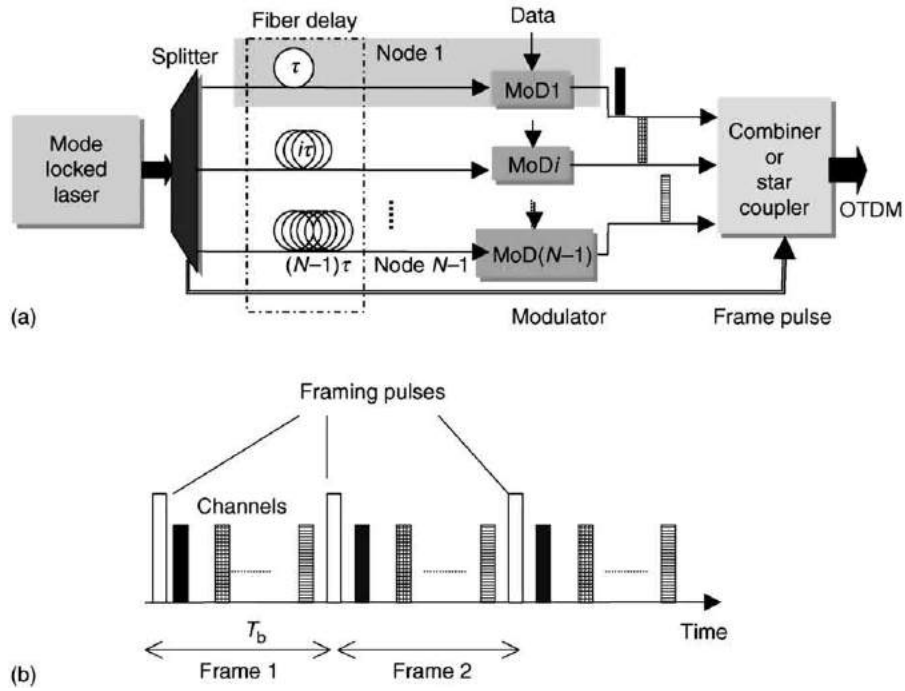


Fig. 3 Bit interleaved OTDM; (a) block diagram and (b) timing waveforms.

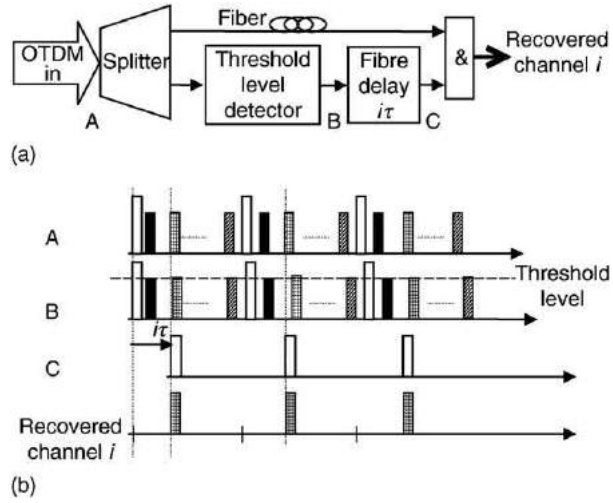


Fig. 4 Bit interleaving demultiplexer: (a) block diagram and (b) typical waveforms.

Packet Interleaved OTDM

Fig. 5(a) shows the block diagram of a system used to demonstrate packet interleaving. Note that the time interval between successive pulses now needs to be much less than the bit interval T. This is achieved by passing the low bit rate output packet of the modulator into a compressor, which is based on a feed forward delay line structure. The feed forward delay lines are based on a cascade of passive M–Z interferometers with unequal arm lengths. This configuration generates the delay times required, i.e., $T-\tau, 2(T-\tau)\ldots(2^{n-1})(T-\tau)$, etc., where $n=\log_2 k$ is the number of stages, T and τ are the incoming bit duration and the outgoing bit duration, respectively, and k is the packet length in bits. Pulse i's location at the output is given by $(2^n-1)(T-t)+(i-1)\tau$. The signal at the input of the compressor is:

$$I_{\text{in}}(t)=\sum_{i=0}^{k-1}\delta(t-iT)A_i \tag{1}$$

Where $\delta(\cdot)$ is the optical carrier pulse shape, and A_i is the ith data bit in the packet.

As shown in Fig. 5(b), each bit is split, delayed, and combined to produce the four bit packets O_i:

$$O_i(t)=\frac{1}{2^{n+1}}\sum_{i=0}^{k-1}I_i[t-i(T-\tau)] \tag{2}$$

Note the factor of $1/(2^{n-1})$ is due to the signal splitting at the 2×2 3-dB coupler at the input of each stage.

The combined signal is shown in Fig. 5(b), and is defined as:

$$I_{\text{out}}=\sum_{i=0}^{k-1}O_i=\frac{1}{2^{n+1}}\sum_{i=0}^{k-1}\sum_{j=0}^{k-1}I_i[t-(i+j)T+j\tau]A_i \tag{3}$$

For a packet delay of $(i+j)T$ packets are being built up. For the case $(i+j=3)$, four bits are compressed into a packet (see Fig. 5(b)). An optical gating device is used to select the only complete usable compressed copy of the incoming packets from the unwanted copies.

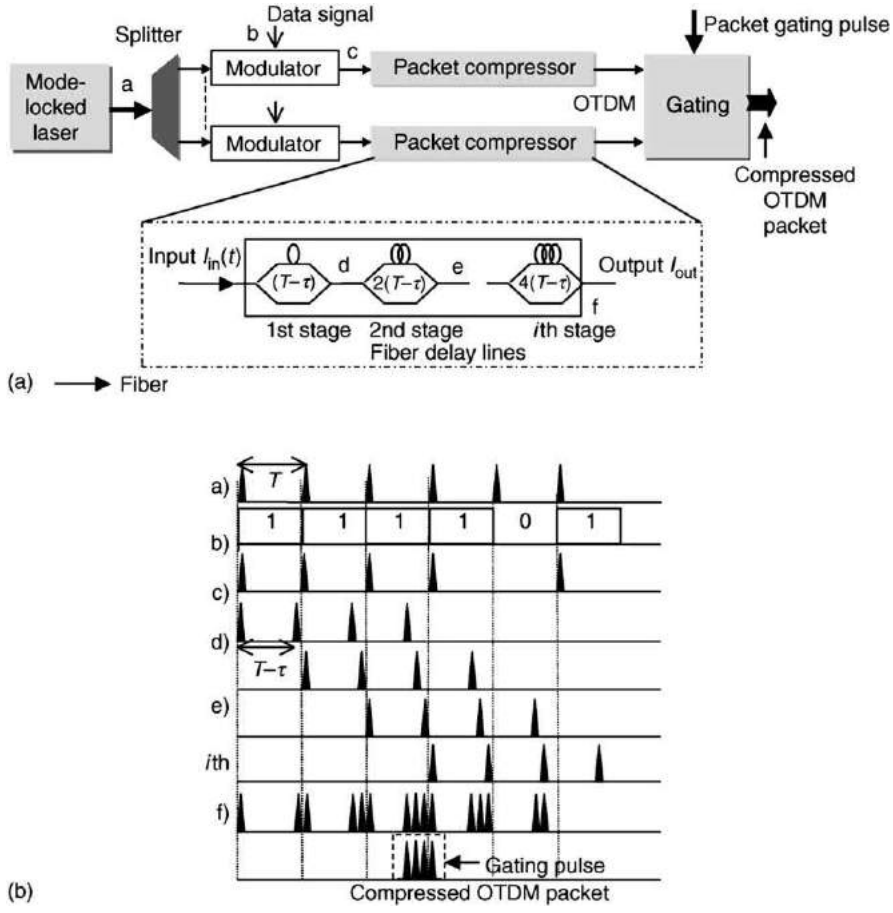


Fig. 5 Packet interleaving multiplexer: (a) block diagram and (b) typical waveforms.

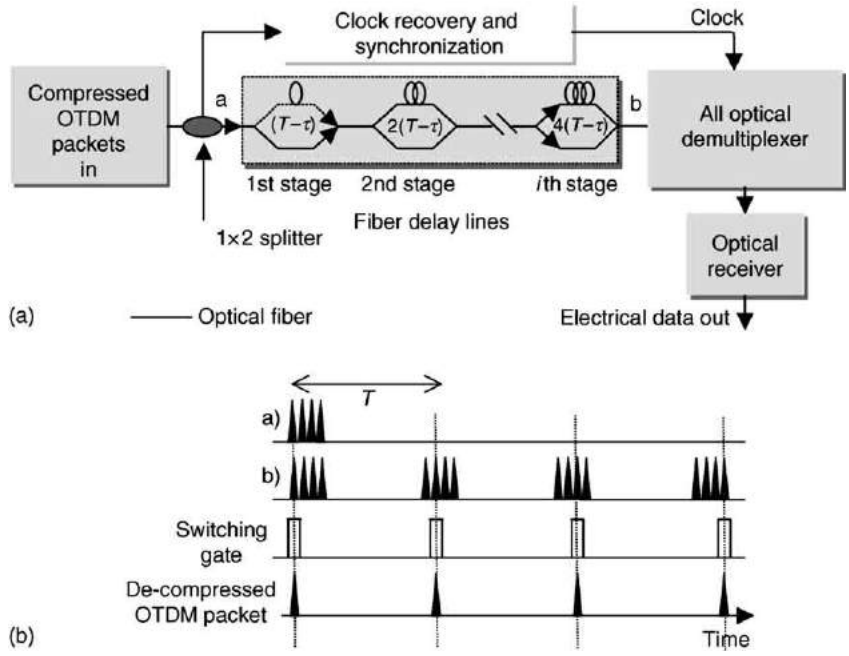


Fig. 6 Packet interleaving demultiplexer: (a) block diagram and (b) typical waveforms.

At the receiving end, demultiplexing is carried out by decompressing the OTDM packet stream using the same delay line structure (see Fig. 6). Multiple copies of the compressed packet are generated, and each is delayed by $(T-\tau)$. An optical demultiplexer placed at the output of the delay lines generates a sequence of switching window at the rate of $1/T$. By positioning the switching pulses at the appropriate position, a series of adjacent packet (channel) bits from each of the copied compressed packets is extracted (see Fig. 6(b)). The output of the demultiplexer is the decompressed packet signal, which can now be processed at a much slower rate using electronic circuitry.

Components of an OTDM System

Optical Sources

In an ultrahigh speed OTDM system, it is essential that the optical sources are capable of generating transform-limited sub-nanosecond pulses having low duty cycles, tuneability, and a controllable repetition rate for synchronization. A number of suitable light sources are: gain-switch distribute feedback laser (DFB), active mode locked lasers (MLL) (capable of generating repetitive optical pulses), and harmonic mode-locking Erbium-doped fiber (EDF) lasers. Alternative techniques include super-continuum pulse generation in a dispersion shifted fiber with EDF pumping, adiabatic soliton compression using dispersion-flattened dispersion decreasing fiber, and pedestal reduction of compressed pulses using a dispersion-imbalanced nonlinear optical loop mirror. As shown in Fig. 1, the laser light source is power split to form the pulse source for each channel. These are subsequently modulated with an electrical data signal (external modulation). External modulation is preferred in an OTDM system as it can achieve narrow carrier linewidth, thus reducing the timing jitter of the transmitted pulse. For ultrahigh bit rate OTDM systems, the optical pulses emerging from the external modulators may also need compressing. One option is to frequency-chirp the pulses and pass them through an anomalous dispersive medium. Using this approach, it is essential that the frequency chirp be linear throughout the duration of the pulse, in addition the midpoint of the linear-frequency chirp should coincide with the center of the pulse. When a frequency-chirped pulse passes through a dispersive medium, different parts of the pulse travel at different speeds, due to a temporal variation of frequency. If the trailing edge travels faster than the leading edge, the result would be pulse compression. An optical fiber cable or a semiconductor laser amplifier (SLA) can be used to create the frequency chirp.

Multiplexers

Multiplexing of the pulses generated by the optical sources can be implemented either passively or actively. The former method is commonly implemented using a mono-mode optical fiber. This method has the advantage of being simple and cost-effective. The latter method uses devices more complex in nature, for example, electro-optic sampling switches, semiconductor optical amplifiers, and integrated optics.

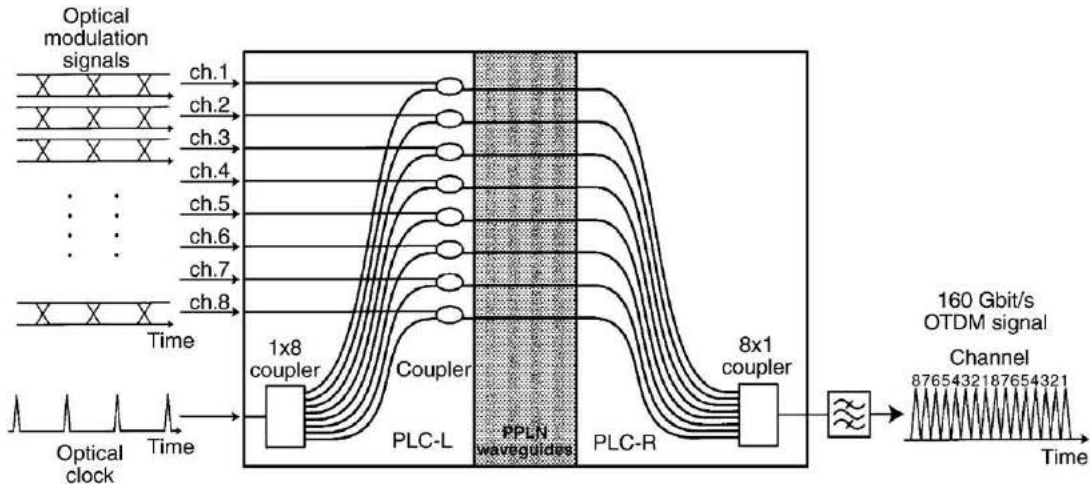


Fig. 7 Configuration of a PLC-OTDM-MUX. Reproduced with permission from Ohara T, Takara H and Shake I, *et al.* (2003). 160-Gb/s optica-time-division multiplexing with PPLN hybrid integrated planar lightwave circuit. *IEEE Photonics Letters* 15(2): 302–304. © 2003 IEEE.

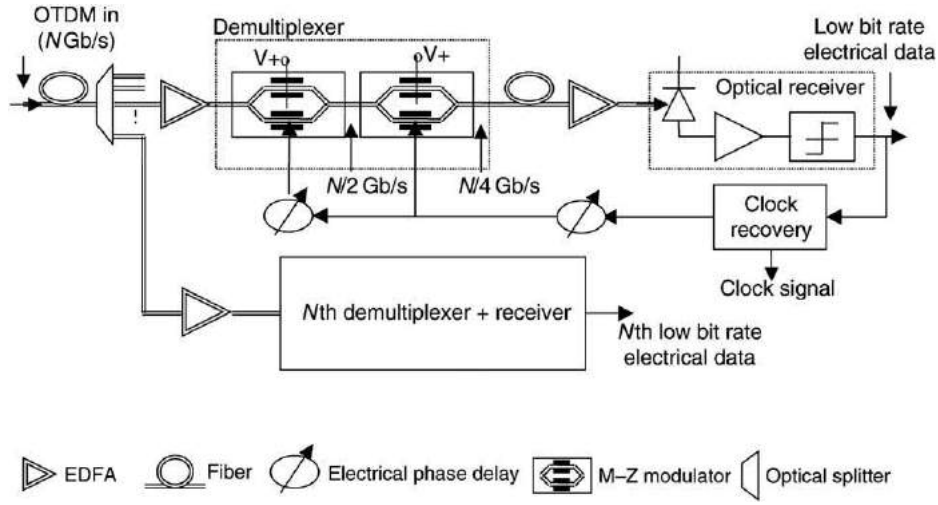


Fig. 8 Electro-optics demultiplexer and receiver system block diagram.

An integrated active multiplexer can be made by integrating semiconductor laser amplifiers (SOA) into a hybrid planar lightwave circuit (PLC). If the optical path length is kept short then temperature control is made easier. An alternative to this approach is to use lithium niobate (LiNb) technology. Periodically poled LiNb (PPLN) based OTDM multiplexers offer compact size and low noise (due to the absence of amplifier spontaneous emission noise and pattern effect, which is a feature of SOA based devices). **Fig. 7** is a schematic of a PLC-based OTDM multiplexer composed of two PLCs (R and L), composed of 1×8 and 8×1 couplers, eight 2×1 couplers, and eight different path length PPLN waveguides. The input clock pulse is split into eight by the 1×8 coupler in PLC-L, and are combined with the modulated optical pulse trains using the 2×1 couplers. The outputs of the 2×1 couplers are then passed through a PPLN waveguide to generate a return-to-zero (RZ) optically modulated signal. These are then combined, using the 8×1 coupler in PLC-R. When the path-length difference between the waveguides is set to one time slot (equal to 1/8 of the base data rate), then the output of the 8×1 coupler is the required high-bit rate OTDM pulse.

Demultiplexers

In contrast to multiplexing, demultiplexing must be performed as an active function. It can be implemented electro-optically or optically. The former method needs to complete demultiplexing of all channels in order to extract a single channel (see **Fig. 8**). The demultiplexer in **Fig. 8** uses two LiNb Mach–Zehender (M–Z) modulators in tandem. The first and second modulators are driven with a sinusoidal signal of amplitudes $2V_\pi$ and V_π, respectively, to down-covert the N-bit rate to $N/2$ and $N/4$, respectively. Channels can be selected by changing either the DC-bias V+ to the M–Zs, or the electrical phase delay. At ultrahigh speed

implementation of an electro-optics demultiplexer becomes increasingly difficult due to the higher drive voltage requirement by the M–Z. An alternative is to use all optical demultiplexing based on the nonlinear effect in a fiber and optical active devices offering switching resolution in the order of picoseconds. There are a number of methods available to implement all optical demultiplexing. The most popular methods that use fast phase modulation of an optical signal are based on M–Z and Sagnac interferometers. Four-wave mixing is another popular method.

Mach–Zehnder (M–Z) Interferometers

The key to interferometric switching is the selection of an appropriate material for the phase modulation. Semiconductor materials, often in the form of an SLA, are suitable devices for this purpose. The refractive index of an SLA is a function of the semiconductor carrier density, which can be modulated optically resulting in fast modulation. If a high- intensity pulse is input to an SLA the carrier density changes nonlinearly, as shown in Fig. 9. Phase modulation is affected via the material chirp index (dN/dn), which represents the gradient of the refractive index carrier density curve for the material. The phase modulation is quite strong and, in contrast to the intensity-based nonlinearity in an optical fiber, (see Kerr effect below), is a consequence of a resonant interaction between the optical signal and the material. The nonlinearity is relatively long-lived and would be expected to limit the switching speed when using semiconductors. Semiconductors, when in excitation from an optical field, tend to experience the effect quite quickly (ps) with a slow release (hundreds of picoseconds) time. Advantage can be taken of this property by allowing the slow recovery to occur during the time between channels of an OTDM signal, as this time may be of the order of hundreds of pico-seconds or more. A Mach–Zehnder configuration using two SLAs, one in each arm of the interferometer placed asymmetrically, is shown in Fig. 10. An optical control pulse entering the device via port 3 initiates the nonlinearity. The transmission equation relating the input signal (port 1) to the output port is given by

$$\frac{I_{\text{out}}(t)}{I_{\text{in}}(t)} = 0.25(G_{\text{SLA1}}(t) + G_{\text{SLA2}}(t) \pm 2\sqrt{G_{\text{SLA1}}(t)G_{\text{SLA2}}(t)}\cos\Delta\phi(t)) \tag{4}$$

where $G_{\text{SLA1}}(t)$ and $G_{\text{SLA2}}(t)$ refer to the gain profiles of the respective SLAs, and $\Delta\phi(t)$ is the time-dependent phase difference between them. Assuming that the gain and phase profiles of an excited amplifier are given by $G(t)$ and $\phi(t)$, respectively, then the

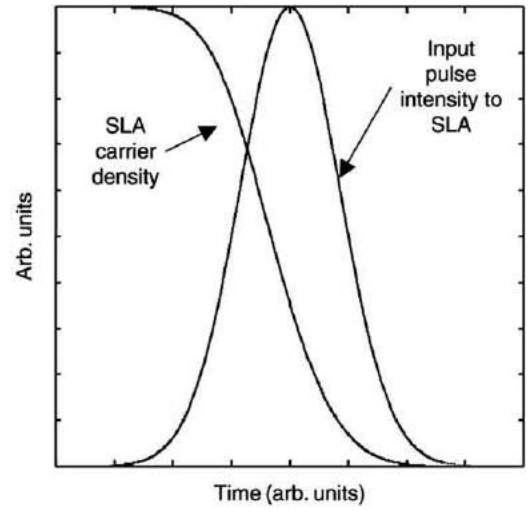


Fig. 9 SLA carrier density response to input pulse.

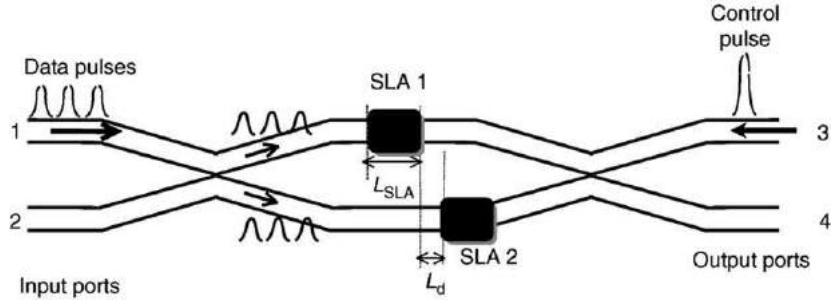


Fig. 10 Asymmetric TWSLA Mach–Zehnder devices.

signals passing through the upper and lower arms experience optical properties of the material given by:

$$G(t), G(t - T_d), \phi(t) \text{ and } \phi(t - T_d) \tag{5}$$

where T_d is given by $2L_dN_{SLA}/c$, L_d is the distance between SLA1 and SLA2, N_{SLA} is the SLA index, and c is the speed of light in a vacuum. As the width of the $\Delta\phi(t)$ profile is dependent on the distance between the SLAs, placing them in close proximity allows high-resolution switching. Less emphasis has been placed on the SLA gain as this is considered to be less effective when phase differences of π are reached. It only remains for the gain and phase modulation to recover, and as indicated previously, this is allowable over a time-scale commensurate with tributary rates.

Sagnac Interferometers

There are two main types; the nonlinear optical loop mirror (NOLM), and the terahertz optical asymmetric demultiplexer (TOAD).

NOLM

In this method of switching the inherent nonlinearity of an optical fiber known as the Kerr effect is used. The phase velocity of any light beam passing through the fiber will be affected by its own intensity and the intensity of any other beams present. When the intrinsic third-order nonlinearity of silica fibers is considered via the intensity-dependent nonlinear refraction component, then the signal phase shift is

$$\Delta\phi_{signal} = \frac{2\pi}{\lambda_{ds}} n_2 L I_{ds} + 2\frac{2\pi}{\lambda_{ds}} n_2 L I_c \tag{6}$$

where n_2 is the Kerr coefficient, I_{ds} is the intensity of the data signal to be switched, I_c is the intensity of the control signal used to switch the data signal, λ_{ds} is the data signal wavelength, and L is the fiber length. The optical loop mirror consists of a long length of mono-mode fiber formed into a fiber coupler at its free ends (see **Fig. 11**). The input to the loop comprises the high-frequency data stream plus a control pulse at the frame rate. The data split at the coupler and propagate around the loop in contra directions (clockwise E_{CW} and counter-clockwise E_{CCW}) recombining back at the coupler. In the absence of a control pulse, the pulse exits via port 1. If a particular pulse in the loop (in this example E_{CW}) is straddled by the control pulse (see **Fig. 12**), then that pulse experiences cross-phase modulation, according to the second term on the right-hand side of **Eq. (6)**, and undergoes a phase change relative to E_{CW}. The difference in the phase between E_{CW} and E_{CCW} causes the pulse to exit via port 2. The phase shift profile experienced by the co-propagating pulse is

$$\Delta\phi(t) = \frac{4\pi}{\lambda_{ds}} n_2 \int_0^L I_c(t)\mathrm{d}x \tag{7}$$

Assuming unity gain around the loop, the transmittance of the NOLM is

$$T_x = 1 - \cos^2\left(\frac{\Delta\phi}{2}\right) = \Delta\phi(t) = 1 - \cos^2\left(\frac{2\pi}{\lambda_{ds}} n_2 \int_0^L I_C(t)\mathrm{d}x\right) \tag{8}$$

As it stands, the switching resolution is determined by the width of the control pulse; however, it is possible to allow the signal and control to 'walk off' each other, allowing the window width to be increased by an amount determined by the 'walk off'. The phase shift is now

$$\Delta\phi(t) = 2\frac{2\pi}{\lambda_{ds}} n_2 \int_0^L I_C(t - T_w x)\mathrm{d}x \tag{9}$$

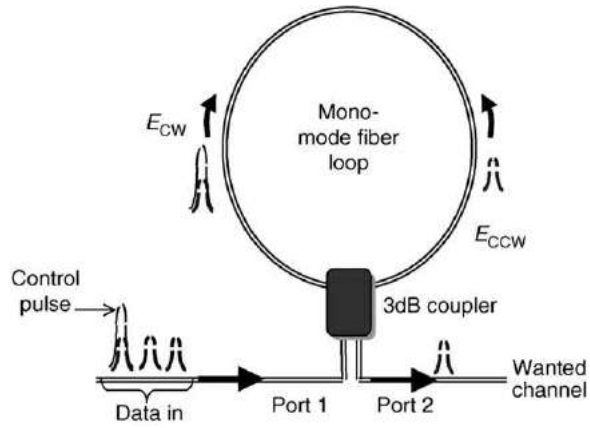


Fig. 11 Nonlinear optical loop mirror.

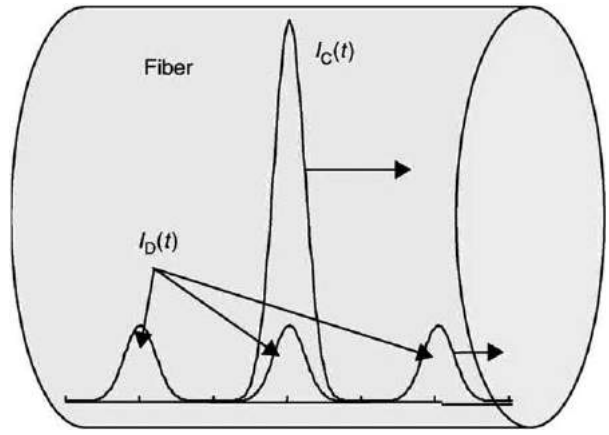


Fig. 12 Control and data pulses propagation in the fiber.

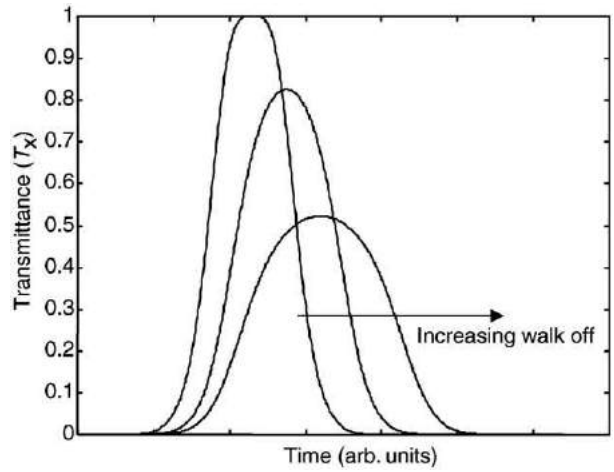


Fig. 13 Transmittance profiles with the walk-off time as a parameter.

where the parameter T_w, is the walk off time per unit length between control and signal pulses. The increased window width is accompanied by a lower peak transmission (see Fig. 13).

TOAD

The TOAD uses a loop mirror architecture that incorporates an SLA and only needs a short length of fiber loop (see Fig. 14(a)). The SLA carrier density is modulated by a high-intensity control pulse, as in the M–Z-based demultiplexer. The operation of the TOAD is similar to the NOLM, where the loop transmittance is determined by the phase difference between CW and CCW traveling pulses. Strictly speaking, the gain of the SLA must also be taken into account; however, without any loss of generality the effect is adequately described by considering only the phase property. Fig. 14(a) shows the timing diagram associated with a TOAD demultiplexer. The control pulse, shown in Fig. 14(b1), is incident at the SLA at a time t_1; the phase profiles for the CCW and CCW data pulses are shown in Figs. 14(b2) and (3), respectively. The resulting transmission window is shown in Fig. 14(b4). The window width is given by $t_{asy} = 2\Delta x/c_f$, where Δx is the SLA offset from the loop center and c_f is the speed of light in the fiber. As in NOLM, if the phase difference is of sufficient magnitude, then data can be switched to port 2. The switch definition is, in principle, determined by how close the SLA is placed to the loop center when the asymmetry is relatively large. However, for small asymmetries the switching window is asymmetric, which is due to the CW and CCW gain profiles being different (see Fig. 15). Assuming the phase modulation dominates the transmission then the normalized transmission for a small asymmetry loop is as depicted in Fig. 16. The gain response (not shown) would have a similar shape and temporal position as the phase response.

Four-Wave Mixing (FWM)

FWM demultiplexing uses a concept whereby two optical signals of different wavelengths are mixed together in a nonlinear medium to produce harmonic components. The nonlinearity in this case arises from the third-order nonlinear susceptibility $\chi^{(3)}$ of

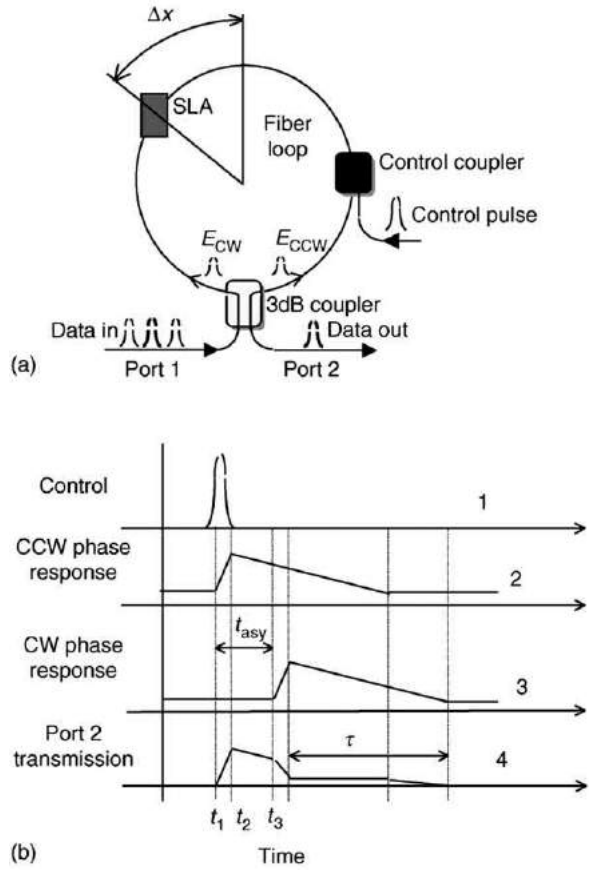


Fig. 14 TOAD: (a) architecture and (b) timing diagrams.

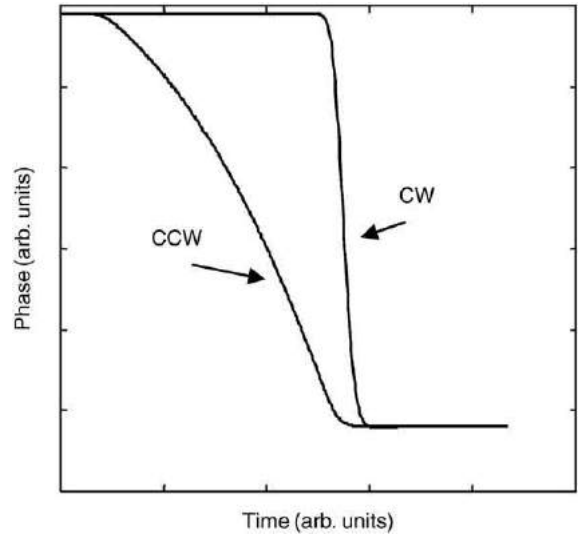


Fig. 15 Phase responses for CW and CCW components of TOAD.

an optical fiber, such that the polarization P induced on a pair of electric fields propagating through the fiber is

$$P = \varepsilon_0 \chi^{(3)} (E_{ds} + E_c)^3 \tag{10}$$

where ε_0 is the vacuum permittivity of the medium, E_{ds} and E_c are the electric fields of data and control signals, respectively. The mixing of the two signals takes place in a long length of fiber in which the control signal propagates alongside the channel to be demultiplexed (see Fig. 17). The control signal is of sufficient intensity to cause the fiber to operate in the nonlinear regime.

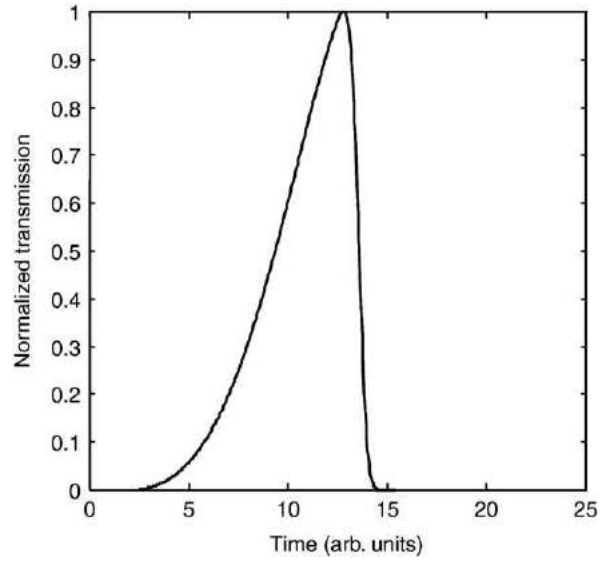


Fig. 16 TOAD transmission window profile for small asymmetry loop.

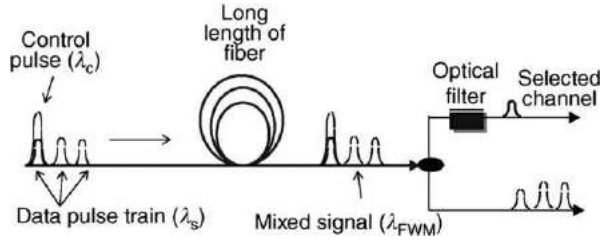


Fig. 17 Block diagram of FWM demultiplexer.

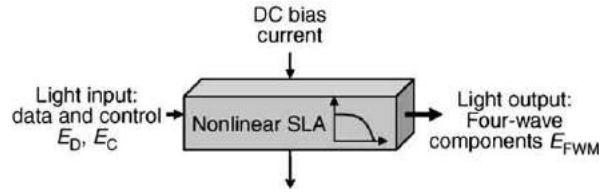


Fig. 18 Four-wave mixing in SOA.

The nonlinear relationship causes a number of frequencies to be generated, with the ones of interest having a frequency given by

$$f_{\mathrm{FWM}} = 2f_{\mathrm{c}} - f_{\mathrm{ds}} \tag{11}$$

An optical filter is then used to demultiplex the required channel from the composite signal. The nonlinearity is detuned from the resonant frequency of the fiber glass and as such tends to be weak, requiring long lengths of fiber to give a measurable effect. The power in the demultiplexed signal, for given data and control signal wavelengths and fiber material, depends on the input power and the fiber length according to

$$P_{\mathrm{FWM}} = kP_{\mathrm{ds}}P_{\mathrm{c}}L_{\mathrm{E}}^2 \tag{12}$$

where k is a constant, P_{ds} is the signal launch power, P_{c} is the control signal launch power, and L_{E} is the fiber effective length. FWM is essentially an inefficient method as power is wasted in the unused frequency components of the four-wave signal. More in line with integrated structures, the nonlinear properties of a semiconductor laser amplifier can be used. Here a relatively high-power optical signal is input to the SLA (see **Fig. 18**). This enables saturation and operation in the nonlinear regime (see **Fig. 19**). Operating the SLA in saturation allows the nonlinear effects to produce the FWM components as in the fiber method.

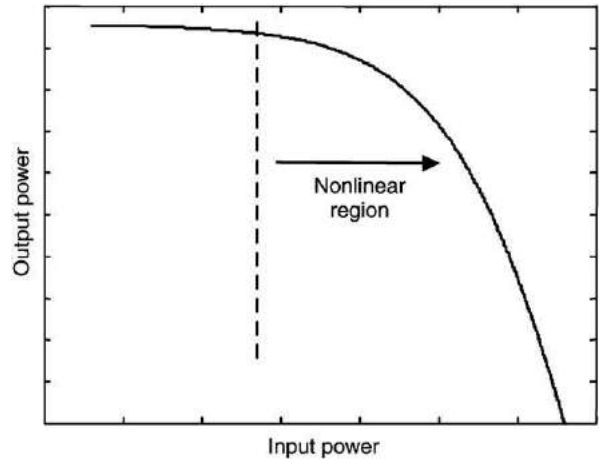


Fig. 19 SLA output–input power curve.

Clock and Data Synchronization in OTDM

In common with electronically based TDM systems, clock recovery is fundamental to the recovery of data in ultrahigh-speed OTDM systems. Two main methods are proposed: (i) clock signal transmitted with the OTDM signal (i.e., multiplexed); and (ii) extraction of clock signal from the incoming OTDM pulse stream. In a packet-based system, synchronization between the clock and the data packet is normally achieved by sending a synch pulse with each packet. However, techniques based on optical-phased locked loops are proving popular and remove the need for a separate clocking signal.

Clock Multiplexing

(i) *Space division multiplexing*: This is conceptually the simplest to implement, where the clock signal is carried on a separate fiber from the data. However, it is susceptible to any differential delay between the different paths taken by clock and data due to temperature variation. It is difficult to justify in systems where the installed fiber base is at a premium.

(ii) *Wavelength division multiplexing*: Here different wavelengths are allocated to the clock, and payload. It is only really practical for predetermined path lengths between nodes in single-hop networks such as point–point links or broadcast-and-select star networks. It also suffers from random delays between the clock and the payload, which is problematic in an asynchronous packet switched based network, where the optical path length a packet may take is nondeterministic.

(iii) *Orthogonal polarization*: This is suitable for small links, where separate polarizations are used for the clock and data. However, in large networks it is quite difficult to maintain the polarization throughout the transmission link due to polarization mode dispersion and other nonlinear effects.

(iv) *Intensity division multiplexing*: This uses higher-intensity optical clock pulses to differentiate it from the data pulses as discussed above. However, in long-distance transmission links, it is difficult to maintain both the clock intensity and its position, due to the fiber nonlinearity.

(v) *Time division multiplexing*: In this scheme a single clock pulse, which has the same wavelength, polarization, and amplitude as the payload pulses, is separated in time, usually ahead of the data pulses.

Synchronization – Optical Phased Locked Loops (PLL)

The PLL is a common technique used for clock recovery in electronic TDM systems. However, the speed of conventional electronic PLLs tends to be limited by the response of the phase comparators used. There are a number of approaches based on opto-electronic PLL. Opto-electronic PLLs based on four-wave mixing in a traveling wave laser amplifier are complex and can suffer from frequency modulation in the recovered clock. However, others based on balanced photo-detectors, result in low timing jitter and good phase stability. In contrast, a number of all optical methods clock recovery scheme exist, one technique based on the TOAD (see above) is depicted in Fig. 20. The high-speed data stream enters the TOAD at a rate $n \times R$ where n is the number of channels per OTDM frame, and R is the frame rate. A pulse generator, such as a mode locked fiber laser (MLFL) clocked by a local oscillator (LO), is used as the TOAD control input (MLFL-C). The OTDM data is switched by the TOAD at a frequency of say, $R + \Delta f$ Hz. Thus, the switching window samples data pulses at a rate higher or lower than $n \times R$ Hz and uses this signal for cross correlation in the PLL unit. The output of the phase comparator is used to regulate a voltage controlled oscillator (VCO) running at R Hz, which in turn feeds the control signal. The PLL circuit locks into the clock frequency, generating the clock signal $S_c(t)$.

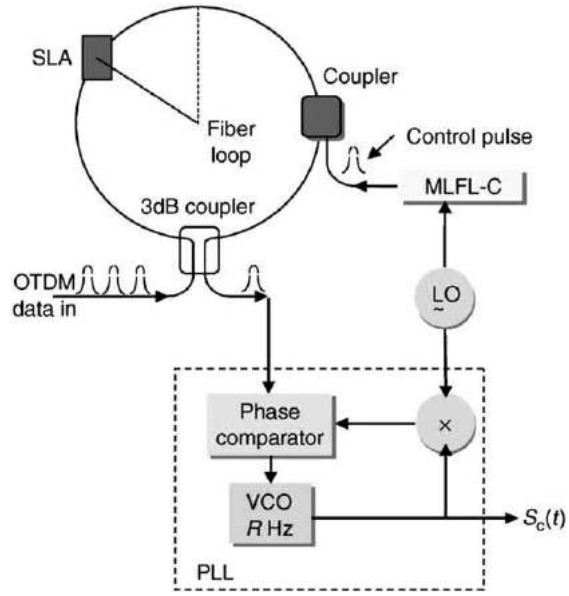


Fig. 20 Ultrafast clock recovery using TOAD.

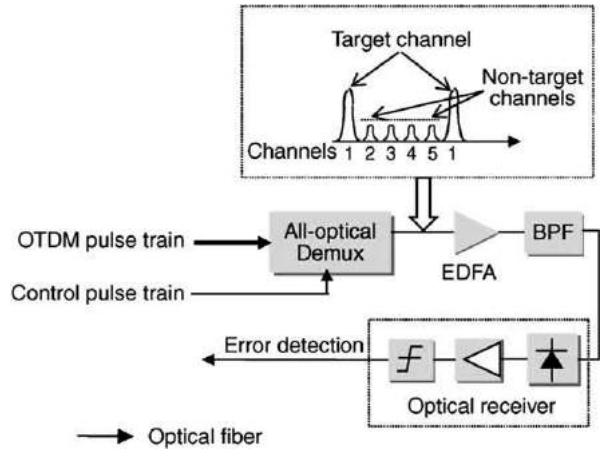


Fig. 21 Block diagram of OTDM receiver.

OTDM Bit-Error Rate (BER) Performance

Fig. 21 shows a typical block diagram of a high-speed optical receiver for an OTDM system, composed of a NOLM or a TOAD demultiplexer, an optical pre-amplifier, an optical bandpass filter, and a conventional optical receiver using a PIN photodiode. Due to the crosstalk introduced in the demultiplexing process, the demultiplexed optical signal contains not only the target channel but may also contain nontarget channels with reduced amplitude (see the inset in Fig. 21). The intensity of the demultiplexed optical signal is boosted by the optical pre-amplifier (EDFA). Amplified spontaneous emission (ASE) from the optical preamplifier adds a wide spectrum of optical fields onto the demultiplexed optical signal. Although an optical bandpass filter (BPF) can reduce the ASE, it still remains one of the major noise sources. The function of the optical filter is to reduce the excess ASE to within the range of the signal spectrum. The PIN photodiode converts the received optical power into an equivalent electrical current, which is then converted to a voltage signal by an electrical amplifier. Finally, the output voltage is sampled periodically by a decision circuit for estimating the correct state (mark/space) of each bit.

The BER performance of an OTDM system deteriorates because of the noise and crosstalk introduced by the demultiplexer and is:

$$\mathrm{BER} = \frac{1}{\sqrt{2\pi}} \frac{\exp(-Q^2/2)}{Q} \tag{13}$$

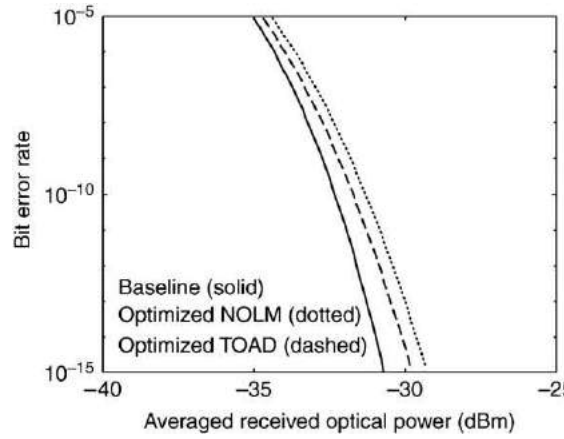


Fig. 22 BER versus the average received optical power for 100 Gb/s (10 channels) OTDM system.

where Q is defined as:

$$Q = \frac{\bar{I}_{\mathrm{m}} - \bar{I}_{\mathrm{s}}}{\sqrt{\sigma^2_{\mathrm{RIN,m}} + \sigma^2_{\mathrm{RIN,s}} + \sigma^2_{\mathrm{amp,m}} + \sigma^2_{\mathrm{amp,s}} + \sigma^2_{\mathrm{rec,m}} + \sigma^2_{\mathrm{rec,s}}}} \tag{14}$$

and $\bar{I}_{\mathrm{m}}$ and $\bar{I}_{\mathrm{s}}$ are the average photocurrents for a mark and a space, respectively. $\sigma_{x,i}$ are the variances of the relative intensity noise (RIN), further optical pre-amplifier and receiver for a mark and a space.

For 100 Gb/s (10 channels) OTDM system, the BER against average received optical power for optimized NOLM and TOAD demultiplexers, is shown in Fig. 22.

See also: Wavelength Division Multiplexing

Further Reading

Agrawal, G.P., 2002. Fiber-Optic Communication Systems, 3rd edn. New York: J Wiley and Sons Inc.
Chang, K. (Ed.), 2003. Handbook of Optical Components and Engineering, 2nd edn. New York: Wiley and Sons Inc.
De Marchis, G., Sabella, R., 1999. Optical Networks Design and Modelling. Dordrecht: Kluwer Academic.
Hamilton, S.A., Robinson, B.S., Murphy, T.E., Savage, S.J., Ippen, E.P., 2002. 100 Gbit/s optical time-division multiplexed networks. Journal of Lightwave Technology 20 (12), 2086–2100.
Jhon, Y.M., Ki, H.J., Kim, S.H., 2003. Clock recovery from 40 gbps optical signal with optical phase-locked loop based on a terahertz optical asymmetric demultiplexer. Optics Communications 220, 315–319.
Kamatani, O., Kawanishi, S., 1996. Ultrahigh-speed clock recovery with phase locked loop based on four wave mixing in a traveling-wave laser diode maplifier. Journal of Lightwave Technology 14, 1757–1767.
Nakamura, S., Ueno, Y., Tajima, K., 2001. Femtosecond switching with semiconductor-optical-amplifier-based symmetric-Mach–Zehnder-type all-optical switch. Applied Physics Letters 78 (25), 3929–3931.
Ohara, T., Takara, H., Shake, I., *et al.*, 2003. 160-Gb/s optica-time-division multiplexing with PPLN hybrid integrated planar lightwave circuit. IEEE Photonics Letters 15 (2), 302–304.
Ramaswami, R., Sivarajan, K.N., 2002. Optical Network; a Practical Perspective. Morgan Kaufman.
Reman, E., 2001. Trends and evolution of optical networks and technology. Alcatel Telecommun. Rev. 3rd Quarter 173–176.
Sabella, R., Lugli, P., 1999. High Speed Optical Communications. Dordrecht: Kluwer Academic.
Seo, S.W., Bergman, K., Prucnal, P.R., 1996. Transparent optical networks with time division multiplexing. Journal of Lightwave Technology 14 (5), 1039–1051.
Sokolloff, J.O., Prucnal, P.R., Glesk, I., Kane, M., 1993. A terahertz optical asymmetric demultiplexer (TOAD). IEEE Photonics Technology Letters 5 (7), 787–790.
Stavdas, A. (Ed.), 2001. New Trends in Optical Network Design and Modelling. Dordrecht: Kluwer Academic.
Ueno, Y., Takahashi, M., Nakamura, S., Suzuki, K., Shimizu, T., Furukawa, A., Tamanuki, T., Mori, K., Ae, S., Sasaki, T., Tajima, K., 2003. Control scheme for optimizing the interferometer phase bias in the symmetric-Mach–Zehnder all-optical switch (invited paper). Joint special issue on recent progress in optoelectronics and communications. IEICE Trans. Electron. E86-C (5), 731–740.

Pulse Characterization Techniques

DJ Kane, Southwest Sciences, Inc., Santa Fe, NM, USA

Nomenclature

$A(t)$ Full electric field
$E(t)$ Complex electric field without carrier frequency
$E_{sig}(t, \tau)$ Complex electric field of the FROG trace versus time and time delay
$E_{sig}(t, \Omega)$ Fourier transform of $E_{sig}(t, \tau)$ with respect to τ
$\tilde{E}(\omega)$ Complex electric field in the frequency domain
$f(t)$ Correlation product
$f(\omega)$ Fourier transform of the correlation product
$g(t-\tau)$ Time gate or filter
G FROG trace error
$h(\omega-\Omega)$ Frequency gate or filter
$I(t)$ Intensity profile of the electric field in the time domain
$\tilde{I}(\omega)$ Spectral intensity
$I_{FROG}(\omega, \tau)$ Intensity of the FROG trace
$\Theta(t)$ Heaviside FUNCTION$=0\ t<0, =1\ t \geqq 0$
$\phi(\omega)$ Frequency domain phase
$S_E(\omega, \tau)$ Spectrogram
$S_E(\Omega, T)$ Sonogram
$S(\omega)$ Spectral interferogram
t, T Variables for time
τ Fourier transform
v_g Group velocity
$\varphi(t)$ Time domain phase
Δw Beam waist
Z Distance metric between the FROG signal field generated by the data constraint and the FROG signal field generated by the mathematical form constraint
Φ Geometric angle between two beams
$\Delta\phi$ Phase difference
$\prod^{(n)}$ nth order nonlinear susceptibility
ω Primary variable for frequency
$\Delta\omega$ Difference in center frequency between two pulses
Ω Secondary variable for frequency
$\Omega(t)$ Instantaneous frequency

Introduction

Ultrashort optical pulses have very short time durations, typically less than a few tens of picoseconds. As a result, these pulses are spectrally broad. Because the index of refraction of materials is a function of wavelength, different wavelengths of light travel at different speeds in optical materials, causing the properties of ultrashort optical pulses to change as they propagate. However, the shape of the pulse can influence how the pulse itself interacts with materials. Thus, the goal of ultrafast laser pulse measurement is to obtain not only the intensity profile of the pulse, but also the actual variation of the frequencies that make up the pulse. This is called measuring the intensity and phase of the pulse, respectively.

Even though the development of techniques to characterize ultrashort optical pulses has not been easy, a myriad of pulse measurement techniques have been developed. In this chapter, we will confine our discussion to the most well-known pulse and generally accepted pulse characterization methods in an effort to provide a basic working knowledge of pulse measurement techniques and a fundamental understanding of the principles behind pulse measurement. After a brief discussion of the mathematical representation of ultrashort optical pulses, we will discuss pulse measurement in (roughly) chronological order, starting with autocorrelation methods. The next section introduces a useful method for the measurement of the relative phase between two pulses called spectral interferometry (SI). Following SI, we introduce the notion of the time-frequency representation of ultrashort optical pulses where techniques such as frequency-resolved optical gating (FROG) are discussed. Following the time-frequency section, a relatively new technique known as spectral phase interferometry for direct electric field reconstruction (SPIDER) is discussed.

Mathematical Representation of an Optical Pulse

The time-dependent variations of the pulse are embodied in the pulse electric field, $A(t)$, which can be written:

$$A(t) = \mathrm{Re}\left[E(t)e^{i\omega_0 t}\right] \tag{1}$$

where ω_0 is the carrier frequency and Re refers to the real part. While we can use $A(t)$ as it stands for calculations, it is much easier to remove the rapidly varying ω_0 part, $e^{i\omega_0 t}$, and use a slowly varying envelope together with a phase term that contains only the frequency variations, not the rapidly varying carrier frequency:

$$E(t) = [I(t)]^{1/2} e^{i\varphi(t)} \tag{2}$$

where $I(t)$ and $\varphi(t)$ are the time-dependent intensity and phase of the pulse. (Note that $E(t)$ is complex.) The frequency variation, $\Omega(t)$, is the derivative of $\varphi(t)$ with respect to time:

$$\Omega(t) = -\,\mathrm{d}\varphi(t)/\mathrm{d}t \tag{3}$$

 doi:10.1016/B978-0-12-809283-5.00842-9

The pulse field can be written equally well in the frequency domain by taking the Fourier transform of Eq. (2):

$$\tilde{E}(\omega) = \left[\tilde{I}(\omega)\right]^{1/2} e^{-i\phi(\omega)} \tag{4}$$

where $\tilde{I}(\omega)$ is the spectrum of the pulse, and $\phi(\omega)$ is its phase in the frequency domain. The spectral phase contains time versus frequency information; that is, the derivative of the spectral phase with respect to frequency yields the time arrival of the frequency – the group delay.

Obtaining the intensity and phase, $I(t)$ and $\varphi(t)$ (or $\tilde{I}(\omega)$ and $\phi(\omega)$ is called full characterization of the pulse. Common phase distortions include linear chirp, where the phase (either the time domain or frequency domain) is parabolic. When the frequency is increasing in time, the pulse is said to have positive linear chirp; negative linear chirp is when the high frequencies lead the lower frequencies. Higher-order chirps are common, but for these, differentiation between spectral and temporal chirp is required because spectral phase and temporal phase are not interchangeable.

Autocorrelation

Traditionally, we measure events using shorter events. Unfortunately, for the ultrafast researcher, shorter events do not exist and modern electronics are not fast enough. Therefore, because the shortest event we have is the event we wish to measure, traditional pulse measurement methods use the pulse itself to determine the approximate duration of the pulse in question giving birth to the ubiquitous intensity autocorrelation (see Fig. 1). While autocorrelations cannot be used to fully characterize ultrashort optical pulses, the methods used in autocorrelations are fundamental to all pulse measurement schemes.

The intensity autocorrelation is measured by combining a pulse and a delayed replica of a pulse in a nonlinear medium such as a second harmonic generation (SHG) crystal. A pulse is sent onto a beamsplitter to produce the two replicas. One pulse is delayed relative to the other and both are focused together into a SHG crystal. As the delay of one pulse relative to the other is varied, the intensity of the second-harmonic signal is recorded. The intensity autocorrelation is not limited to using SHG; any nonlinearity may be utilized. Indeed, two-photon absorption in a semiconductor LED or photodiode often acts as a convenient nonlinearity. Sometimes a third-order nonlinearity is used to provide some direction of time information about the pulse. When the generation of the signal involves phase matching, such as second harmonic generation, care must be taken to use a thin crystal. Typically, 10 μm to 1 mm thick SHG crystals are used, depending on the bandwidth of the pulse to be measured. The exact thickness depends on the SHG crystal, the pulse width, and the requirements of the measurement.

Regardless of the nonlinearity used, an autocorrelation yields only the approximate duration and shape of the pulse. The structure of the pulse is smeared, producing a smooth, featureless profile for even a complex pulse. Thus, the intensity autocorrelation alone does not determine the intensity profile of the pulse, $I(t)$. Only the interaction of the intensity of the pulse is recorded, producing no phase information for the pulse. Some qualitative information can be gleaned from the spectrum and the autocorrelation, but usually only that the pulse is chirped, and higher-order chirps and complex pulse structures elude such an analysis.

Another useful form of intensity autocorrelation is the single-shot autocorrelator. In this case, fairly large beams are used, and these beams are set to intersect at an angle of 2Φ, tilting the pulse fronts, so that delay is mapped onto a spatial coordinate (see Fig. 2). The beams are then focused into a nonlinear medium using a cylindrical lens. The interaction region in the SHG crystal is imaged onto a linear array or CCD camera. Typically, when second-harmonic generation is used, the SHG crystal is oriented for Type I phase matching. The time window is proportional to $\Delta w(\sin \Phi)/\Delta v_g$, where Δw is the beam waist at the SHG crystal, and v_g is the group velocity of the pulse in the crystal. Like all SHG interactions, care must be exercised to insure that the phase matching is sufficient to mix the entire bandwidth of the pulse.

Jean-Claude Diels improved the intensity autocorrelation by the development of the interferometric autocorrelation. In this configuration, a Michelson interferometer is typically used so that the beams are collinear when arriving at the SHG crystal, which allows the beams to interfere. As a result, fringes appear on the autocorrelation. By examining the shape and extent of the fringes,

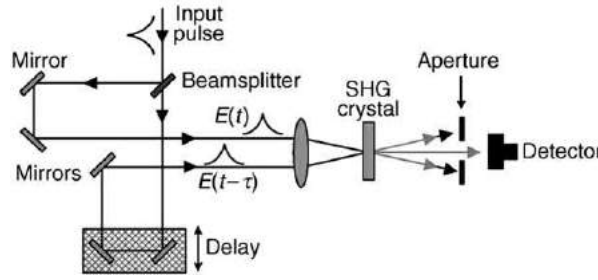


Fig. 1 An intensity autocorrelator used to determine the approximate duration of a pulse. The pulse is split into two replicas that are sent into delay lines. The outputs from both delay lines are focused into a doubling crystal. The second harmonic output is monitored as a function of delay between the two pulses. A plot of the signal versus time is the intensity autocorrelation of the pulse.

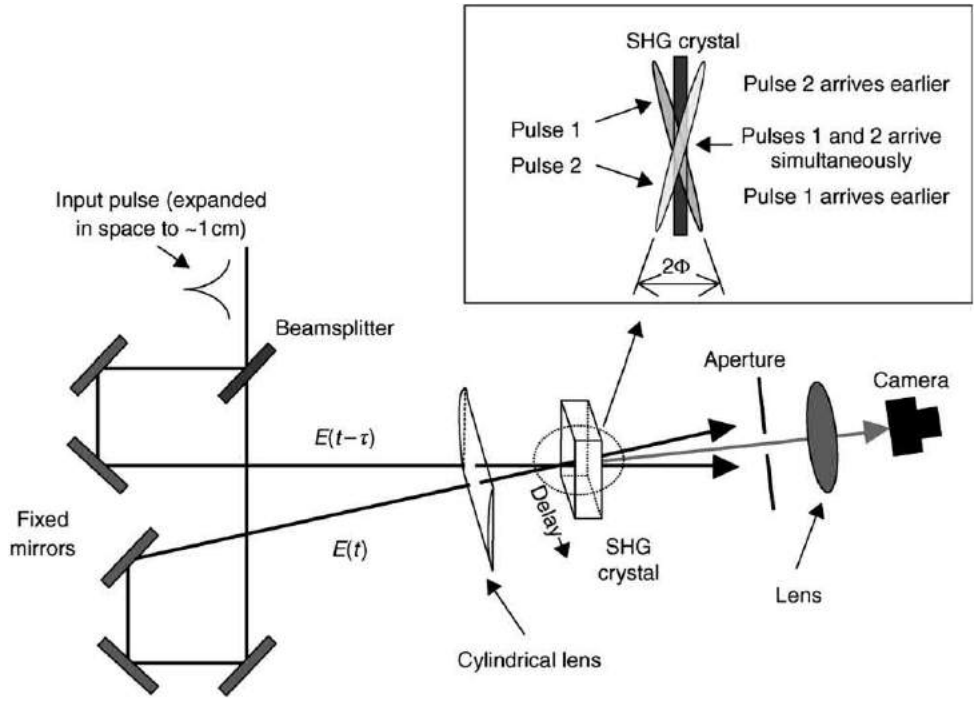


Fig. 2 A single shot intensity autocorrelator. For this type of autocorrelator, the beam size is large, on the order of 1 cm in diameter, and the pulse fronts are tilted with respect to each other in order to map relative delay to position. The two beams are focused using a cylindrical lens and the output is recorded using an array detector. The inset shows how delay is mapped into position. Inset: the tilted pulse fronts cause time delay to be mapped spatially.

some information about the pulse chirp can be obtained. While this does add to the information provided by an intensity autocorrelation, the added information is only the spectrum of the second harmonic. Unfortunately the addition of the second harmonic spectrum does not provide enough information for full retrieval of the intensity and phase.

Even though autocorrelation alone does not completely determine pulse intensity and phase, it is a very simple and useful technique to determine approximate pulse duration. Sometimes, cross-correlation can be used to determine the duration of a long pulse, if a shorter one is available, or a pulse too weak (or at an inappropriate wavelength) to measure with autocorrelation. The addition of the spectrum and/or the second harmonic spectrum can provide more, albeit, incomplete information about pulse chirp. More importantly, the experimental techniques used by correlation methods are fundamental to all pulse measurement methods.

Spectral Interferometry – Relative Phase Measurements

While determination of the absolute intensity and phase of an ultrashort laser pulse is difficult, determination of the relative phase is not. Invented in the late 1800s, spectral interferometry (SI), measures the relative phase between two pulses. SI, a linear technique, can also be modified (albeit, a some-what complex modification) to be self-referencing for full characterization of the ultrashort optical pulse (See the section on SPIDER below). In addition, when combined with a full pulse characterization technique, SI, because it is a linear technique, provides a method by which very weak optical pulses can be characterized (see below).

Typically, an SI apparatus uses a Mach–Zender interferometer (see **Fig. 3**). A pulse is split into two replicas; one is sent through a region or medium to measure and the other through a known path. The two pulses remain separated by some known time and are sent, collinearly, into a spectrometer; no time scanning is required. An analysis of the fringe pattern yields the relative phase between the two pulses. In other words, SI provides $\phi_{\text{unk}}(\omega) - \phi_{\text{ref}}(\omega)$, where $\phi_{\text{unk}}(\omega)$ is the unknown pulse phase versus frequency. The spectrum of the two pulses in the frequency domain is

$$\tilde{I}_{\text{SI}}(\omega) = |\tilde{E}_0(\omega) + \tilde{E}_{\text{unk}}(\omega)|^2$$

$$\tilde{I}_{\text{SI}}(\omega) = \tilde{I}_0(\omega) + \tilde{I}_{\text{unk}}(\omega) + 2\sqrt{\tilde{I}_0(\omega)}\sqrt{\tilde{I}_{\text{unk}}(\omega)}\cos(\phi_{\text{unk}}(\omega) - \phi_0(\omega) - \omega\tau) \tag{5}$$

where $\tilde{I}_{\text{SI}}(\omega)$ is the output spectrum, $\tilde{E}_0(\omega)$ is the reference pulse electric field, $\tilde{E}_{\text{unk}}(\omega)$ is the unknown pulse electric field, $\tilde{I}_0(\omega)$ is the reference pulse spectrum, $\tilde{I}_{\text{unk}}(\omega)$ is the unknown pulse spectrum, $\phi_{\text{unk}}(\omega)$ is the unknown pulse phase, $\phi_0(\omega)$ is the reference

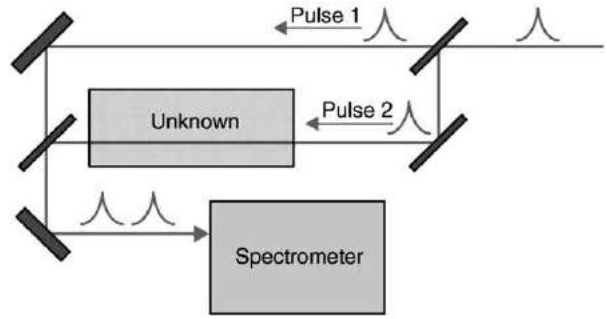


Fig. 3 A Mach–Zender interferometer is the typical arrangement for spectral interferometry. One path acts as the reference arm and the sample to be measured is placed in the other arm. Unlike standard interferometry, the delay is fixed in a spectral interferometry experiment.

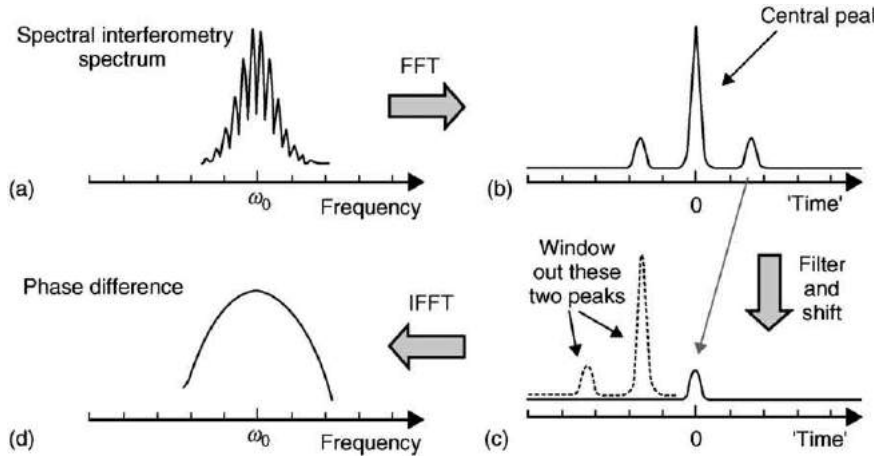


Fig. 4 The steps required for the analysis of spectral interferograms. Part (a) shows a sample SI interferogram. The dominant frequency of the fringes is the delay multiplied by the speed of light. The actual phase differences of interest are the perturbations on this frequency. Part (b) is the Fourier transform of the SI interferogram in Part (a). (Note that this is not the true time domain.) The central peak contains only spectral information. All of the phase information is contained in the satellite peaks. The next step is to mask out the central peak and one of the satellite peaks (Part (c)). The phase of the inverse Fourier transform of Part (c) yields the relative phase between the two pulses (Part (d)).

pulse phase, and τ is the delay between the two pulses. To facilitate extracting the phase difference from the SI output, τ is chosen to yield fringes in the sum spectrum. The spectral fringes, which have a period inversely proportional to the optical path difference between the two beams (see Fig. 4), contain all of the phase difference information.

Fig. 4 shows the steps required to obtain the phase difference between the two pulses. The first step is to subtract out the spectra of the individual pulses in order to isolate the spectral interferogram, $S(\omega)$ (Fig. 4(a)), where

$$S(\omega) = 2\sqrt{\tilde{I}_0(\omega)}\sqrt{\tilde{I}_{\text{unk}}(\omega)}\cos(\phi_{\text{unk}}(\omega) - \phi_0(\omega) - \omega\tau) \tag{6}$$

By Fourier transforming $S(\omega)$, we obtain

$$\mathfrak{F}^{-1}[S(\omega)] = f(t-\tau) + f(-t-\tau) \tag{7}$$

where $f(t)$ is the correlation product between the reference and the unknown electric field (see Fig. 4(b)). The Fourier transform of Eq. (7) multiplied by a Heaviside function, $\Theta(t)$ (to remove $f(-t-\tau)$, recovers the amplitude and phase of $f(\omega)$, the Fourier transform of $f(t-\tau)$, which contains the relative spectral phase between the unknown pulse and the unmodified pulse. Some care, however, needs to be taken to correctly remove the linear phase due to the time delay between the two pulses, $\omega\tau$. Also, for best results, the spectrometer should be well calibrated and care should be taken to properly window the spectrum before taking the Fourier transform.

Spectral interferometry is a straightforward and simple technique that always provides a relative phase. For this very reason, care must be taken to make sure the beams are mode-matched as well as collinear; also, interferometric stability must be maintained over the course of the measurement. Furthermore, SI is nongating; that is, CW background in the laser can add to the interferogram, masking transient effects. Nevertheless, spectral interferometry is a useful technique that has been applied to the measurement of the linear and nonlinear spectral phase introduced by optical fibers. More recently, SI has been applied to phase-locking and to phase-resolved pump probe experiments.

Time-Frequency Representation

In 1971, E.B. Treacy laid the foundations for a revolution in pulse measurement by introducing the idea of measuring the intensity versus time for different spectral splices of a pulse. Because his measurements provided both time and frequency information simultaneously, these measurements could be thought of as applying in a hybrid time-frequency domain. At first, this may seem confusing, but similar methods have been used to visualize sounds patterns (a musical score, for example) and speech. A time-frequency plot of this type, where spectral slices of a pulse are plotted versus time, is called a sonogram. The mathematical formalism of a sonogram is:

$$S_{\mathrm{E}}(\Omega, T) = \left| \int_{-\infty}^{\infty} \tilde{E}(\omega) h(\omega - \Omega) \mathrm{e}^{-\mathrm{i}\omega T} \mathrm{d}\omega \right|^2 \tag{8}$$

where $\tilde{E}(\omega)$ is the electric field of the pulse to be measured in the frequency domain and $h(\omega - \Omega)$ is a frequency gate that varies with frequency. The magnitude squared of the inverse Fourier transform of spectral slices yields the sonogram.

In 1991, Chilla and Martinez showed that the sonogram could be used to reconstruct the full intensity and phase of the pulse. That is, under certain conditions, the approximate group delay could be determined as a function of frequency from the sonogram by finding the peak time arrival of each spectral slice; integration of the group delay yields the spectral phase, which together, with the pulse spectrum, provides the intensity and phase of the pulse in the frequency domain. They labelled this technique as frequency domain phase measurement (FDPM).

An experimental diagram of an FDPM apparatus is shown in Fig. 5. The pulse is split into two replicas, and one of the pulse replicas is spectrally filtered. The spectrally filtered pulse is cross-correlated with the original pulse to find the 'time arrival' of the spectrally filtered pulse – defined as the peak of the cross-correlated pulse. Because the spectrally filtered pulse has a much longer time duration than the original pulse, the small perturbation caused by the finite length of the original pulse is neglected.

The main difficulty of the Chilla–Martinez method is that the frequency gate must be very narrow, reducing the filtered pulse energy significantly, which greatly reduces the measured signal strength. If the spectral phase is not well behaved, the filtered pulse may not be much longer than the original pulse. In addition, for a given frequency, the sonogram may have two or more peaks in time, that is, the group delay may not be a function. However, a method called 2D phase retrieval, discussed in the next section, can alleviate many of these issues.

A spectrogram is a relative of the sonogram. Rather than determining the time arrival of spectral slices of a pulse, a spectrogram is obtained when the spectrum of time slices of the pulse are measured. The spectrogram is experimentally much easier to obtain than the sonogram; however, the measured quantity is not quite as mathematically useful as the sonogram because the time slicing of the pulse occurs in the time domain. Thus, if a narrow time gate is used, the time domain phase is obtained. The time domain phase together with the intensity of the pulse, $I(t)$, provides the full intensity and phase of the pulse. Unfortunately, the intensity profile is not as readily measurable as the pulse spectrum. Consequently, to obtain the full intensity and phase of a pulse from its spectrogram requires an iterative 2D phase retrieval algorithm. This is the basis of a pulse measurement technique called frequency-resolved optical gating (FROG).

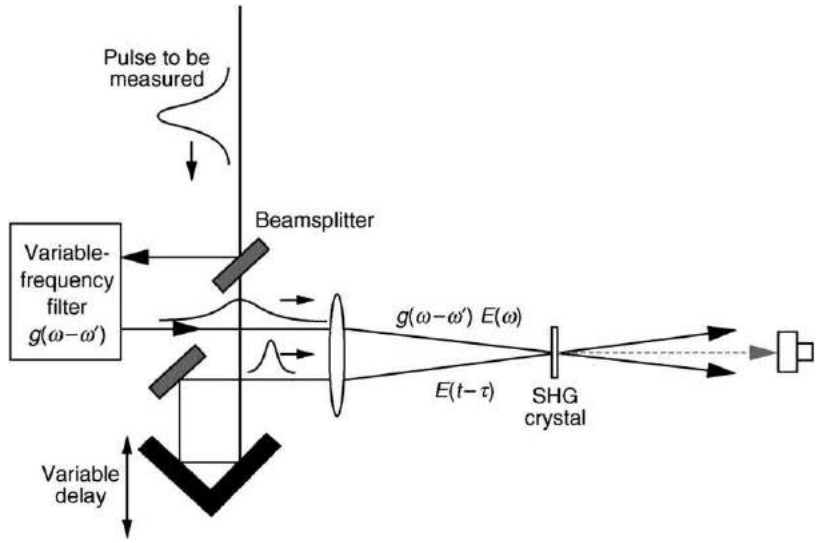


Fig. 5 Measuring a sonogram requires determining the time arrival of spectral slices of the pulse. The pulse is split into two replicas using a beamsplitter. One replica is spectrally filtered using a tunable filter. The other pulse is cross-correlated with the spectrally filtered pulse to determine the time arrival.

Frequency-Resolved Optical Gating

Frequency-resolved optical gating (FROG), developed by Kane and Trebino, measures the spectrum of a particular temporal component of the pulse (see **Figs. 6** and **7**) by spectrally resolving the signal pulse in an autocorrelation-type experiment using an instantaneously responding nonlinear medium. As shown in **Fig. 6**, FROG involves splitting a pulse and then overlapping the two

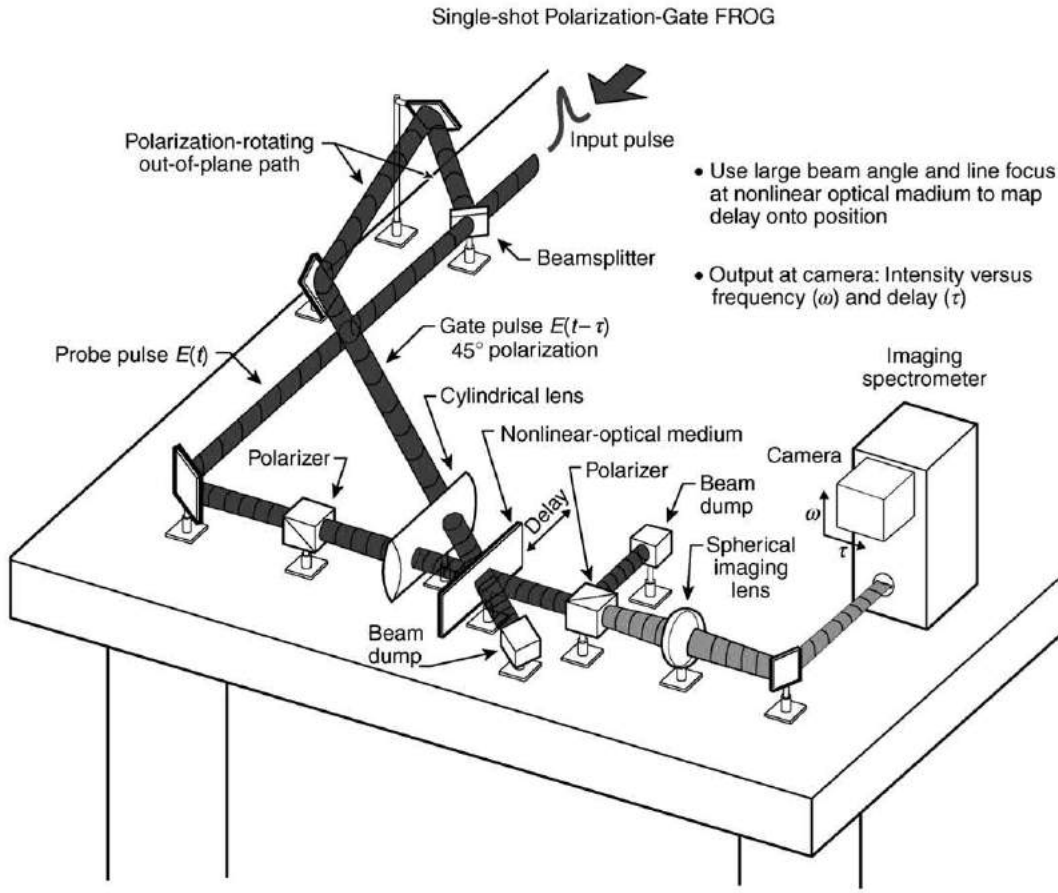


Fig. 6 Measuring the spectrogram of a pulse is easier than measuring its sonogram – a spectrogram is a spectrally resolved autocorrelation. In this figure, the optical Kerr–effect is used (polarization-gate) as the nonlinearity. Adapted with permission from Kane DJ and Trebino R (1993) Single shot measurement of the intensity and phase of an arbitrary ultrashort pulse by using frequency-resolved optical gating. *Optics Letters* 18(10): 823–825.

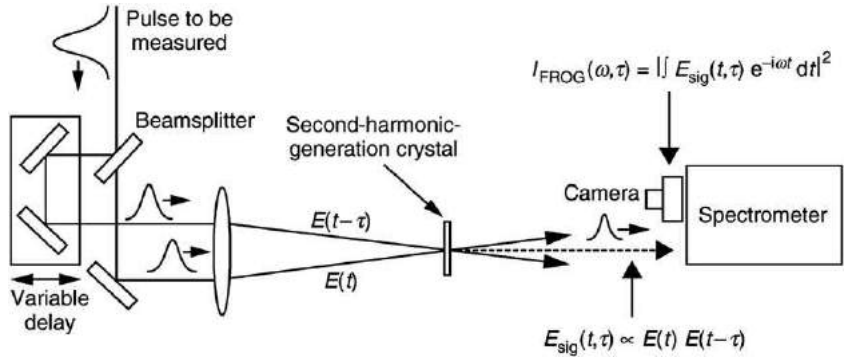


Fig. 7 This figure shows a schematic of an SHG FROG device. The SHG signal from the autocorrelation is spectrally resolved. SHG FROG is very simple and sensitive, but it has a direction-of-time ambiguity. For example, if the pulse has chirp, only the magnitude of the chirp is determined – the sign of the chirp remains unknown.

resulting pulses in an instantaneously responding $\chi^{(3)}$ or $\chi^{(2)}$ medium. Even though any instantaneous nonlinear interaction may be used to implement FROG, perhaps the most intuitive is the polarization-gating configuration. In this case, induced birefringence, due to the electronic Kerr effect, is used as the nonlinear-optical process. In other words, the 'gate' pulse causes the $\chi^{(3)}$ medium, which is placed between two crossed polarizers, to become slightly birefringent. The polarization of the 'gated' probe pulse (which is cleaned up by the first polarizer) is rotated slightly by the induced birefringence allowing some of the 'gated' pulse to leak through the second polarizer. This is referred to as the signal. Because most of the signal emanates from the region of temporal overlap between the gate pulse and the probe pulse, the signal pulse contains the frequencies of the 'gated' probe pulse within this overlap region. The signal is then spectrally resolved, and the signal intensity is measured as a function of wavelength and delay time τ. The resulting trace of intensity versus delay and frequency is a spectrogram, a time- and frequency-resolved transform that intuitively displays time-dependent spectral information of a waveform.

The spectrogram can be expressed as:

$$S_{\mathrm{E}}(\omega,\tau)=\left|\int_{-\infty}^{\infty}E(t)g(t-\tau)\mathrm{e}^{-\mathrm{i}\omega t}\mathrm{d}t\right|^2 \tag{9}$$

where $E(t)$ is the measured pulse's electric field, $g(t-\tau)$ is the variable-delay gate pulse, and the subscript E on S_{E} indicates the spectrogram's dependence on $E(t)$. The gate pulse $g(t)$ is usually somewhat shorter in length than the pulse to be measured, but not infinitely short. This is an important point: an infinitely short gate pulse yields only the intensity $I(t)$ and conversely, a CW gate yields only the spectrum $I(\omega)$. On the other hand, a finite-length gate pulse yields the spectrum of all of the finite pulse segments with duration equal to that of the gate. While the phase information remains lacking in each of these short-time spectra, having spectra of an infinitely large set of pulse segments compensates for this loss. The spectrogram has been shown to nearly uniquely determine both the intensity $I(t)$ and phase $\varphi(t)$ of the pulse, even if the gate pulse is longer than the pulse to be measured (although if the gate is too long, sensitivity to noise and other practical problems arise).

In FROG, when using optically induced birefringence as the nonlinear effect, the signal pulse is given by:

$$E_{\mathrm{sig}}(t,\tau)\propto E(t)|E(t-\tau)|^2 \tag{10}$$

So the measured signal intensity $I_{\mathrm{FROG}}(\omega,\tau)$, after the spectrometer is:

$$I_{\mathrm{FROG}}(\omega,\tau)=\left|\int_{-\infty}^{\infty}E(t)|E(t-\tau)|^2\mathrm{e}^{-\mathrm{i}\omega t}\mathrm{d}t\right|^2 \tag{11}$$

We see that the FROG trace is a spectrogram of the pulse $E(t)$ although the gate, $|E(t-\tau)|^2$, is a function of the pulse itself.

To see that the FROG trace essentially uniquely determines $E(t)$ for an arbitrary pulse, it is first necessary to observe that $E(t)$ is easily obtained from $E_{\mathrm{sig}}(t,\tau)$. Then it is simply necessary to write **Eq. (11)** in terms of $E_{\mathrm{sig}}(t,\Omega)$, the Fourier transform of the signal field $E_{\mathrm{sig}}(t,\Omega)$, with respect to delay variable τ. We then have what appears to be a more complex expression, but one that will give us better insight into the problem:

$$I_{\mathrm{FROG}}(\omega,\tau)=\left|\int_{-\infty}^{\infty}E_{\mathrm{sig}}(t,\Omega)\mathrm{e}^{-\mathrm{i}\omega t-\mathrm{i}\Omega\tau})\mathrm{d}t\mathrm{d}\Omega\right|^2 \tag{12}$$

Eq. (12) indicates that the problem of inverting the FROG trace $I_{\mathrm{FROG}}(\omega,\tau)$, to find the desired quantity $E_{\mathrm{sig}}(t,\Omega)$, is that of inverting the squared magnitude of the two-dimensional (2D) Fourier transform of $E_{\mathrm{sig}}(t,\Omega)$. This problem, which is called the 2D phase-retrieval problem, is well known in many fields, especially in astronomy, where the squared magnitude of the Fourier transform of a 2D image is often measured. At first glance, this problem appears unsolvable; after all, much information is lost when the magnitude is taken. It is well known that the one-dimensional (1D) phase retrieval problem is unsolvable (for example, infinitely many pulse fields give rise to the same spectrum). Intuition fails in this case, however; two- and higher-dimensional phase retrieval essentially always yields unique results.

The Simplified FROG – GRENOUILLE

A simplified version of the FROG device, known as grating eliminated no-nonsense observation of ultrafast incident laser light e-fields (GRENOUILLE), can be constructed using the doubling crystal as both the gating medium and the spectrometer (see **Fig. 8**). In this case, the phase matching condition in certain thick doubling crystals causes a wavelength dependent output from the doubling crystal. A cylindrical lens is used in a Fourier transform configuration to image the wavelength variation onto a CCD camera. A cylindrical lens set at 90 degrees to the Fourier transform lens images the time axis onto the CCD camera. A Fresnel biprism replaces the traditionally used Mach–Zender interferometer to produce two pulses propagating at an angle with respect to each other. Presently, the GRENOUILLE technique works with BBO, in the wavelength region around 500 nm–1200 nm, and with proustite in the wavelength region around 1200 nm–2000 nm.

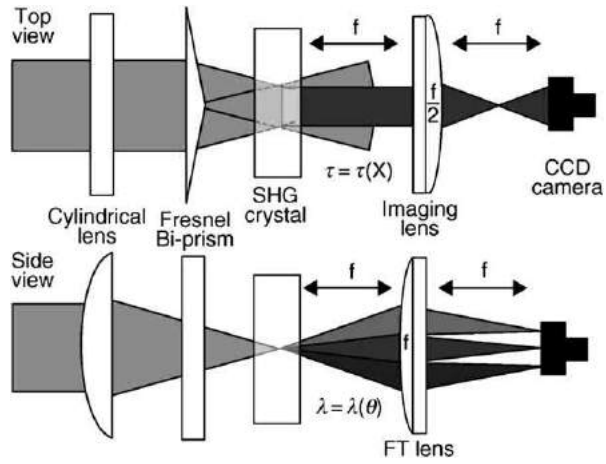


Fig. 8 Fig. showing a schematic of a GRENOUILLE FROG device which uses a thick doubling crystal as both the gating nonlinearity and the spectrometer. Like single-shot autocorrelators, a cylindrical lens focuses a spatially large input beam into an SHG crystal in only one dimension. Unlike most autocorrelators, the Mach–Zender interferometer is replaced with a Fresnel biprism that forces each half of the input beam to propagate, at an angle, toward the SHG crystal. An imaging cylindrical lens images the spatial dependence of the delay onto a CCD camera. A Fourier transform cylindrical lens maps the angular dependence of wavelength from the SHG crystal to a spatial coordinate on the CCD camera. Adapted with permission from O'Shea P, Kimmel M, Xun G and Trebino R (2001) Highly simplified device for ultrashort-pulse measurement. *Optics Letters* 26(12): 932–934.

FROG Inversion Algorithms

An iterative 2D phase retrieval algorithm is required to extract the pulse information from the measured FROG trace (see **Figs. 9** and **10**). This algorithm converges to a pulse that minimizes the difference between the measured and the calculated FROG trace. While this aspect of FROG has been its Achilles heel in the past, in reality, new generalized projections algorithms (together with faster computers) converge quickly and can track pulse changes at rates up to 30 Hz and beyond, making FROG a true, real-time pulse measurement technique. Indeed, excellent algorithms for the analysis of FROG traces in real-time are commercially available.

The original FROG inversion algorithm, commonly referred to as the vanilla algorithm, is simple and iterates quickly, but tends to stagnate, giving erroneous results, especially for geometries that use a complex gate function such as SHG or self-diffraction. An improved algorithm, incorporating several different algorithms including brute force minimization, was developed to alleviate stagnation programs at the expense of both speed and convergence time. By the mid-1990s, a significant advance in both speed and stability was made by the addition of a numerical method called generalized projections. This algorithm determines the next guess by constructing a projection – minimizing the error (distance) between the FROG electric field, $E_{\mathrm{sig}}(t, \tau)$, obtained immediately after the application of the intensity constraint, and the FROG electric field calculated from the mathematical form constraint.

The first generalized projections algorithm used a standard minimization procedure to find the electric field for the next iteration, which can still be slow. For the most common FROG geometries, PG and SHG, a different algorithm that determines the next iteration directly has been developed. This algorithm, called the principal components generalized projections (PCGP) algorithm, converts the generalized projections algorithm into an eigenvector problem. Using this algorithm, pulse measurement rates of at least 30 Hz have been achieved.

The goal of the phase retrieval algorithm is to find the $E(t)$ that satisfies two constraints. The first is the FROG trace itself which is the magnitude squared of the 1D Fourier transform of $E_{\mathrm{sig}}(t, \tau)$:

$$I_{\mathrm{FROG}}(\omega, \tau) = \left| \int_{-\infty}^{\infty} E_{\mathrm{sig}}(t, \tau) \mathrm{e}^{-\mathrm{i}\omega t} \mathrm{d}t \right|^2 \tag{13}$$

The other constraint is the mathematical form of the signal field, $E_{\mathrm{sig}}(t, \tau)$, for the nonlinear interaction used. The signal forms for a variety of FROG beam geometries are:

$$E_{\mathrm{sig}}(t, \tau) \propto \left\{ \begin{array}{l} E(t)|E(t-\tau)|^2 \text{PG FROG} \\ E(t)^2 E^*(t-\tau) \text{ SD FROG} \\ E(t)E(t-\tau) \quad \text{SHG FROG} \\ E(t)^2 E(t-\tau) \quad \text{THG FROG} \end{array} \right\} \tag{14}$$

where PG is polarization gate, SD is self-diffraction, SHG is second harmonic generation, and THG is third harmonic generation FROG.

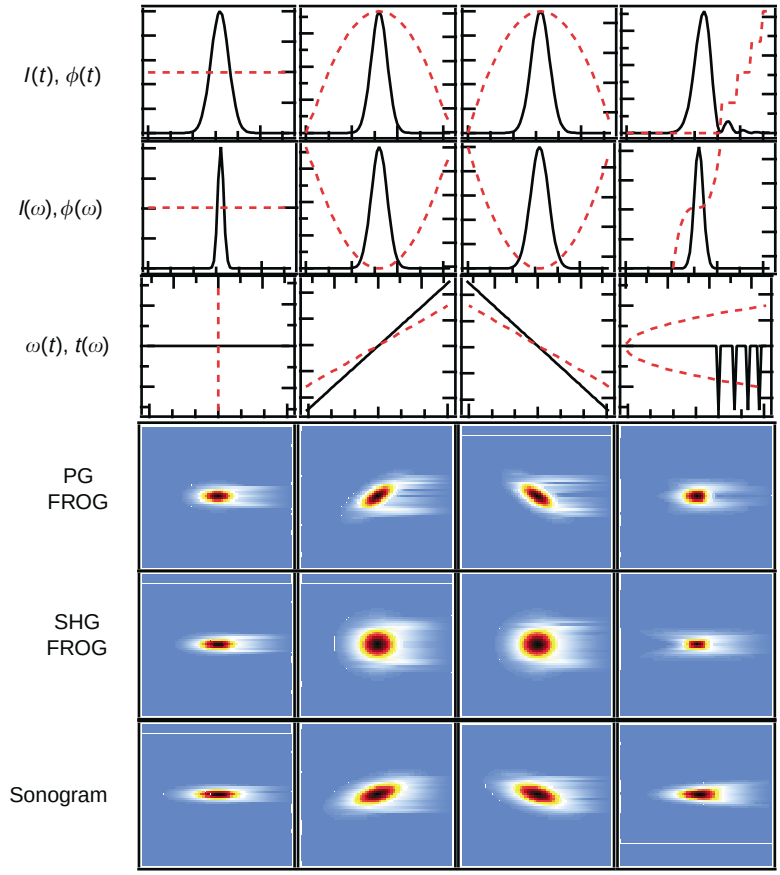


Fig. 9 Example FROG traces and sonograms are shown for four different pulses: a transform limited pulse, a positively chirped pulse, and a negatively chirped pulse. The top series of plots are the time domain representation of the pulse intensity (solid line) and phase (dashed line). The second row of four plots is the frequency domain representation of the sample pulse's intensity and phase. The next four plots show the instantaneous frequency and the group delay. The vertical axis is frequency and the horizontal axis is time. The remaining rows show the PG FROG traces, the SHG FROG traces and sonograms of the pulse shown in the first two rows.

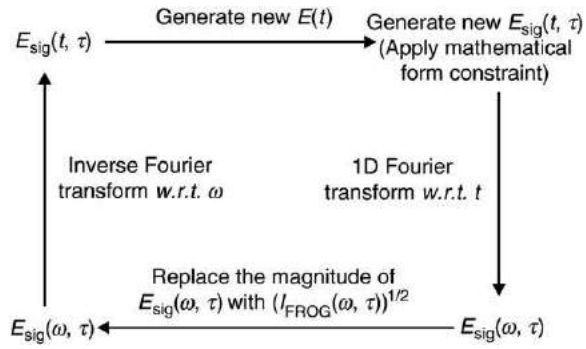


Fig. 10 Phase retrieval algorithm for the inversion of FROG spectrograms. Start with an initial guess for the pulse to generate an initial $E_{sig}(t,\tau)$. A 1D Fourier transform generates the FROG trace. The next step is to replace the magnitude of the calculated FROG trace with the square root of the measured FROG trace. Inverse Fourier transform with respect to ω to produce the new signal field and generate a new guess for $E(t)$. Interestingly, it is only the step that produces $E(t)$ from $E_{sig}(t,\tau)$ that differentiates all the FROG algorithms.

All FROG algorithms work by iterating between two different data sets: the set of all signal fields that satisfy the data constraint, $I_{FROG}(\omega, \tau)$, and the set of all signal fields that satisfy **Eq. (14)**. The difference between the FROG algorithms is how the iteration between the two sets is completed. In the case of generalized projections, the $E(t)$'s are chosen such that the distance between the $E(t)$ on the magnitude set and the $E(t)$ on the mathematical form set is minimized. This is accomplished by minimizing

the following equation:

$$Z = \sum_{i,j=1}^{N} |E^{(k)}_{\mathrm{sig(DC)}}(t_i, \tau_j) - E^{(k+1)}_{\mathrm{sig(MF)}}(t_i, \tau_j)|^2 \tag{15}$$

where $E^{(k)}_{\mathrm{sig(DC)}}(t_i, \tau_j)$ is the signal field generated by the data constraint, and $E^{(k+1)}_{\mathrm{sig(MF)}}(t_i, \tau_j)$ is the signal field produced from one of the beam geometry equations in **Eq. (14)**. For the normal generalized projection, the minimization of Z is completed using a standard steepest descent algorithm; the derivative of Z with respect to the signal field is computed to determine the direction of the minimum. The computation of the derivatives are tedious; they are tabulated elsewhere.

An alternative to an algorithm that minimizes **Eq. (15)** is the PCGP algorithm. This algorithm converts the computation of the next guess into an eigenvector problem, reducing the computation of the next guess to simple matrix–vector multiplications. This algorithm works for both the PG and SHG beam geometries, it is robust and the fastest FROG algorithm. Indeed, the PCGP algorithm was used in the original real-time FROG work by D.J. Kane.

Self Checks in FROG Measurements

Unlike any other pulse measurement techniques, FROG can provide a great deal of feedback about both the quality of the measurement (systematic errors) and the quality of the algorithm's performance. The most common check for convergence is the FROG trace error together with a visual comparison between the retrieved FROG trace and the measured FROG trace. The FROG trace error is given by:

$$G = \sqrt{\frac{1}{N^2} \sum_{i,j=1}^{N} |I_{\mathrm{FROG}}(\omega_i, \tau_j) - \alpha I^{(k)}_{\mathrm{FROG}}(\omega_i, \tau_j)|^2} \tag{16}$$

where α is a renormalization constant, I_{FROG} is the measured trace, and $I^{(k)}_{\mathrm{FROG}}(\omega, \tau)$ is computed from the retrieved electric field. Typically, the FROG trace error of a PG measurement should be less than 2% for a 64×64 pixel trace, while the FROG trace error of a 64×64 pixel SHG FROG trace should be about 1% or less. Acceptable FROG trace errors decrease as FROG trace size increase for smaller FROG traces. These values are only rules of thumb only; for example, acceptable retrievals of large and very complicated FROG traces can produce larger FROG trace errors.

In a good FROG measurement, the spectrum of the retrieved pulse should faithfully reproduce the salient features of the pulse's measured spectrum. SHG FROG even provides an additional check called the frequency marginal. The sum of an SHG FROG trace along the time axis yields the autoconvolution of the pulse's spectrum, providing two ways the FROG measurement can be checked. First, the autoconvolution of an independently measured spectrum can be compared to the sum of the FROG trace along the time axis, providing an indication of how well the measurement was made. For example, if the doubling crystal was too thick in the pulse measurement, the FROG trace's frequency marginal will be narrower than the autoconvolution of the measured spectrum. Second, comparing the autoconvolution of the retrieval pulse spectrum with the FROG trace marginal can provide a test of algorithm convergence in addition to a test of the measurement. Phase matching problems appear as a mismatch between the FROG trace frequency marginal and the autoconvolution of the retrieved spectrum.

Because FROG is a spectrally resolved autocorrelation, summing any FROG trace along the frequency axis yields the autocorrelation of the measured pulse. This autocorrelation can be compared to an independently measured autocorrelation, or a comparison can be made between the frequency sum of the FROG trace and the autocorrelation calculated from the retrieved pulse to determine algorithm convergence and the quality of the measurement.

Measuring Pulses Directly – SPIDER

SPIDER is a novel technique that measures the pulse directly – no iterative phase retrieval method is required. To accomplish this feat, spectral interferometry is conducted on two pulse replicas that are shifted in frequency with respect to one another. The original pulse is split into two replicas. The two pulse replicas are sent into a phase modulator that shifts the center frequency of each pulse slightly. The frequency-shifted pulses are sent into a spectrometer and a standard spectral interferometry analysis yields the derivative of the phase. Integration of the phase derivative, together with the spectrum of pulse, provides the full intensity and phase.

Thus, analysis of the spectral interferogram gives:

$$\Delta\phi = (\phi(\omega_1) - \phi(\omega_2))\Delta\omega \tag{17}$$

where $\Delta\phi$ is the measured phase difference from the spectral interferogram, $\Delta\omega$ is the difference in the center frequency between the two pulses, and $\phi(\omega)$ is the pulse phase as a function of frequency. Consequently, the phase difference must be divided by the frequency difference between the two pulses to determine the true derivative. It is also important to subtract out the linear phase difference resulting from the delay between the two pulses (the delay produces the fringes in the resulting interferogram) as this integrates to a linear chirp.

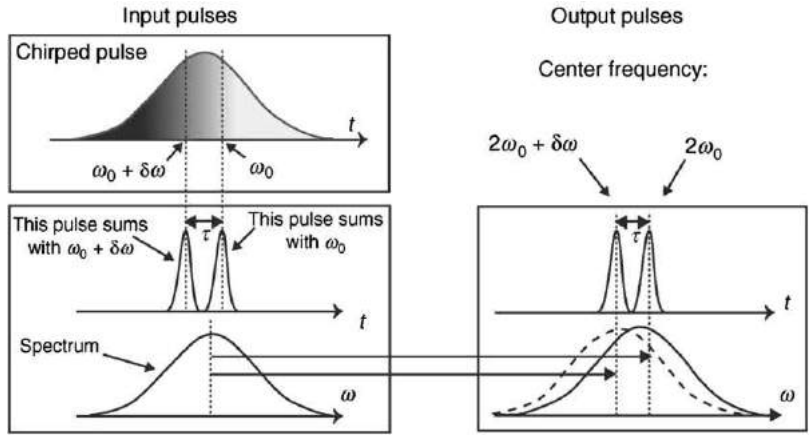


Fig. 11 Fig. showing how a SPIDER device works. The highly chirped pulse has a frequency that varies with time. Mixing the chirped pulse with two time delayed replicas produces two pulses, separated in time, that have slightly different frequencies, but are otherwise identical. By interfering the two pulses in a spectrometer, the relative phase between $\omega_0 + \delta\omega$ portions of the pulse and ω_0 portions of the pulse are determined.

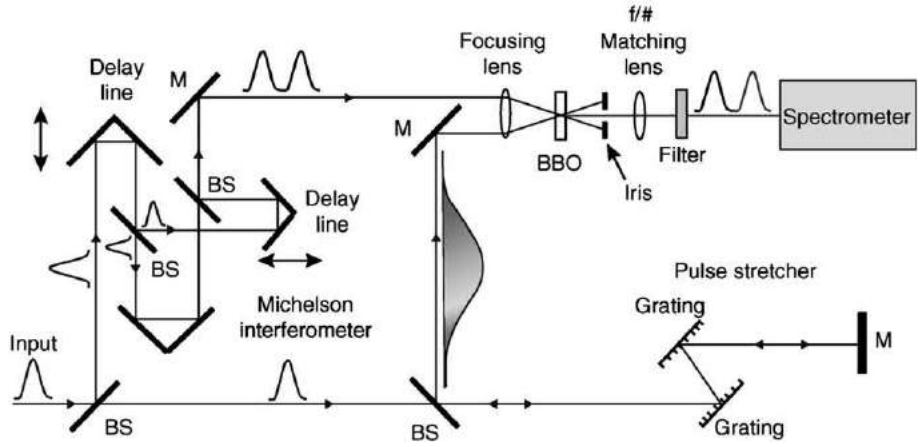


Fig. 12 SPIDER apparatus. The input pulse is first split into two pulses. One pulse is stretched in a pulse stretcher. The other pulse is sent into a Michelson interferometer to produce two time-delayed pulses. The two time-delayed pulses are mixed with the stretched pulse to produce two pulses at a slightly different center frequency. The two pulses are sent into a spectrometer to produce a spectral interferogram. Analysis of the spectral interferogram yields the derivative of the pulse phase.

Because phase modulators are too slow to be used in the femtosecond region, a different method is used to shift the frequency of the two pulses – they are mixed with a highly chirped pulse (see Fig. 11). Because the pulses are slightly time delayed, each one aligns with a slightly different frequency in the chirped pulse. The chirped pulse must be so highly chirped that the pulse frequency does not change significantly over the duration of the pulse to be measured.

A schematic of a SPIDER device for the measurement of femtosecond laser pulses is shown in Fig. 12. The input pulse is first split into two replicas. The first replica is sent to a pulse stretcher to produce to highly chirped pulse. The other replica is sent into a Michelson interferometer (or a simple étalon may sometimes suffice) to create two time-delayed replicas. The time-delayed replicas are mixed with the chirped pulse. Because each replica is mixed with a slightly different frequency in the chirped pulse, each replica produces a pulse with a slightly different center frequency. When the time-delayed replicas are spectrally resolved, each frequency of the pulse interferes with a frequency-shifted portion of the same pulse. Thus, the measured phase difference is proportional to the derivative of the phase of the original pulse.

Measuring Weak Pulses

Because virtually all the pulse measurement techniques, that do not require a reference pulse, use a nonlinearity, measuring weak ultrafast pulses directly is difficult. Fortunately, because most weak pulses are generated from a stronger pulse, spectral interferometry can be used to determine the relative phase between the weaker pulse and the stronger pulse (for regions where there is spectral overlap between the two pulses). If the stronger pulse is characterized using another pulse measurement technique, then

the weaker pulse phase can be determined from the relative phase measurement and the known phase of the stronger pulse. For example, the Trebino group termed the combination of FROG and SI as TADPOLE (Temporal Analysis by Dispersing a Pair Of Light E-fields).

Which Pulse Measurement Method Should I Use?

Choosing a pulse measurement method depends on the required measurement as well as the precision and accuracy required. All the pulse measurement techniques have advantages and disadvantages, but excellent measurements have been made with all of them. Autocorrelation is adequate when all that is required is the approximate pulse duration without any phase information. Spectral interferometry is quite useful for measuring transient changes in a sample in pump-probe experiments. FROG is perhaps the most commonly used and general-purpose pulse measurement technique, providing a great deal of feedback about the quality of the pulse measurement. Inexpensive and convenient FROG devices, together with software, are available commercially, adding to their acceptance. While SPIDER is experimentally complex and does not have the cross checks that FROG has, it provides the quickest answer, requiring no iterative phase retrieval algorithm to obtain pulse intensity and phase.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation under Grant numbers DMI-9801116 and DMI-0091454. The author would also like to acknowledge Rick Trebino for the use of figures from his lectures.

Further Reading

Diels, J.C., Rudolph, W., 1996. Ultrashort Laser Pulse Phenomena. San Diego, CA: Academic Press Inc.

Dorrer, C., Belabas, N., Likforman, J.-P., Joffre, M., 2000. Spectral resolution and sampling issues in Fourier-transform spectral interferometry. Journal of the Optical Society B 17, 1795.

Kane, D.J., Weston, J., Chu, K.-C.J., 2003. Real-time inversion of polarization gate frequency-resolved optical gating spectrograms. Applied Optics 42, 1140–1144.

Iaconis, C., Walmsley, I.A., 1999. Self-referencing spectral interferometry for measuring ultrashort optical pulses. IEEE Journal of Quantum Electronics 35, 501–509.

Lepetit, L., Cheriaux, G., Joffre, M., 1995. Linear techniques of phase measurement by femtosecond spectral interferometry for applications in spectroscopy. Journal of the Optical Society of America B 12, 2467–2474.

Rulliere, C. (Ed.), 2000. Femtosecond Laser Pulses. New York: Springer Verlag.

Shuman, T.M., Anderson, M.E., Bromage, J., Iaconis, C., Waxer, L., Walmsley, I.A., 1999. Real-time SPIDER: ultrashort pulse characterization at 20 Hz. Optics Express 5, 134–143.

Treacy, E.B., 1971. Measurement and interpretation of dynamic spectrograms of picosecond light pulses. Journal of Applied Physics 42, 3848–3858.

Trebino, R., 2000. Frequency Resolved Optical Gating: The Measurement of Ultrashort Laser Pulses. Boston, MA: Kluwer Academic Publisher.

Trebino, R., DeLong, K.W., Fittingholf, D.N., Sweetser, J.N., Krunbugel, M.A., Richman, B.A., Kane, D.J., 1997. Measuring ultrashort laser pulses in the time-frequency domain using frequency-resolved optical gating. Review of Scientific Instruments 68, 3277–3295.

Various authors, 1999. Feature issue on ultrashort-laser-pulse measurement and applications. *IEEE Journal of Quantum Electronics* 35: 4.

Optical Time Division Multiplexing

LP Barry, Dublin City University, Dublin, Ireland

Nomenclature

Dispersion [ps km^{-1} nm^{-1}]
Transmission data rate [bit s^{-1}]
ADM Add drop multiplexer
CW Continuous wave
FWM Four-wave mixing
NOLM Nonlinear optical loop mirror
OTDM Optical time division multiplexing
SDH Synchronous digital hierarchy
SOA Semiconductor optical amplifier
STM Synchronous transport module
TOAD Terahertz optical asymmetric demultiplexer
WDM Wavelength division multiplexing

Introduction

As the demand for bandwidth continues to grow, driven by the massive increase in Internet usage, so will the necessity to have communications networks that can handle very high data rates. The use of optical fiber networks is the clear choice for such systems given the huge available bandwidth of the fiber transmission medium. However, in a basic optical communication system comprising a laser transmitter, an optical fiber transmission medium, and a receiver, the capacity is essentially limited by the speed at which light can be modulated at the transmitter. To overcome this limitation, which is basically due to the speed of available electronics, it is necessary to use optical multiplexing techniques, and one such technique, known as optical time division multiplexing (OTDM), will be the subject of this article.

Principles of Time Division Multiplexing

Time division multiplexing (TDM) has long been the traditional method for electrically combining information channels. If we take the most fundamental data rate to be that of a simple voice call at 64 kbit/s, then the transmission of data at higher bit rates is achieved by electrically multiplexing a large number of 64 kbit/s channels in the time domain. With the evolution of standards, a number of different data rates have been specified as standard transmission rates. In Europe the synchronous digital hierarchy (SDH) has a basic data rate of 155.52 Mbit/s, which is known as the synchronous transport module – Level 1 (STM-1). This particular data rate is essentially obtained by electrically multiplexing over 2000 voice calls in the time domain (with some of the capacity required for overhead information). By subsequently multiplexing a number of STM-1 channels together we can obtain transmission at the higher standard data rates of STM-4 (622 Mbit/s), STM-16 (2.48 Gbit/s), and STM-64 (9.88 Gbit/s), as outlined in **Table 1**.

Fig. 1 illustrates how basic electrical TDM is used in standard optical communication systems. Clearly as we approach the higher data rates of STM-64, a serious level of electrical multiplexing is required, and as the data rates increase so does the cost and complexity of the electrical equipment at the transmitter and receiver. Indeed the present state of the art in electronics seems to suggest 40 Gbit/s as the limit for electrically multiplexed communication systems. However, as we stated earlier, communications traffic has been growing explosively over the last decade and will continue to do so. In order to meet this demand for capacity, and better exploit the massive available bandwidth of optical fiber, it is necessary to use optical multiplexing techniques for communications systems. The two main optical multiplexing techniques available are wavelength division multiplexing (WDM) and optical time division multiplexing (OTDM). WDM essentially involves transmitting data at a number of different wavelengths on the same fiber. Although electrical multiplexing may be limited to data transmission rates of around 40 Gbit/s using one laser, by multiplexing together N different wavelength channels each carrying 40 Gbit/s, we can achieve overall data rates up to and beyond a terabit/s.

The second of these optical multiplexing techniques, OTDM, is the subject of this article. Whereas WDM multiplexes optical data channels in the wavelength domain, the basic principle of OTDM communications is to increase the system

Table 1 Standard data rates for SDH transmission systems

SDH standard	*Data rate (Mbit/s)*
STM-1	155.52
STM-4	622.08
STM-16	2488
STM-64	9953
STM-256	39 813

 doi:10.1016/B978-0-12-809283-5.00674-1

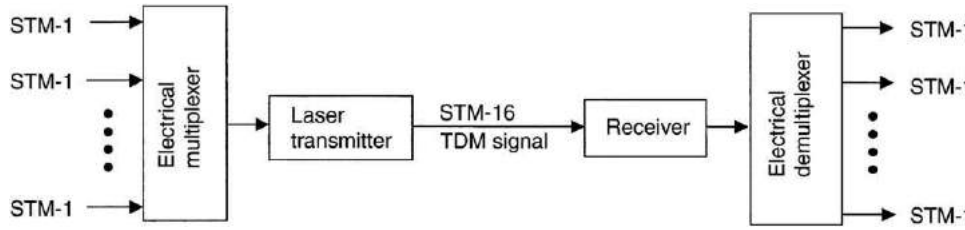


Fig. 1 Electrical time division multiplexing (ETDM) in a basic optical communication system.

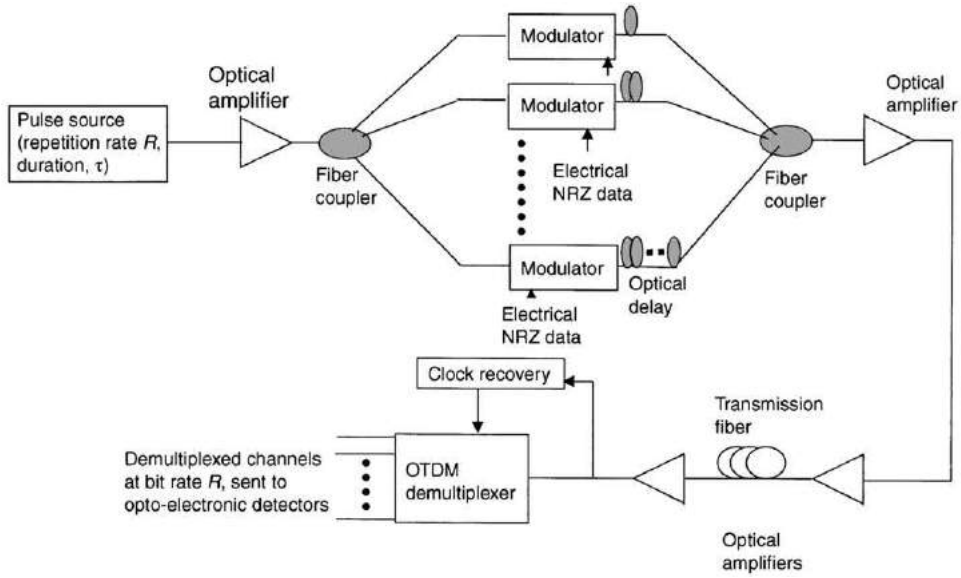


Fig. 2 Basic configuration for an OTDM transmission system comprising transmitter, transmission fiber, and receiver.

capacity by multiplexing optical data channels in the time domain. This multiplexing is usually achieved by a process known as bit-interleaving. This process can be explained by considering Fig. 2, which shows a basic schematic of an OTDM based transmission system. The main component of the overall system is a source of ultrashort optical pulses (pulse duration τ) at a certain repetition rate, R. The optical pulse source is initially split into N channels using a passive fiber coupler, and each pulse train is subsequently modulated by electrical data which is at a data rate of R. The resulting output from each modulator is essentially an optical data channel where the data are represented using ultrashort optical pulses (return-to-zero data format). The data from each modulator then passes through a fixed fiber delay line which delays each channel by a time equal to $1/RN$ relative to its adjacent channel (as shown in Fig. 2). The N modulated and delayed optical data channels are then recombined in another passive fiber coupler to form the OTDM data signal. We can consider that the delay lines essentially assign each data channel to a specific bit slot (of width $1/RN$) in the overall multiplexed signal. The multiplexed data signal may then be transmitted over optical fiber before arriving at the receiver which is responsible for demultiplexing the optical signal into its discrete channels.

The duration of the optical pulse source is extremely important in determining the maximum overall data rate which can be achieved. The overall data rate is basically defined by the temporal separation between channels in the multiplexor, but in order to avoid cross-channel interference, this separation must be significantly greater than the pulse duration. Thus to increase the overall data rate we must use shorter optical pulses. However, as we reduce the optical pulse width to raise the data rate we need to take into account the problems that may be encountered as this high-speed data signal propagates over optical fiber, and also the difficulty in demultiplexing a high-speed OTDM data signal. The following section will look in greater detail at these key elements of the OTDM communication system, namely the optical pulse source, the transmission of the ultrahigh capacity signal, and the demultiplexing at the receiver.

Key Elements of an OTDM Communication System

Ultra-Short Optical Pulse Sources

As stated earlier, the optical pulse source is a key element in any OTDM-based communication system. The important characteristics of the pulse source that will affect its usefulness in an OTDM system are the pulsewidth, the spectral width, and the

temporal jitter. The pulse duration clearly has to be short enough to support the desired overall transmission rate. For example, if we wish to design an OTDM system with an aggregate data rate of 100 Gbit/s (which will have a delay of 10 ps between each channel in the multiplexed signal), then the pulsewidth should normally be less the 30% of the channel spacing to avoid cross-channel interference (i.e. around 3 ps). For terabit/s OTDM systems we would require pulsewidths less than 0.5 ps. The spectral width of the pulse source is also important, as it will have a major impact on how the pulse will evolve during propagation in the fiber. A standard figure of merit which is employed is the time–bandwidth product, and ideally we require the pulse source to be transform limited, which implies that the spectral width is as small as possible for the associated pulsewidth. The impact of temporal jitter on the pulses can be easily understood by considering the above example of a 100 Gbit/s system, employing 3 ps pulses, with the multiplexed channels spaced by 10 ps. Obviously if the jitter on the pulses becomes significant in relation to the channel spacing, then this can also lead to interference between adjacent channels in the overall OTDM system. A large number of techniques have been employed to develop ultrashort pulse sources suitable for use in OTDM systems, but the three main methods which will be described below are active mode-locking, gain-switching, and external modulation of a CW light signal.

Active mode-locking of laser diodes normally involves modulating the amplitude of the optical field inside the laser cavity at a frequency which is equal to the mode spacing of the laser. This can be achieved by applying an electrical sinusoidal signal at the correct frequency, and results in the generation of optical pulses at the repetition rate of the applied signal. This technique has been successfully employed in the generation of subpicosecond pulses at repetition rates up to and beyond 40 GHz, with excellent spectral and temporal jitter characteristics. However, an inherent problem in all mode-locked pulse sources is the difficulty in synchronizing the mode-locked frequency to a specific SDH standard data rate.

Gain switching of semiconductor laser diodes is probably the simplest and most reliable technique to generate optical pulses. The technique, which is presented in **Fig. 3**, involves applying a high-power electrical pulse (or electrical sinusoidal signal) to the laser in conjunction with a certain bias current. By ensuring that the electrical pulse signal and the bias signal have the correct level, the relaxation oscillation phenomenon of the laser results in the production of optical pulses with durations between 10 and 30 ps, at the repetition rate of the applied electrical signal. The frequency of the electrical signal applied to the laser is essentially arbitrary (provided it is not larger than the modulation bandwidth of the diode), thus making it straightforward to synchronize the optical pulse train to a SDH line-rate. The main problem with this technique is that the spectral width of the pulses generated is such that the pulses are far from transform-limited, which would affect their subsequent propagation in the fiber. In addition, temporal jitter on gain-switched pulses can also be a problem. However, by using novel arrangements such as external injection into the gain-switched laser, this difficulty can be overcome.

The third pulse generation technique mentioned above involves external modulation of a cw light signal with an electro-absorption modulator. The experimental configuration for this pulse generation technique is shown in **Fig. 4**. By biasing the modulator around its null point, and applying an electrical sinusoidal signal to it, the cw light passing through the modulator becomes shaped into optical pulses. The optical pulse train which is generated due the nonlinear response of the electro-absorption modulator is at a repetition rate of twice the applied electrical signal. The pulses generated are normally transform limited with extremely low temporal jitter, and the repetition rate is arbitrary as in the gain-switching technique (with the limit being ultimately determined by the modulation bandwidth of the modulator). This method may be readily used to generate pulses at repetition rates up to 40 GHz with pulsewidth around 5 ps, and it has the advantage that the optical source and the modulator can be integrated in a single device to form a compact pulse source suitable for OTDM communication systems.

It should also be noted that in addition to the various techniques that have been outlined, it is possible to use pulse compression in order to reduce the pulse width. The main issues concerned with pulse compression are the shape and spectral width of the optical pulses after compression. One of the most attractive methods of achieving pulse compression involves using nonlinear compression in dispersion-decreasing fiber, with the main advantage of this technique being the ability to maintain a transform limited pulse after compression. By employing this compression scheme, optical pulse sources at 10 GHz with pulsewidths below 200 fs have been developed, and such pulsewidths would be suitable for use in Tbit/s OTDM systems.

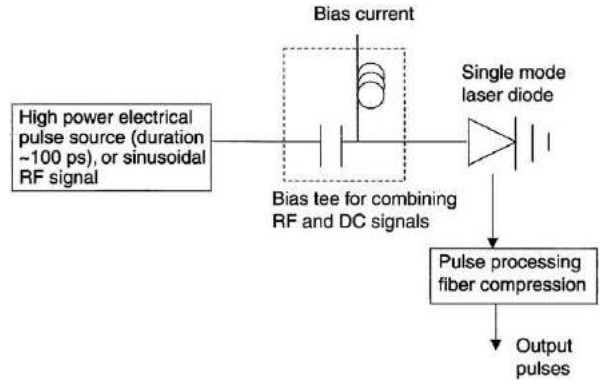


Fig. 3 Pulse generation using the gain-switching technique followed by pulse compression.

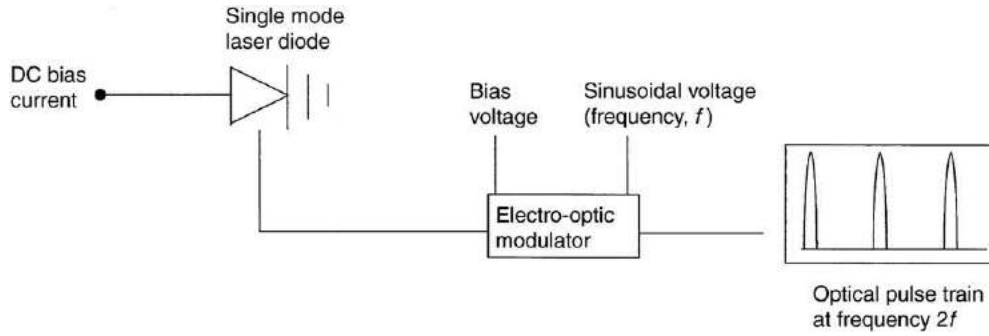


Fig. 4 Pulse generation based on shaping of cw light using the nonlinear response of an external modulator biased about its null point.

Transmission of an OTDM Signal over Fiber

The transmission performance of an OTDM data signal is vital in determining the distance over which the data can be transmitted successfully. The main fiber parameters, which will affect the signal transmission, are attenuation and dispersion. If we consider OTDM systems operating at a wavelength of around 1550 nm (minimum loss wavelength), to overcome the fiber attenuation and maintain a suitable optical power budget for the system, optical amplifiers are normally employed. If we thus assume that the amplifiers overcome the fiber loss problem, then the maximum transmission distance will be limited by the fiber dispersion. In a very basic OTDM system operating at 1550 nm, with transmitter and receiver linked using standard fiber (dispersion parameter of about 16 ps km^{-1} nm^{-1}), the maximum transmission distance will be limited by the overall data rate, and the pulsewidth and spectral width of the optical pulse source. For example, consider a 40 Gbit/s OTDM data signal which is formed using 8 ps optical pulses with a spectral width of 40 GHz (0.32 nm). From the spectral width we can calculate the signal broadening due to dispersion to be around 5 ps km^{-1}, and as the pulses broaden and spread into adjacent channels of the OTDM signal then this interference will make it increasingly difficult to correctly detect the signal at the receiver. In this case after propagation through 5 km of fiber, the signal pulses will have already dispersed to around 25 ps duration, the same value as the temporal bit slot into which each channel is placed. This will clearly result in serious loss of signal integrity. A straightforward possibility to increase the transmission distance of OTDM systems is to employ dispersion shifted fiber; the dispersion parameter is now around 1–2 ps km^{-1} nm^{-1}, at an operating wavelength of 1550 nm. This reduction in dispersion will obviously increase the allowed transmission distance by about an order of magnitude for the example described above. However, to develop ultrahigh-speed, long-haul OTDM communications we require more complex transmission schemes. Two possibilities for this include dispersion compensation, and soliton transmission techniques.

Dispersion compensation basically involves compensating for the dispersion that has been encountered during transmission by using some fiber with a total dispersion of opposite sign but equal magnitude to the transmission fiber. Dispersion compensation may also be achieved using a suitably designed fiber grating. For a long-haul OTDM system, a dispersion compensator may be used every 50 or 60 km, in conjunction with the optical amplifier, thus allowing us to compensate both fiber loss and attenuation periodically along the link. The main limitation, however, with the dispersion compensation technique is caused by the dispersion slope of the fiber, as for ideal compensation it is necessary to compensate completely for the dispersion slope in addition to the overall fiber dispersion. The second technique that may be employed to greatly extend the transmission distance of high-speed OTDM communication systems is the use of soliton transmission. The basic principle of soliton transmission is to use optical data pulses with a particular shape, pulsewidth, and peak power, such that as the pulse propagates, the effects of fiber dispersion and nonlinearity counterbalance to allow the signal to propagate undistorted. By using optical amplifiers to ensure that the optical pulse peak power does not vary too much along the transmission link, OTDM transmission at data rates greater than 100 Gbit/s, over distances approaching 500 km have been demonstrated.

Demultiplexing of OTDM Signal at Receiver

For OTDM communication systems with data rates above 40 Gbit/s, it is not feasible to use electrical switching. Indeed for ultrahigh-speed systems operating at 100 Gbit/s and beyond, the only solution for demultiplexing is to use all-optical switching in which optical control signals are used to switch the OTDM data signal. All-optical switches normally use nonlinear optical effects either in optical fiber or semiconductors. A typical scenario (as presented in Fig. 5) involves the injection of both the OTDM data signal and an optical control signal into the nonlinear device. The control signal consists of high-power ultrashort pulses at the repetition rate of the individual channels within the temporally multiplexed signal, and by synchronizing it with one of the OTDM channels it is possible to demultiplex this channel from the high-speed OTDM signal.

Two of the most popular all-optical demultiplexers are the nonlinear optical loop mirror (NOLM), and the terahertz optical asymmetric demultiplexer (TOAD). The NOLM is based on the nonlinear refractive index of optical fiber. It essentially consists of a 2×2 fiber coupler with its two outputs joined using a certain length of fiber. When an OTDM signal is injected into an input port of the coupler, it splits into two counterpropagating components in the fiber loop, and when these components recombine and

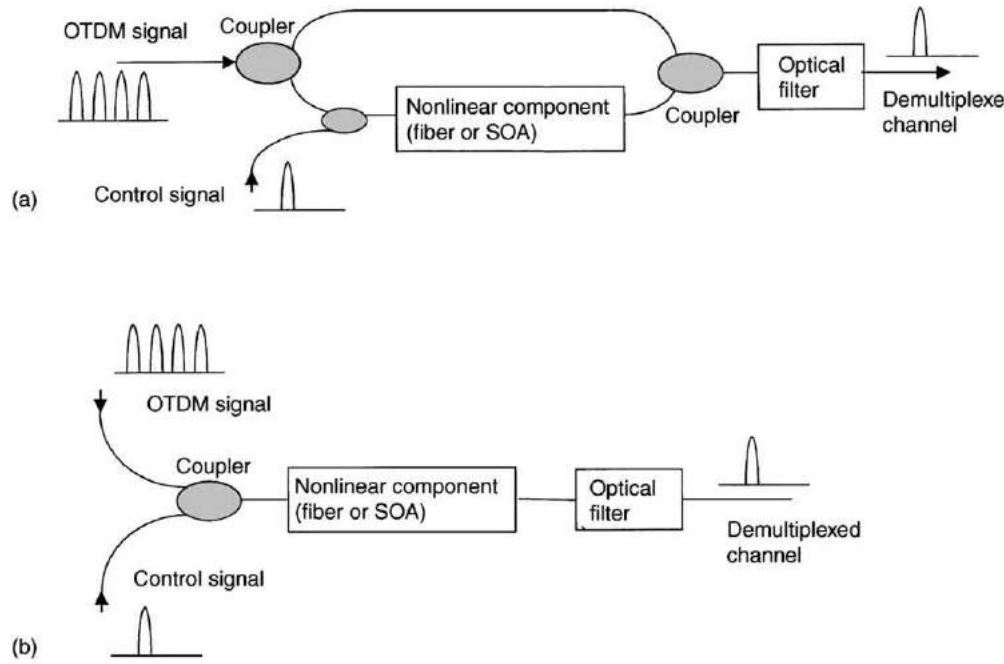


Fig. 5 Configuration of OTDM demultiplexers using (a) an interferometer based on nonlinear phase shift in fiber or semiconductor optical amplifiers, and (b) four-wave mixing in fiber or semiconductor optical amplifiers.

interfere at the coupler, the overall signal is output through its initial input port. If we now inject a high-power control signal directly into the loop, such that it propagates unidirectionally, and is synchronized with one of the OTDM data channels, then the phase shift induced by the control on the copropagating signal channel (via the nonlinear refractive index of the fiber) results in that particular channel being switched out to the second input port of the coupler. The remaining data channels of the OTDM signal are once again output through the same port that they entered the coupler. As this particular nonlinear effect in fiber (known as the Kerr effect) has a femtosecond response time, it should be possible to realize extremely high-speed switching with the NOLM. Demultiplexing at data rates in excess of 100 Gbit/s has already been achieved using this technique. One disadvantage of this method is that the nonlinear index coefficient of standard fiber is very small, so in order to achieve the required phase shift from the control pulse, it is necessary to use a fiber loop of around 1 km in length (depending on the power of the control pulse used). This length may affect the overall stability of the demultiplexer and its ability to be readily integrated into high-speed OTDM systems. However, it should be noted that recent developments in the design of high-nonlinearity fibers may reduce this problem.

The second all-optical switch mentioned above is known as a TOAD. Whereas the NOLM is based on the nonlinear refractive index of the fiber, the TOAD is based on a nonlinear optical effect in semiconductor optical amplifiers (SOA). The overall setup for a TOAD is very similar to that for a NOLM, except that a semiconductor amplifier, as shown in Fig. 5, replaces the length of fiber in the loop. As for the NOLM, the OTDM data signal is injected into the TOAD using one input of the coupler, such that it splits into two counterpropagating signals, while the control signal propagates unidirectionally in the loop containing the SOA. By synchronizing the control with one channel in the OTDM signal, this particular channel is switched out. Although the response time of a TOAD-based switch may not be as fast as the NOLM, it does have the major advantage that it can be developed into a very compact demultiplexer, suitable for deployment in OTDM-based communication systems.

In addition to the NOLM and the TOAD structures, which are both interferometric-based switches, it is also possible to use four-wavemixing (FWM) in optical fiber or semiconductor optical amplifiers to carry out the demultiplexing of OTDM signals. In this case, the high-power control pulse is synchronized with one of the OTDM channels such that the wavelength of this channel is shifted as it passes through the nonlinear device. Optical filtering may then be used to select out the demultiplexed channel.

As we discussed in the section on demultiplexing, whatever specific all-optical switching technique is employed, we need to generate an optical clock signal for use as the control pulses. This implies that it is necessary to extract a clock at the base rate of the individual data channels from the high-speed OTDM signal. Depending on the base data rate, which may be anywhere from 2.5 to 40 Gbit/s for typical OTDM systems, different clock recovery techniques have been demonstrated. These range from basic electronic clock recovery schemes to more advanced techniques based on self-pulsating laser diodes and optical phase-locked loops.

OTDM Networking Issues

The previous section has predominantly dealt with the basic elements required to implement a high-speed OTDM transmission link. However, if we are to use OTDM technology for high-speed networking, then additional elements will be required. One of

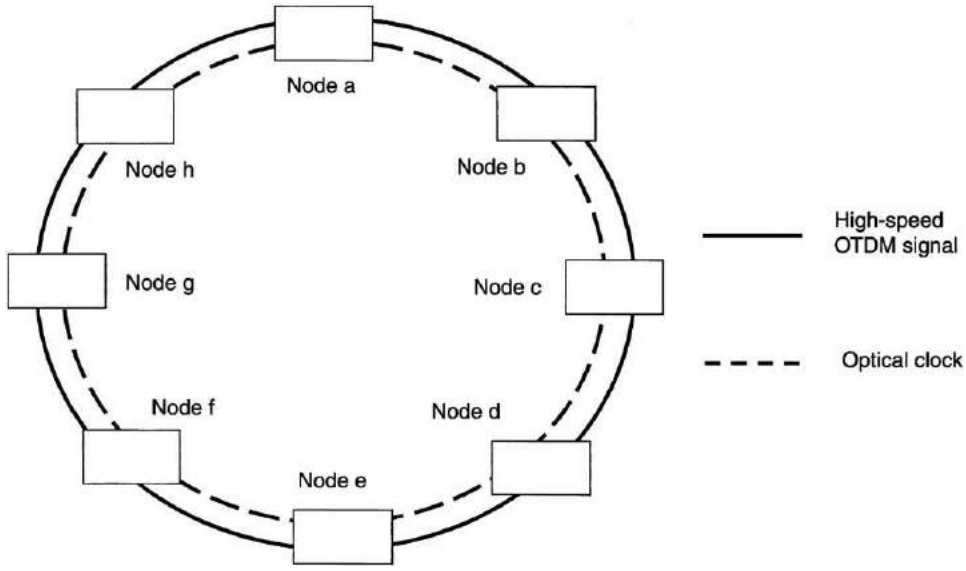


Fig. 6 Schematic representation of a ring network employing OTDM with a separate fiber for clock distribution.

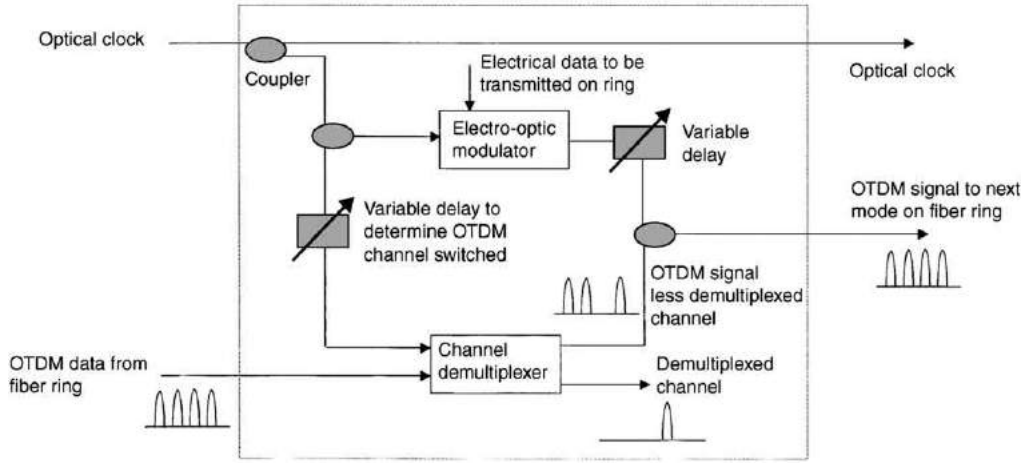


Fig. 7 Configuration of an add/drop multiplexing node for use in OTDM-based communication systems.

the most important of these is an add/drop multiplexer (ADM). The ADM allows us to switch out one channel from the OTDM signal, and then insert another data channel for transmission in the vacant bit slot. If we consider employing OTDM in a ring configuration network, then a typical architecture for the network will be as shown in Fig. 6. The high-speed OTDM data signal propagates around the closed fiber loop, and each node in the network is responsible for switching out the data it requires, in addition to inserting the data it wishes to transmit onto the fiber ring. Add/drop multiplexing is thus imperative in developing such an OTDM ring network.

The configuration for each node in an OTDM ring network is shown in Fig. 7. At each node it is initially necessary to recover the base rate clock signal. The recovered clock is then used in conjunction with the OTDM signal to switch out the required data channel from the multiplexed signal. When an add/drop multiplexer is used for this purpose we obtain two outputs, one being the demultiplexed data, and the other being the OTDM signal less the switched-out data channel. A simple fiber coupler can subsequently be used to allow this particular node to insert the data it wishes to send into the free bit slot of the overall OTDM signal. This obviously requires control of the relative delay between the OTDM signal and the local data channel to be inserted onto the fiber ring, to ensure that the local data are inserted into the vacant bit slot. A number of techniques have been used for implementing add/drop multiplexers in OTDM-based networks. These have used either fiber interferometers, or interferometers using semiconductor optical amplifiers, in which we obtain both the demultiplexed data signal and the OTDM signal less the switched-out channel (to which the local data channel is then added). The continuing development and enhancement of ADM devices for OTDM systems will be vital for the future implementation of OTDM communication systems.

Table 2 Main elements of an OTDM communications system

Device	*Principal function*	*Technical implementation*
Optical pulse source	Generate ultrashort optical pulses suitable for transmission in high-speed OTDM systems	Gain-switching, mode-locking, or external modulation techniques
Multiplexer	Multiplex individual pulse data channels in the temporal domain using appropriate delays to achieve an OTDM signal	Passive couplers with fixed fiber delay lines, or planar waveguide circuits (for accurate control of delays)
Optical amplifier	Amplify the OTDM signal during fiber transmission to overcome fiber loss	Fiber doped with rare earth materials (e.g. erbium) in conjunction with pump laser
Transmission fiber	Transmit the OTDM data signal from transmission site to required receiver	Standard single-mode fiber, dispersion shifted fiber, or other specialty fiber may be used
Clock recovery	Extract clock (at base rate of individual channels) from OTDM data signal at receiver, for use in demultiplexing	Electronic or optical phase-locked loops, self-pulsating lasers, or mode-locked ring laser techniques
Demultiplexer	Demultiplex one of the channels from the overall OTDM signal	Interferometric or four-wave mixing techniques based on fiber or semiconductor nonlinearity
Add/drop demultiplexer	Extract one channel from OTDM data signal, and insert a local channel into the empty time slot for transmission	Similar technical implementations to demultiplexer described above

Conclusion

In this article we have explained the main ideas involved in building OTDM communication systems, and also introduced the principal technologies which are required to do so. OTDM is an extremely wavelength-efficient technique for delivering high-capacity data signals, and it may be used both for long-haul transmission and networking around metropolitan areas. In addition, unlike WDM, it does not require a different wavelength source for each channel in the multiplexed signal. Despite these and other advantages, OTDM is still way behind WDM when it comes to commercial maturity, with no OTDM systems currently available in the telecommunications market. However, the technologies required to implement high-speed OTDM systems (outlined in **Table 2**), including ultrashort pulse sources, clock extraction circuitry, and demultiplexing devices, are developing rapidly. It therefore seems likely that in the future, OTDM technology may be used for enhancing the overall capacity of optical communication systems.

One possibility for employing OTDM systems would be for upgrading installed WDM networks. The current method for improving WDM networks involves using more wavelength channels, but it may also be feasible to increase the overall capacity by developing hybrid WDM/OTDM communications systems. In such a system, each wavelength channel may carry aggregate data rates in excess of 100 Gbit/s, achieved using OTDM. The design of these hybrid networks would permit switching of the optical data signals to be carried out at two different levels in the overall network. The WDM signals could be switched coarsely using passive filtering devices, and the OTDM data channels could be switched with fine granularity using one of the all-optical switching techniques available.

See also: Lightwave Transmitters

Further Reading

Agrawal, G.P., 1997. Fibre Optic Communications, 2nd edn. New York: Wiley.
Cole, M., 2000. Introduction to Telecommunications: Voice, Data, and the Internet. Englewood Cliffs, NJ: Prentice Hall, chapter 5.
Islam, M., 1992. Ultrafast Fiber Switching Devices and Systems. Cambridge, UK: Cambridge University Press.
Kawanishi, S., 1998. Ultrahigh-speed optical time-division-multiplexed transmission technology based on optical signal processing. IEEE Journal of Quantum Electronics 34, 2064–2079.
Midwinter, J.E., Guo, Y.L., 1992. Optoelectronics and Lightwave Technology. Chichester, UK: Wiley.
Mukherjee, B., 1997. Optical Communications Networks. New York: McGraw-Hill.
Nakazawa, M., Kubota, H., Suzuki, K., Yamada, E., Sahara, A., 2000. Ultrahigh-speed long distance TDM and WDM soliton transmission technologies. IEEE Journal of Selected Topics in Quantum Electronics 6, 363–396.
Prucnal, P.R., Santoro, M.A., Sehgal, S.K., 1986. Ultrafast all-optical synchronous multiple access fiber networks. IEEE Journal of Selected Areas of Communication 4, 1484–1493.
Spirit, D.M., Blank, L.C., Ellis, A.D., 1995. Optical time division multiplexing for future high capacity network applications. In: Smith, D.W. (Ed.), Optical Networking Technology. London: Chapman and Hall, pp. 54–77.
Spirit, D.M., Ellis, A.D., Barnsley, P.E., 1994. Optical time division multiplexing: systems and networks. IEEE Communications Magazine 32 (12), 56–62.
Tucker, R.S., Einstein, G., Korotky, S.K., 1988. Optical time-division multiplexing for very high bit rate transmission. IEEE Journal of Lightwave Technology 6, 1737–1749.

Lightwave Transmitters

JG McInerney, National University of Ireland-Cork, Cork, Ireland

Introduction

Optical fiber telecommunications has changed human society forever, providing the capacity for affordable and ubiquitous communication. It provides data rates and transport economies far in excess of those available using purely electronic means. It is doubtful whether the Internet and worldwide web, or the vast infrastructure of wireless cellular telephony, would be viable without optical fiber technology. Its rapid development and acceptance has been driven largely by contemporaneous successes in manufacturable low-loss optical fiber cables, sensitive pin-FET photoreceivers, and reliable high-performance laser diode-based transmitters. The earliest developments in silica glass fibers at Standard Telecommunication Laboratories, UK, (then part of the US-based ITT Corporation) and Corning Glass Works (USA) in the 1960s and 1970s, were pivotal in determining the directions of modern fiber technology, as were later key inventions of rare-earth doped fiber amplifiers at the University of Southampton (UK) and AT&T Bell Laboratories (USA) in the 1980s and early 1990s. However, perhaps the greatest contributors to the success of fiber optics have been semiconductor injection lasers, or laser diodes, which in various forms have made up the vast majority of all fiber optic transmitters, and the entirety of long-haul transmitters, as well as pump sources for the doped-fiber amplifiers which enable modern networks. The fascinating story of the development of fiber optics is to a large extent driven by the dramatic progress in the development of high-performance, reliable semiconductor lasers. These emitted at wavelengths near 1300 and 1550 nm, which respectively correspond to the dispersion and attenuation minima of single-mode silica optical fibers.

Semiconductor Laser Principles

Semiconductor lasers were first demonstrated in research laboratories at General Electric, IBM Corporation and the MIT Lincoln Laboratory (all USA) as early as 1962, although it took almost two decades for the basic science and engineering of materials, fabrication, reliability, and performance design issues to be developed sufficiently for their use in practical communication systems. Early successes were obtained using the GaAs/AlGaAs lattice-matched material system, in which a lightly doped GaAs or ternary compound AlGaAs active layer is sandwiched between two lattice-matched heterojunctions with n- and p-doped AlGaAs layers with higher Al fractions than the active layer, so that their refractive indices are lower and their bandgaps higher, providing optical and charge-carrier confinement. In all cases the material is epitaxially grown, meaning that the entire laser chip forms a single crystal. The gain, and hence the laser emission, occurs at the direct bandgap of the active layer, at wavelengths in the range 750–870 nm depending on the Al fraction. Later evolutions used the quaternary compound GaInAsP clad by InP guiding layers, where the In and P fractions are chosen to match the lattice constant to the InP substrate.

In general, optical gain in a laser diode is generated by creating an electron-hole plasma in the vicinity of the forward-biased junction, a situation corresponding to population inversion in a conventional laser, and with sufficient forward current the material becomes transparent, in that the gain exactly equals the absorption and scattering losses, at a particular wavelength. When this arrangement is enclosed within a suitable optical cavity, and the forward current is increased further, the system begins to lase, that is to oscillate continuously to produce coherent optical radiation at a wavelength which satisfies two basic conditions: it must be within the set of wavelengths at which the material produces sufficient gain, and it must be close to one of the electromagnetic modes at which the optical cavity is resonant. These conditions, which define lasing threshold, are characterized by the equality of the optical gain and the total losses, that is the material losses (absorption, scattering, and free-carrier plasma) and the cavity losses which include the light output.

Early laser diodes in both the GaAs/AlGaAs and GaInAsP/InP systems were of the edge-emitting type, in which the light is emitted from the edge of the laser chip, perpendicular to the growth direction in the plane of the wafer. Although this geometry complicates laser production flow, particularly the testing function which requires that the wafer be scribed into bars to define the laser output facets, edge-emitting lasers (EELs) make up the vast majority of optical fiber transmission sources. An alternative laser diode structure is the vertical cavity surface-emitting laser (VCSEL), in which emission occurs in the growth direction perpendicular to the plane of the wafer. VCSELs offer major advantages in spectral stability, beam quality, manufacturability and cost, but their output powers are low (a few mW) and they are not yet available at the key telecommunication windows around 1300 nm and 1550 nm, due to difficulty in forming the necessary high-reflectivity Bragg mirrors in the GaInAsP/InP material system. Major initiatives are currently underway to produce long-wave VCSELs in the GaAs/AlGaAs system using InAs quantum dots, or dilute nitrides such as InGaAsN, or hybrid approaches such as InP-based gain media fused to high-reflectivity mirrors using GaAs/AlGaAs, dielectric coatings, air gaps, and others. At present, however, all long-haul optical fiber transmission is based on edge-emitting laser diodes fabricated in GaInAsP/InP.

General Structure and Requirements of Optical Fiber Communication Systems

The first optical fiber communication links were simple point-to-point affairs, consisting essentially of series-connected transmitters, fibers, and receivers. The transmitters were simple Fabry–Perot laser diodes, onto which data were encoded by direct digital modulation of the injection current at rates ~100 Mbit/s. The first commercial optical fiber link was built in 1976, a single fiber cable linking two switching centers of the Illinois Bell Telephone Co just 2.5 km apart in the Chicago metropolitan area, using 850 nm GaAs laser technology. In 1988, the first transatlantic fiber link, AT&T's TAT-8, was completed linking endpoints in New Jersey, England, and France with three fiber pairs carrying signals generated by 1300 nm GaInAsp single mode lasers. In these early systems, data were regenerated frequently by repeaters each consisting of a receiver, signal conditioner and transmitter. Each fiber in TAT-8 carried a single optical frequency modulated at 280 Mbit/s and the construction cost of the system was ~$50 k/km, resulting in an economic figure of merit of over $600 k/Mbit/s. Twenty-five years later, current systems cost the same per kilometer to build in absolute dollars despite inflation, a major effective cost reduction enabled largely by replacement of the expensive repeaters by doped-fiber optical amplifiers. Moreover, each fiber is now highly multiplexed, with the potential for hundreds of wavelengths each carrying 10 Gbit/s signals generated by 1550 nm laser transmitters, so that the unit cost has plummeted to ~$200/Mbit/s, a reduction of three and a half orders of magnitude.

In the remainder of this article, we will describe transmitters for both short- and long-haul systems where the channel spacing is only tens of GHz in the optical carrier frequency. Modern short-haul systems either feed client networks as part of bi-directional transceivers, or drive metro and local access networks connected directly by photonic switches to terminals of the long-haul systems.

DWDM Transceivers

Modern practice uses modular units containing transmitters and receivers allowing bi-directional data flow. In the send mode, data are applied to a long-haul transmitter at the local (client) source and coupled through the external fiber network to the remote long-haul receiver. In the receive mode, data arrive from a remote long-haul transmitter and are coupled (via a regenerator including signal conditioning, clock recovery, and error correcting steps) to a local short-haul transmitter leading to the client's internal network. Each transceiver card therefore contains a long-haul transmitter and receiver set (known collectively as dense wavelength division multiplexing (DWDM) transport devices) and a short-haul transmitter and receiver set (known as client interfaces or local transport devices).

Transmitter Requirements

Short-haul transmitters may be simple Fabry–Perot edge-emitting laser diodes, or vertical cavity laser diodes, or even superluminescent LEDs for sub-Gbit/s data rates. They are generally not required to be at one of the minimum-loss telecom windows near 1550 nm, as the transmission distances are short. However, as client networks grow into large metropolitan networks, it is likely that 1300 nm sources will be used to minimize dispersion. Power requirements are modest, a few mW peak, at data rates ~1–10 Gbit/s. Single mode laser emission and polarization control are not required by present day systems, although polarization-mode dispersion may limit transmission at higher data rates. Both return-to-zero (RZ) and nonreturn-to-zero (NRZ) coding may be used, and direct modulation of the laser amplitude is the norm.

For long-haul transmitters the situation is dramatically different. Operation in one of the standard erbium-doped fiber amplifier (EDFA) bands near 1550 nm is mandatory: the C (conventional) band extends from 1530–1565 nm, the S (short-wave) band from 1460–1530 nm) and the L (long-wave) band from 1565–1625 nm. In the C-band, output powers ~10 mW are generally sufficient, and powers above ~30 mW are limited by nonlinear effects such as self-phase modulation and four-wave mixing in the fiber. In the S- and L-bands, where the available gain from EDFAs is less, lasers may need to operate close to the nonlinear limit, several tens of mW, although various schemes (differential pumping, gain-flattening filters, etc.) are being evaluated to flatten the global EDFA gain spectrum.

In current practice, a single fiber utilizing DWDM, using only the C-band, can carry up to 1 Tbit/s comprising 100 channels at 10 Gbit/s on each channel, the current standard. For such performance a channel spacing of 50 GHz in optical frequency is required. Using all three bands with similar density would enable ~3 Tbit/s, which could be doubled by using 25 GHz channel spacings. Ultimately, even higher channel densities and hence overall system data rates are possible, with a practical limit ~10 Tbit/s. The channel spacing is a key parameter in that it determines the laser design, specifically its spectral stability and linewidth, and the overall transmitter design in that laser chirp (modulation induced dynamic spectral shift) must be less than half the channel spacing. In practice, only single-wavelength and coarse wavelength division multiplexing (CWDM) systems may be modulated directly. All modern DWDM systems require external modulation, for example using electro-absorption or Mach–Zehnder interferometric devices outside the laser cavity. For the latter cases, stabilization of the laser wavelength is required for example by on-chip gratings to form distributed feedback (DFB) or distributed Bragg reflector (DBR) lasers, or by external fiber gratings.

Directly Modulated Lasers

Direct modulation of semiconductor lasers is convenient and effective: by modulating the laser injection current, one modulates the carrier density and hence the optical gain, resulting in modulation of the laser output up to ~ 10 Gbit/s. The actual modulation bandwidth is determined by interactions between photons and carriers, whose respective decay lifetimes, τ_s and τ_p, are determined by cavity losses and total recombination rates. The simplest form of such interactions is described by the photon and carrier rate equations for the injected carrier density N and photon density S in a single lasing mode:

$$dN/dt = J/ed - gNS - N/\tau_s \tag{1}$$

$$dP/dt = gNS + \beta N/\tau_s - S/\tau_p \tag{2}$$

where J is the injected current density, e the electronic charge, d is the thickness of the active region (hence J/ed is the rate of injection of carriers per unit volume), g is the gain coefficient per unit length and β is the fraction of the spontaneous emission coupled into the lasing mode, typically $\sim 10^{-2}$–10^{-5} for edge-emitting lasers (product of geometric and spectral overlap factors). These equations simply state that the rate of change of the carriers or photons is given by the rate of generation less the loss rate. Additional terms are required in the presence of optical feedback or coupling (in which case the photon density is replaced by the complex amplitude and phase of the optical electric field) or for extremely rapid modulation where the traveling wave nature of the disturbances in the photon fields is significant. These so-called traveling-wave rate equations for the photons are of the form:

$$dS_+/dt + cdS_+/dz = gNS_+ + \beta N/\tau_s \tag{3}$$

$$dS_-/dt - cdS_-/dz = gNS_- + \beta N/\tau_s \tag{4}$$

with the same carrier rate equation as [1] using the total photon number S made up of the forward-traveling component S_+ and the backward-traveling component S_- : $S = S_+ + S_-$.

Both sets of rate equations are based on approximations such as homogeneous gain broadening, neglect of transverse field effects, and diffusion. These are valid in most situations but care is required when analyzing lasers with large transverse apertures, or when ultrashort (picosecond) pulses are involved.

Small-Signal Modulation

Small-signal modulation may be analyzed by a linear perturbation analysis of the rate equations, leading to a conjugate pair of poles in the response function, with damped relaxation oscillations at frequency:

$$f_R = (1/2\pi)\sqrt{(gS_0/\tau_p)} \tag{5}$$

or, in terms of the injected current density J (assuming $S \sim G(J - J_{th})/ed$ with Γ the electromagnetic confinement factor of the mode):

$$f_R = (1/2\pi)\sqrt{[\Gamma g(J - J_{th})/ed]}$$

f_R can be regarded as the resonant frequency for the interaction between the carriers and photons. As a practical matter, lasers can be modulated up to $\sim 2f_R$ but the modulation response rolls off rapidly with increasing frequency above f_R. Using explicit expressions for the threshold gain, we can write f_R as:

$$f_R = (1/2\pi)\sqrt{\left[\left(GN_T g\tau_p + 1\right)\left(J/J_{th} - 1\right)/\tau_s\tau_p\right]} \tag{6}$$

where N_T is the transparency carrier density $\sim 10^{18}$ cm^{-3}. f_R can take values in excess of 10 GHz for a well-designed laser. It should be noted, however, that high modulation bandwidth almost always requires high differential gain dg/dN and short τ_p. Increasing dg/dN requires p-doping (which increases optical loss) and/or some special quantum-confining structure such as wells, wires, or dots. Decreasing τ_p inevitably leads to greater optical loss and thus a higher threshold.

The damping rate of the relaxation oscillations is also important in limiting modulation bandwidth, and also (when increased) in reducing the sensitivity of the laser to optical feedback. For bulk or quantum well lasers, the condition for critical damping is $A^2 = 4B$ with $A = (S/g\tau_s - \sigma[N - N_T]/g\tau_p$ and $B = \sigma[1 + (S_0/g\tau_s) - (1 - \beta)(N - N_T)]$. For quantum dot lasers there are additional damping terms due to carrier transport and thermalization which may restrict small-signal modulation bandwidths to a few GHz.

Experimentally, small-signal modulation properties may be determined using a simple sampling oscilloscope, low-noise tunable signal generator and spectrum analyzer. A combination of these elements with automated frequency sweeping may be found in a scalar network analyzer acting as an *s*-parameter test set. The modulated laser is connected to port 1 and a high speed photodetector connected to port 2, then a swept-frequency measurement of the transfer characteristic s_{21} is performed. In addition to the limitations on modulation response due to the carrier-photon resonance, practical lasers are limited by RC parasitics in the laser chip (junction impedance) and its package. For the highest modulation speeds, packages need to be designed as microwave transmission line components with effective impedance matching and low back reflections, as characterized by the voltage standing wave ratio (VSWR).

Large-Signal Modulation

Large-signal direct modulation can be analyzed by numerical integration of the rate equations. In general, its results are beyond the scope of this review, but for effective high-speed response, digital modulation of laser diodes should be carried out with a pre-bias close to threshold. Modulation bandwidth also increases with laser power, which is limited for systems considerations by fiber nonlinearities, and for reasons of laser reliability. In pulse code modulation, care must be taken that the laser is near critical damping, to minimize thermal patterning effects due to long strings of ones (high power) or zeros (low power). Experimentally, large signal modulation is analyzed by constructing eye diagrams, in which pulse traces on a fast oscilloscope are continuously overlaid when the laser is modulated using pseudorandom binary pulses. When the signal quality is high, the high and low digital levels are easily distinguishable, resulting in an open 'eye' in the accumulated traces.

Requirements for Externally Modulated Lasers

The requirement for external modulation occurs when the laser chirp exceeds half the desired channel spacing in the communication system. Chirp occurs due to changes in the refractive index, and hence the optical phase, during modulation. This dynamic phase shift then results in an instantaneous frequency shift. For direct modulation we have instantaneous wavelength

$$\lambda(t) = \lambda(0)\mu(t)/\mu(0) \tag{7}$$

Where $\mu(t)$ is the refractive index at time t. μ shifts with carrier density N due to combinations of several factors (free carrier plasma dispersion, thermal bandgap shrinkage, bandgap renormalization, dynamic Burstein–Moss effects) which give $\mathrm{d}\mu/\mathrm{d}N \sim -3\times10^{-20}\mathrm{cm}^3$, which for large signal digital modulation can result in tens of GHz. The actual degree of chirp depends on the so-called antiguiding or linewidth enhancement factor a given by

$$\alpha = -2k_0(\mathrm{d}\mu/\mathrm{d}N)/(\mathrm{d}g/\mathrm{d}N) \tag{8}$$

where $k_0 = 2\pi/\lambda_0$ is the free-space wavenumber. α as defined is positive, typically in the range 2–7 depending on the material, operating carrier density, and wavelength relative to the gain peak. Distributed feedback (DFB) or distributed Bragg reflector (DBR) lasers, which use integrated gratings to control the lasing wavelength, can have lower chirp by detuning from the gain peak but therefore are likely to require careful temperature stabilization. Quantum dot lasers, based on 3D nanoclustered active regions, have the potential for very low chirp, narrow linewidths, and reduced wavelength shifts, so that directly modulated DWDM transmitters may be possible.

Integration of laser diodes with electro-absorption or Mach–Zehnder type modulators is achieved in the transmitter optical subassembly and may in some cases be accomplished by monolithic integration on the same chip. Using separate modulators allows each device to be optimized independently but requires optical assembly, alignment, and retention. Mach–Zehnder modulators in X-cut lithium niobate, for example, have essentially zero chirp and can operate to 40 Gbit/s and above, with dynamic extinction ratios (ratio between fully on and fully off) of ~20 dB. Integrated semiconductor M-Z modulators have chirp an order of magnitude less than those of directly modulated lasers, so that channel spacings ~25 GHz are possible in DWDM systems with suitable temperature and wavelength controls. Although the details of such modulators are beyond the scope of this article, the laser requirements are to generate ~10 mW of continuous power with stable center wavelength and narrow linewidth. Such lasers are usually mounted in hermetic packages (e.g., the current 14-pin butterfly standard) with integrated thermo-electric cooler, power monitor photodiode, an optical isolator to suppress optical backreflections which cause instabilities and self-pulsing, and optional wavelength locking optics.

Reliability of Lasers in Fiber Optic Systems

From the early days, when laser diode lifetimes were measured in seconds even at cryogenic temperature, enormous progress has been made in achieving materials purity and reducing crystalline defects. Today's fiber optic laser transmitters have projected lifetimes of decades. In common with electronic devices, diode lasers fail at a rate given by a 'bathtub' curve, that is the failure rate $r(t)$ defined as the probability of failure per unit time at time t, has relatively high values at low t (early failures) and high t (wearout failures) and very low values in between. In terms of the population of lasers $n(t)$ we have:

$$r(t) = (-1/n(t))\mathrm{d}n(t)/\mathrm{d}t \tag{9}$$

If Δn is the number of samples which fail in time Δt, then assuming Δt begins at $t=0$, the effective or average failure rate r_{eff} over the interval is

$$r_{\mathrm{eff}}(t) = \Delta t\Delta n/n(0) \tag{10}$$

In reliability science it is customary to define r_{eff} in FITs (failures integrated in time) with the time-span chosen to yield statistically significant numbers. For electronic and photonic devices, it is customary to select the time interval $\Delta t = 10^9$h (1 billion operating hours), so that if 1% of the devices fail in 10 years ($\sim10^5$ hours) we have $r_{\mathrm{eff}} \sim 100$ FITs, which is the order of magnitude required by modern fiber optic laser sources.

In terms of actual statistical models, failure rates can be estimated using failure probability density functions $f(t)$. This is related to $r(t)$ by $f(t)=r(t)S(t)$ where $S(t)$ is the cumulative probability of surviving until t. When $r(t)$ is nearly constant, as in early failures due to material defects or process errors, the statistics are approximately exponential:

$$f(t) + (1/\tau)\exp(-t/\tau) \tag{11}$$

or Weibull-distributed (an exponential is a Weibull distribution with $\beta=1$):

$$f(t) = \left(\beta/t^{\beta}\right)t^{b-1}\exp\left[-\left(t/\tau\right)^{b}\right] \tag{12}$$

but this description is not suited to wearout failures for which $f(t)$ shows a significant increase as the population ages. For such cases, the lognormal distribution is often applicable:

$$f(t) = (1/\sigma t\sqrt{(2\pi)})\exp[-(\ln t - \ln\tau)^2/2\sigma^2] \tag{13}$$

where in each case τ is approximately the mean time to failure (MTTF).

Laser diodes undergo standardized qualification tests to demonstrate sufficient reliability prior to being used in commercial systems. Operationally, failure statistics must be tabulated and the best fit obtained. Burn-in procedures are used to screen early failures. Chip and module failures should be distinguished clearly, as the latter includes many additional factors such as thermal and power management, light coupling, mechanical or chemical integrity.

Given that effective lifetimes of decades are required, it is clearly necessary to accelerate aging to produce statistically meaningful failure rates in reasonable times $\sim 10^3$ h. For the most common modes of laser chip degradation, due to recombination-induced aggregation of defects in the crystalline epitaxial material, these are thermally activated and current driven, so that it is customary to use a modified Arrhenius law at temperature T:

$$\tau \sim (1/J^n)\exp(E_a/k_B T) \tag{14}$$

where J is the operating current density, n is an index ~ 2, E_a is the activation energy, and k_B is Boltzmann's constant. In terms of thermal acceleration, we have a factor $F=\tau(T_A)/\tau(T_B)$ with:

$$F = \exp[(E_a/k_B)(1/T_A - 1/T_B)] \tag{15}$$

Typical values of E_a for laser diodes are ~ 1 eV so that acceleration factors $\sim 10^3$ are possible.

Finally, it should be noted that laser diodes are generally subject to catastrophic failure in the event of overdriving or static discharge, the failure mode being thermal facet damage at ~ 1–10 MW/cm^2 for continuous-wave operation, the exact value depending on the material, surface preparation and specifically surface state absorption and its temperature coefficient.

Conclusions

Laser diodes are ideal sources for optical fiber communication systems and have propelled the development of fiber optics from its origins in the 1960s to the present day. They are compact, rugged, efficient, and reliable sources of light at key wavelength ranges such as 1300–1310 nm (short haul high speed links) and 1500–1600 nm (long haul amplified systems). They are capable of direct modulation at gigabit rates for simple systems but require external modulation and careful wavelength control for dense wavelength division multiplexed terabit systems. Transmitter lasers for real systems must satisfy stringent reliability conditions.

See also: Basic Concepts of Optical Amplifiers

Further Reading

Adams, M.J., Thomas, B., 1977. Detailed calculations of transient effects in semiconductor injection lasers. IEEE Journal of Quantum Electronics QE-13, 580.
Agrawal, G.P., 2002. Fiber Optic Communication Systems, 3rd edn. New York: Wiley.
Alferov, Z.I., Andreev, V.M., Portnoi, E.L., Trukan, M.K., 1970. AlAs-GaAs heterojunction injection lasers with a low room-temperature threshold. Soviet Physics – Semiconductors 3, 1107.
Basov, N.G., Krokhin, O.N., Popov, Y.M., 1961. Production of negative-temperature states in p-n junctions of degenerate semiconductors. Soviet Physics – JETP 13, 1320.
Bernard, M.G.A., Duraffourg, G., 1961. Laser conditions in semiconductors. Physica Status Solidi 1, 699.
Bhattacharya, P., 1966. Semiconductor Optoelectronic Devices, 2nd edn. Englewood Cliffs, NJ: Prentice-Hall.
Cheng, J., Dutta, N.K. (Eds.), 2000. Vertical Cavity Surface-Emitting Semiconductor Lasers: Technology and Applications. London: Taylor & Francis.
Chuang, S.-L., 1995. Physics of Optoelectronic Devices. New York: Wiley.
Coldren, L.A., Corzine, S.W., 1995. Laser Diodes and Photonic Integrated Circuits. New York: Wiley.
Dupuis, R.D., 1987. An introduction to the development of the semiconductor laser. IEEE Journal of Quantum Electronics QE-23, 651.
Fukuda, M., 1991. Reliability and Degradation of Semiconductor Lasers and LEDs. Norwood, MA: Artech House.
Fukuda, M., 1998. Optical Semiconductor Devices. Wiley.
Guekos, G. (Ed.), 1998. Photonic Devices for Telecommunications. New York: Springer.
Hall, R.N., Fenner, G.E., Kingsley, J.D., Soltys, T.J., Carlson, R.O., 1962. Coherent light emission from GaAs junctions. Physical Review Letters 9, 366.
Hayashi, I., Panish, M.B., Foy, P.W., Sumski, S., 1970. Junction lasers which operate continuously at room temperature. Applied Physics Letters 17, 109.
Holonyak Jr, N., Bevacqua, S.F., 1962. Coherent (visible) light emission from Ga(As1-xPx) junctions. Applied Physics Letters 1, 82.

Ito, R., Nakashima, H., Kishino, S., Nakada, O., 1975. Degradation sources in GaAs-AlGaAs double-heterostructure lasers. IEEE Journal of Quantum Electronics QE-11, 551.
Kaminow, I.P., Koch, T.L. (Eds.), 1997. Optical Fiber Telecommunications III (two parts). London: Academic Press.
Kaminow, I.P., Li, T. (Eds.), 2002. Optical Fiber Telecommunications IV (two parts). Amsterdam: Elsevier.
Kapon, E., 1998. Semiconductor Lasers, (two parts). London: Academic.
Kartalopoulos, S.V., 1999. Introduction to DWDM Technology. New York: Wiley.
Kartalopoulos, S.V., 1999. Understanding SONET/SDH and ATM. IEEE Press.
Kasap, S.O., 2001. Optoelectronics and Photonics: Principles and Practices. Englewood Cliffs, NJ: Prentice-Hall.
Kazovsky, L., Benedetto, S., Willner, A., 1996. Optical Fiber Communication Systems. Norwood, MA: Artech House.
Keiser, G., 1991. Optical Fiber Communications. New York: McGraw-Hill.
Kressel, H., Lockwood, H.F., 1974. A review of gradual degradation phenomena in electroluminescent diodes. Journal de Physique 35, C3–223.
Lasher, G., Stern, F., 1964. Spontaneous and stimulated recombination radiation in semiconductors. Physical Review 133, A553.
Milonni, P.W., Eberly, J.H., 1988. Lasers. New York: Wiley.
Morthier, G., Vankwikelberge, P., 1997. Handbook of Distributed Feedback Laser Diodes. Norwood, MA: Artech House.
Nathan, M.I., Dumke, W.P., Burns, G., Dill Jr, F.H., Lasher, G., 1962. Stimulated emission of radiation from GaAs p-n junctions. Applied Physics Letters 1, 62.
Numai, T., 2004. Fundamentals of Semiconductor Lasers. New York: Springer.
Paoli, T.L., 1977. Changes in the optical properties of cw (AlGa) as junction lasers during accelerated aging. IEEE Journal of Quantum Electronics QE-13, 351.
Paoli, T.L., Ripper, T.E., 1970. Direct modulation of semiconductor lasers. Proceedings of IEEE 58, 1457.
Pearsall, T.P., 2003. Photonics Essentials: An Introduction with Experiments. New York: McGraw-Hill.
Piprek, J., 2003. Semiconductor Optoelectronic Devices: Introduction to Physics and Simulation. London: Academic.
Quist, T.M., Rediker, R.H., Keyes, R.J., Krag, W.E., Lax, B., McWhorter, A.L., Zeiger, H.J., 1962. Semiconductor maser of GaAs. Applied Physics Letters 1, 91.
Reinhart, F.K., Hayashi, I., Panish, M.B., 1971. Mode reflectivity and waveguide properties of double-heterostructure injection. Lasers 42, 4466.
Saleh, B.E.A., Teich, M.C., 1991. Fundamentals of Photonics. New York: Wiley.
Senior, J.M., 1992. Optical Fiber Communications, 2nd edn. Engelwood cliffs, NJ: Prentice-Hall.
Siegman, A.E., 1986. Lasers. University Science Books.
Telcordia Technologies (formerly Bellcore): document family FR-796 on Reliability and Quality Generic Requirements for telecommunications equipment. Latest issue is 003 dated Jul 2003. The specific document related to laser reliability is GR-468 – Generic Reliability Assurance Requirements for Optoelectronic Devices used in Telecommunications Equipment. Standards for other network photonic, electronic and mechanical components may be found in the document family FR-2063 Network Equipment-Building System Family of Requirements, issue 003 dated Feb 2003 and specifically FR-357 – Network Equipment-Building System Family of Requirements: Component Design for Reliability.
Verdeyen, J.T., 1994. Laser Electronics, 3rd edn. Englewood Cliffs, NJ: Prentice-Hall.
Yariv, A., 1989. Quantum Electronics, 3rd edn. New York: Wiley.
Zachos, T.H., Ripper, J.E., 1969. Resonant modes of GaAs junction lasers. IEEE Journal of Quantum Electronics QE-5, 29.

Broadband Passive Optical Access Networks

Elaine Wong and Maluge P Imali Dias, The University of Melbourne, Parkville, VIC, Australia
Zhengxuan Li and Lilin Yi, Shanghai Jiao Tong University, Shanghai, P.R. China

Introduction

The exponential growth in Internet and mobile traffic along with bandwidth-intensive applications, has continued to drive the deployment of fiber networks deeper into the access network segment (Wong, 2012). In future-proofing against the forecasted increase in bandwidth demands, Fiber-To-The-X (X=home, curb, building) networks have been deployed in various parts of the world. Offering direct fiber connection to or close to the home, FTTx deployments can assume point-to-multipoint (PtMP) or point-to-point (PtP) physical topologies, as illustrated in **Fig. 1** (Wong, 2012). Driven by low cost and low energy consumption per bit, most deployed FTTx models are based on the passive optical network (PON) topology (Wong, 2012). Typically, a PON has a PtMP topology and comprises a central office (CO) which houses multiple optical line terminals (OLTs), optical network units (ONUs) residing in subscriber premises, and a passive splitting optical distribution network (ODN) consisting of a feeder fiber, passive optical splitter(s), and multiple distribution fibers.

Commercial PONs such as Broadband PON (BPON) (ITU-T, 2005), Gigabit PON (GPON) (ITU-T, 2008a) and Gigabit Ethernet PON (GE-PON) (IEEE, 2004) are deployed around the world. GPON systems are specified by the ITU-T G.984 Gigabit PON (GPON) standard (ITU-T, 2008a) whereas GE-PON is specified by the IEEE 802.3ah Gigabit Ethernet PON (GE-PON) standard (IEEE, 2004). Both GPON and GE-PON are classified as time division multiplexed/time division multiple access (TDM/TDMA) PONs. In TDM/TDMA PONs, encrypted information is broadcast to all ONUs in timeslots at a line-rate of 2.5 Gb/s for GPON and 1.25 Gb/s for GE-PON in the downstream direction. Due to the power-splitting ODN, all time slots are received by each ONU but only particular time slots addressed to the ONU through its media access control (MAC) destination address, are detected and received. In the upstream direction, burst mode TDMA is used between ONUs to share the aggregate 1.25 Gb/s upstream bandwidth (for both GPON and GE-PON). If required, a dedicated radio frequency (RF) channel can be added to broadcast TV/video data (Wong, 2012).

One major disadvantage of the power-splitting nature PONs is the limitation in further scaling of bandwidth, reach, and subscriber count. Specifically, the splitting loss of the ODN is a strong function of the number of supported subscribers thereby constraining any possible increase in user count, reach, and/or average user data rate. To overcome this limitation and to support future business, residential, back- and front-hauling bandwidth needs, next-generation PONs commonly referred to as NG-PON1, with a maximum line rate of 10 Gb/s, were recently standardized. Defined by both IEEE (2009) and ITU-T (2010), these standards allowed for backward compatibility and co-existence with the current-generation GPONs and GE-PONs, thus enabling progressive upgrades with minimal capital investment on the external plant and operational impact on existing users.

The PtP network can overcome the limitation in bandwidth, reach, and subscriber count of power-splitting PONs. In a PtP network, dedicated fiber and hence bandwidth is assigned between the CO and each ONU. Other beneficial features include enhanced data privacy and security. Nonetheless, PtP suffers from high fiber count and terminations at the CO (one per ONU) and thus requires high density fiber management. While the focus of this article is on PONs that adopt the PtMP topology, PtP Ethernet

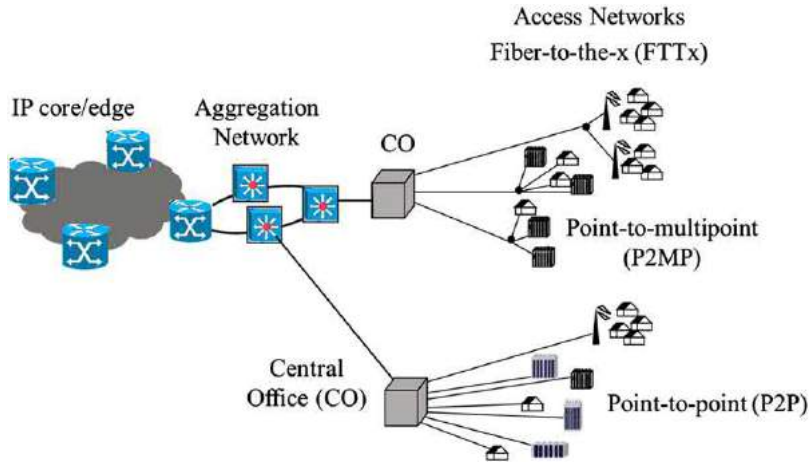


Fig. 1 Schematic of network structure showing the physical topology options of access networks. Reproduced from Wong, E., 2012. Next-generation broadband access networks and technologies. IEEE/OSA J. Lightw. Technol. 30(4), 597–608.

is popular in some parts of the world, e.g., Japan, China and Sweden, and has been deployed to connect high revenue fixed access and mobile business customers that require guaranteed bandwidths (Förster, 2006).

New access technologies beyond NG-PON 1, namely 10 Gb/s TDM/TDMA systems, are being explored to support symmetrical average data rates of ~1 Gb/s per user, an extended system reach of 60–100 km, a high subscriber count of up to 1000, and heterogeneous service convergence, while meeting the cost constraints of the access market. The standardization of time and wavelength division multiplexed PON (TWDM-PON) as the selected NG-PON2 technology, specifies a 40-Gb/s downstream capacity using an underlying hybrid time and wavelength division multiplexing scheme (ITU, 2013). Each wavelength in a TWDM-PON is shared between multiple ONUs by employing time division multiplexing and multiple access mechanisms.

Beyond the promises of NG-PON2, wavelength division multiplexed (WDM) PONs are being explored as the next-generation solution. The idea of exploiting wavelength division multiplexing in access networks was conceived in the late 1980s (Oakley, 1988) but significant advancement of key enabling technologies such as the athermal arrayed waveguide grating (AWG), WDM filter, reflective semiconductor optical amplifier (RSOA) and Fabry-Perot laser diode (FP-LD) has fuelled the resurgence of WDM-PON research in the last 10 years. A number of commercialized WDM-PON systems have been deployed to service customized business and wireless/wireline backhaul markets, e.g., (see Section Relevant Websites). In a WDM-PON, each end user is given dedicated capacity through the assignment of unique upstream and downstream wavelength channels. Concurrently, extended-reach PONs (ER-PONs) are being explored to consolidate metropolitan area and access networks (Davey *et al.*, 2009). Serving an increased number of clients, accommodating high bandwidth applications, and an increased network span, ER-PON deployments promise lower capital expenditure (CAPEX) and operational expenditure (OPEX) due to reduced number of active network interfaces and elements in the field. More recently, the use of orthogonal frequency division multiplexing (OFDM) in combination with WDM has been vigorously explored for application in the access segment (Cvijetic, 2012). In an OFDM-PON, data is carried by multiple orthogonal subcarriers, and a flexible 3D bandwidth allocation can be realized in time, frequency and modulation formats domain.

This review article is organized as follows. In Section Time-Division Multiplexed (TDM) Passive Optical Networks, we will introduce the early standardized TDM-PONs including BPONs, GPONs and GE-PONs. Then, the standardized 10G-PONs including ITU-T NG-PON1 and IEEE 10 G-EPON will be introduced in Section NG-PON1. In Section NG-PON2, the TWDM-PON structure based NG-PON2 will be introduced from the viewpoint of both physical layer and protocol layer. High capacity TWDM-PONs with advanced techniques to enable 100 Gb/s or beyond capacity will also be discussed. In Section Future PONs, future PONs with potential standardization possibility including WDM-PONs, extended-reach PONs and OFDM-PONs will be discussed. Finally, we will conclude this article in Section Summary and Conclusion.

Time-Division Multiplexed Passive Optical Networks

BPONs

The Broadband Passive Optical Network (BPON) was firstly developed by the Full Service Access Network (FSAN) working group in 1995 and was standardized by the International Telecommunications Union (ITU) in 2001 (ITU-T, 2008a). BPON is based on the asynchronous transfer mode (ATM) protocol and has a maximal downstream rate of 1244.16-Mbit/s and upstream rate of 622.08-Mbit/s. As discussed in Section Introduction, downstream signal is broadcast to all ONUs on the PON. To avoid collisions in the upstream direction, upstream transmission time slots from each ONU are controlled and granted by the OLT in the CO through Time Division Multiple Access (TDMA) (ITU-T, 2008a). The line code employed by BPON in both downstream and upstream transmissions is the Non-Return Zero (NRZ) modulation format. The typical operating wavelength for downstream and upstream is 1550 nm and 1310 nm, respectively.

The BPON MAC frame consists of different numbers of fixed size ATM cells, the actual number depending on the rate of transmission, as shown in **Fig. 2**. For the downstream MAC frame, a Physical Layer Operation, Administration, and Maintenance (PLOAM) is inserted at the beginning of every 28 ATM cells. PLOAMs are also used to carry grant information relating to an ONU's upstream access. Through PLOAM, an OLT in the CO can allocate and divide the upstream access bandwidth amongst different ONUs. Downstream transmission from the OLT is broadcast to every connected ONU. Each ONU will selectively detect its own cell according to the bandwidth allocation scheduled in OLT. In the upstream direction, transmission occurs in ATM cells with 3 overhead bytes per cell in its own time slot. Each ONU can only transmit its ATM cells during pre-determined time slots. Due to ONUs not being equidistance to the OLT, the OLT will receive burst-mode signals at differing power levels, each from a different ONU.

Though being the first-standardized PON, the uptake of BPON was slow and is currently rarely deployed due to the high cost per unit bandwidth. Further, the use of ATM cells within this MAC frame results in complex processing and large overhead (a 5-byte header to each 48-byte payload), thus limiting large-scale deployment.

Gigabit PONs

The Gigabit PON (GPON) was first standardised by the FSAN/ITU-T under ITU-T G.984 listing different aspects of the GPON such as general characteristics, physical medium, and management and control (ITU-T, 2008a). The standardized GPON supports data

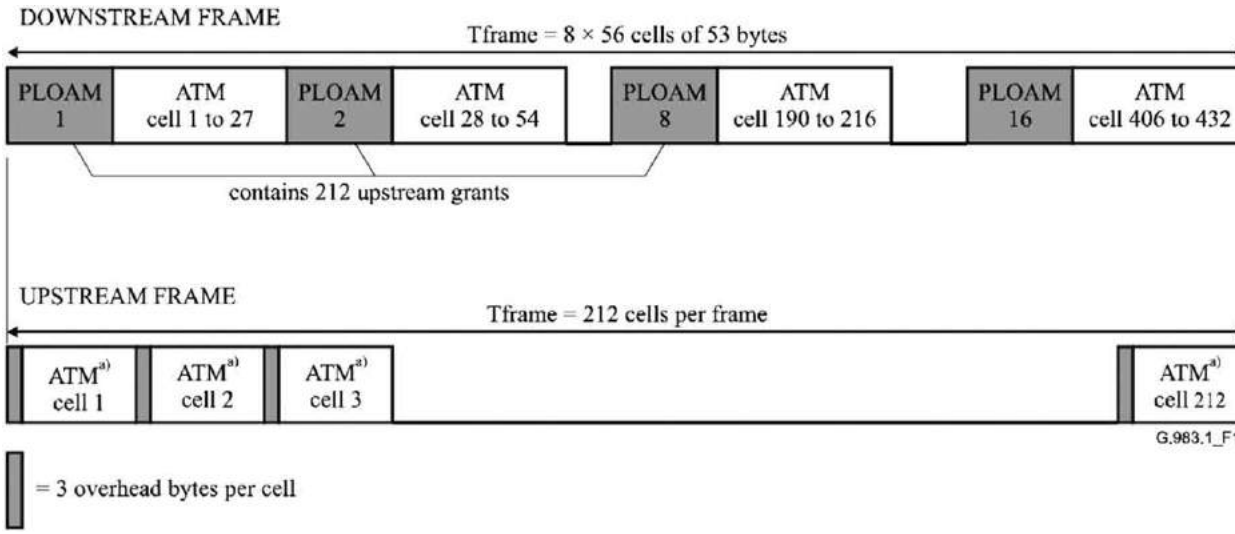


Fig. 2 Frame structure of BPON for 1244.16/622.08-Mbit/s. Adapted from ITU-T, 2008a. G.984 – Series recommendation, Gigabit-capable passive optical networks (G-PON).

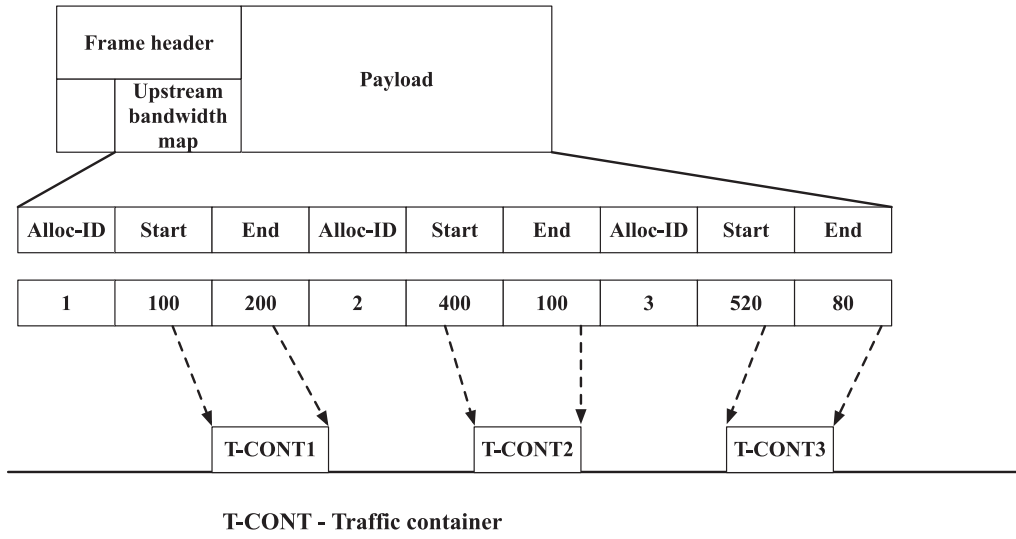


Fig. 3 Downstream GPON packet structure. Adapted from CommScope, 2013. EPON-GPON comparison,CommScope Solutions Marketing.

rates of 2.5 Gb/s and 1.25 Gb/s in the downstream and upstream directions, respectively. Today, GPON is one of the most popular technology of choice in some areas of Europe and North America.

Before entering the physical medium, Ethernet frames are encapsulated into GPON Transmission Control (GTC) layer encapsulation method (GEM) frames. GEM frames are, in turn, encapsulated into GPON Transmission Convergence (GTC) frame (in both upstream and downstream directions) (CommScope, 2013). **Fig. 3** illustrates a downstream GTC frame in detail (CommScope, 2013). In GPON, the GTC layer is responsible for aggregating traffic generated from different services, e.g., voice and video, into a common service-independent framework. The GTC header carries an upstream bandwidth map, which includes information on transmission time slots allocated to each ONU. In the GPON, the bandwidth grant is per traffic container (T-CONT). A T-CONT is an ONU object representing a group of logical connections that appear as a single entity for the purpose of upstream bandwidth assignment on the PON. For example, a GPON may have multiple T-CONTs per ONU, each dedicated for a different traffic class, e.g., best-effort and high priority. The allocation ID (Alloc-ID) is used to identify uniquely a T-CONT.

Fig. 4 illustrates the upstream frame structure of GPON (ITU-T, 2008b). The frame length is as same as that of the downstream, 125 μs and 19,440 bytes long. Each frame contains a number of transmissions from one or multiple ONUs, where the bandwidth map discussed under downstream frame structure schedules the arrangement of these transmissions. In any given upstream transmission, the ONU can transmit one of the four types of overheads and the payload. The four types of overheads are physical

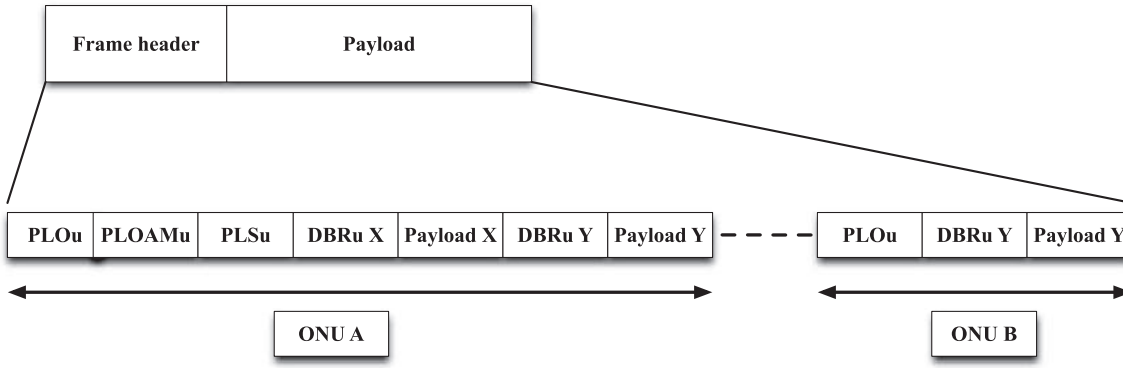


Fig. 4 Upstream GPON packet structure. Adapted from ITU-T, 2008b. G.984 x-series recommendations: Gigabit-capable passive optical networks (G-PON): Transmission convergence layer specification (G.984.3).

layer overhead (PLOu), physical layer operations administrations and management upstream (PLOAMu), power levelling sequence upstream (PLSu), and dynamic bandwidth report upstream (DBRu). The ONUs utilize the dynamic bandwidth report (DBRu) field, located in the header of the upstream GTC frame, to inform the queue length of each T-CONT to the OLT.

Gigabit Ethernet PONs

The IEEE standardized the Gigabit Ethernet PON (GE-PON) under IEEE 802.3 ah (IEEE, 2004). Integrated with conventional Ethernet technology, GE-PON enables seamless integration with IP and Ethernet technologies. The GE-PON supports 1.25 Gb/s link rate in both upstream and downstream directions and a 20 km network span. The deployment of GE-PON is most prevalent in the Asian region, especially in countries such as Japan and Korea.

In GE-PON, Ethernet packets are transmitted natively across the PON. As Ethernet technology is widely deployed in current local area networks (LANs), incorporating Ethernet into PONs requires minimal protocol conversion as far as hardware and technical know-how are concerned. As a TDM-PON, the ONUs in a GE-PON follow a time division multiple access technique, where the upstream bandwidth is divided into distinct time slots. In order to obtain bandwidth requirement information as well as to grant bandwidth from and to each ONU, a GE-PON employs the Multi-Point Control Protocol (MPCP). MPCP relies on control messages to facilitate two modes of operation: auto-discovery and normal operation. The primary MPCP control messages used for each mode of operation are listed below:

- Auto-discovery: GATE, REGISTER REQ, REGISTER, and REGISTER ACK.
- Normal operation: GATE and REPORT.

Auto-discovery

Fig. 5 illustrates the auto-discovery process, an initialization process implemented at both OLT and ONUs. Using auto-discovery, the OLT identifies the MAC addresses and round trip times of ONUs that have recently joined the network (IEEE:802.3ah). For this purpose, MPCP uses four control messages, GATE, REGISTER REQ, REGISTER, and REGISTER ACK. When an OLT discovers new ONUs, it allocates a discovery window, where no other registered ONU can transmit. A discovery GATE message with a local OLT time stamp will be broadcast, which only the uninitialized ONUs will respond to. When an ONU clock reaches the start time of the discovery window, it will wait a random amount of time and send its REGISTER REQ to the OLT. The random delay minimizes potential collisions caused by multiple uninitialized ONUs simultaneously sending REGISTER REQ messages. The REGISTER REQ message contains the MAC address and the local time stamp of the ONU, through which the OLT learns the round trip time to the ONU. Once the ONU is registered, the OLT will assign a unique identification, link layer ID (LLID), to the ONU. The OLT will then send the unique LLID through REGISTER to the ONU. This is then followed by a normal GATE message. Upon receiving the REGISTER and normal GATE messages, the ONU sends the REGISTER ACK to the OLT during the time slot allocated through the normal GATE message.

Normal operation

Once ONUs have established their connection with the OLT, bandwidth requests from and allocations to each ONU are performed under normal operation mode, as illustrated in Fig. 6 (Kramer, 2005). The GATE message, sent from the OLT to ONUs, is used to grant bandwidth (in the form of transmission time slots) to the ONUs. The GATE specifies the start time and the length of a transmission window and is sent per LLID. As discussed, an LLID which is attached to the preamble of the downstream packets, uniquely identifies an ONU in the network. The REPORT message, sent from each ONU to the OLT, contains bandwidth requirements of the ONUs. It is important to note that the standardized MPCP is not unique to a particular bandwidth allocation algorithm, rather, it is a supporting protocol that encourages the execution of different bandwidth allocation schemes in the GE-PON depending on latency, energy-efficiency, and capacity requirements of the network/service provider.

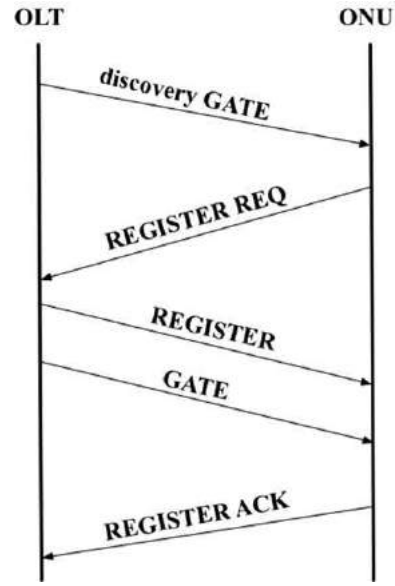


Fig. 5 Auto-discovery process of MPCP. IEEE:802.3ah, Ethernet in the first mile task force.

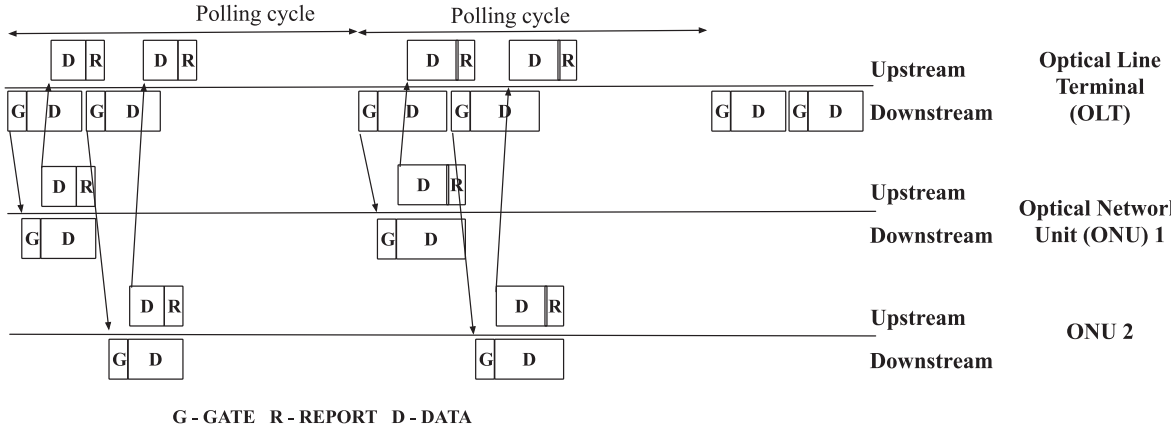


Fig. 6 Normal operation of MPCP. Reproduced from Kramer, G., 2005. Ethernet Passive Optical Networks. New York, NY: McGraw-Hill.

NG-PON1

Overview of 10GE-PON and XG-PON

2The IEEE and the ITU-T with the Full Services Access Network (FSAN) group have defined their respective 10 Gb/s solution to address future bandwidth growth over existing ODNs that exploit time-division multiplexing (TDM) in the downstream direction and time division multiple access (TDMA) in the upstream direction. These solutions are specified in the IEEE Std. 802.3av 10 GE-PON (IEEE, 2009) and ITU-T XG-PON (ITU-T, 2010) standards respectively. The 10 GE-PON defines symmetric (10.3 Gb/s) and asymmetric (10.3 Gb/s downstream and 1.25 Gb/s upstream) line-rate operations, the latter to support asymmetric traffic of IP video services. In contrast, XG-PON1 supports an asymmetrical bandwidth capacity of 10 Gb/s downstream and 2.5 Gb/s upstream (ITU-T, 2010). In XG-PON1, new features including enhancing security through authentication of management messages and minimizing energy consumption through powering down parts or all of the ONU are defined. Specific solutions to minimize the energy consumption include (a) powering down user network interfaces that are not actively in use (b) operating in "doze" mode whereby the ONU transmitter is powered down when the user has no real data to send, and (c) operating in "sleep" mode whereby both transmitter and receiver are powered down when the ONU is idle. A symmetric version of XG-PON1, known as XG-PON2 with 10 Gb/s downstream/upstream capacity is also specified (ITU-T, 2010).

Forward error correction (FEC) is defined in both standards to compensate for the reduced receiver sensitivity from using 10 Gb/s optical receivers. NRZ line coding in conjunction with FEC ((RS(248,216) block code) is mandatory in the downstream

direction for XG-PON systems but optional in the upstream direction (ITU-T, 2010). For 10 GE-PON, FEC ((RS(255,223) block code) is mandatory for the symmetric network. As for the asymmetric 10 GE-PON network, the upstream 1 Gb/s links can optionally use the IEEE GE-PON FEC (IEEE, 2009). To enhance coding efficiency in 10 GE-PON, 64b/66b line coding is used instead of the conventional 8b/10b coding in GE-PON.

Coexistence and Wavelength Allocation

Table 1 summarizes the main characteristics of currently standardized TDM-PONs discussed in Sections Time-Division Multiplexed (TDM) Passive Optical Networks and NG-PON1. Coexistence between these different standards over the same ODN can be achieved by observing the wavelength allocation schematically shown in Fig. 7. In the figure, the upstream and downstream wavelength bands for XG-PON, 10 GE-PON, GPON, GE-PON, and RF video overlay are indicated. Specifically, GPON and GE-PON use 1480–1500 nm for downstream transmission and 1550–1560 nm for RF video overlay (ITU-T, 2008a; IEEE, 2004). For upstream transmission, GPON uses the 1290–1330 nm waveband whereas GE-PON uses the entire O band. Both XG-PON and 10 GE-PON specify the O-minus band (1260–1280 nm) for upstream transmission and the L-band (1575–1580 nm) for downstream transmission (IEEE, 2009; ITU-T, 2010), thus facilitating coexistence with legacy systems and those implemented with RF video overlay.

NG-PON2

Overview of Time and Wavelength Division Multiplexed-PON

The standardization process of NG-PON2 was implemented by ITU-T study group 15 (SG15) (ITU, 2013). To meet the general requirement of 40-Gb/s downstream capacity, an underlying hybrid time and wavelength division multiplexing scheme is adopted in NG-PON2, giving rise to the time and wavelength division multiplexing PON (TWDM-PON). Fig. 8 depicts a typical schematic diagram of a TWDM-PON system (Luo *et al.*, 2013). Several wavelength pairs are multiplexed into a single PON system, whereby each wavelength is shared between multiple ONUs by employing time division multiplexing and multiple access mechanisms. NG-PON2 provides at least 40/10-Gb/s downstream/upstream capacity by stacking 4 TDM channel pairs with

Table 1 Comparison of standardized TDM-PONs

	BPON	*GPON*	*GE-PON*	*XG-PON1*	*10GE-PON*
Standard	ITU-T G.983	ITU-T G.984	IEEE803.2ah	ITU-T XGPON	IEEE 802.3av
Bandwidth					
Down	622 Mb/s	upto 2.5 Gb/s	Symmetric up to 1.25 Gb/s	10.3 Gb/s	10 Gb/s
Up	Up - 155 Mb/s	upto 2.5 Gb/s		10.3 Gb/s	1.25 Gb/s up to 10 Gb/s
Downstream, λ_d	1490 and 1550 nm	1490 nm and 1550 nm	1490 nm	1575–1580 nm	1575–1580 nm
Upstream, λ_u	1310 nm	1310 nm	1310 nm	1270 nm	1270 nm

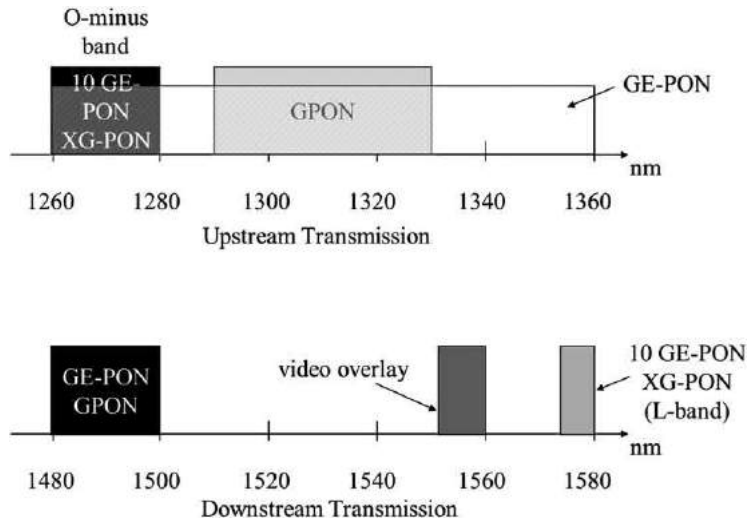


Fig. 7 Spectrum overlay of 10GE-PON and XG-PON showing coexistence with GE-PON and GPON. Reproduced from Wong, E., 2012. Next-generation broadband access networks and technologies. IEEE/OSA J. Lightw. Technol. 30(4), 597–608.

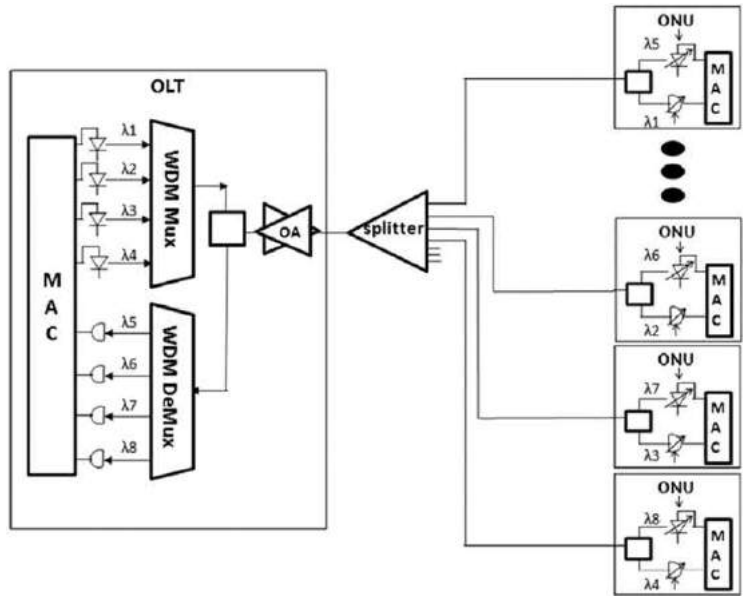


Fig. 8 TWDM-PON system diagram. Adapted from Luo, Y., Sui, M., Effenberger, F., 2012. Wavelength management in time and wavelength division multiplexed passive optical networks (TWDMPONs). In: Proceedings of IEEE Global Telecommunications Conference (GLOBECOM).

10-Gb/s downstream and 10/2.5-Gb/s upstream capacity per channel. For seamless network upgrade and user migration, the optical technologies specified for NG-PON2 must be compatible to legacy PON systems, including the coexisting wavelength band plan and reusing the power splitting ODNs. As such, L band (1596–1603 nm) and C-band (1524–1544 nm) have been assigned for downstream and upstream transmission, respectively (ITU, 2013).

In addition, PtP-WDM channel overlay is defined in NG-PON2 for supporting services such as fronthaul and backhaul wireless network transmission with an expanded spectrum option of 1524–1625 nm and a shared spectrum option of 1603–1625 nm (ITU, 2013). Regardless, all specified PtP-WDM wavebands must be chosen from unused spectrum by the TWDM-PON and legacy systems. As for link budget, four ODN path loss classes are specified, namely Class N1, N2, E1 and E2, which corresponds to 29 dB, 31 dB, 33 dB and 25 dB respectively. The line code for both downstream and upstream links is scrambled NRZ. To improve receiver sensitivity, FEC support is mandatory at both OLT channel termination and ONU. For 2.5 Gb/s and 10 Gb/s, the FEC codes are RS (248, 332) and RS (255, 223), respectively.

In view of the multi-wavelength operation property of TWDM-PON, additional network and protocol requirements beyond that of legacy PON systems, are critical in TWDM-PONs, e.g., the ability to tune the ONU transmitter and receiver. In the upstream direction, the ONU is tuned to emit at the desired wavelength channel. In the downstream direction, a tunable ONU receiver is required to select the proper wavelength channel. Also, to eliminate the chromatic dispersion induced signal distortion, electrical dispersion compensation (EDC) may be used in the OLT/ONU to achieve the specified loss budget.

Energy-Efficient Dynamic Bandwidth Allocation

Aside from its increased bandwidth capacity, TWDM-PON is also an attractive solution from an energy perspective (Valcarenghi *et al.*, 2014; Dixit *et al.*, 2012). Due to multiple line cards and wavelengths channels, the power consumption of the OLT is comparatively higher in TWDM-PON compared to networks specified under NG-PON1. Nevertheless, the large number of ONUs serviced by the TWDM-PON results in lower energy per user. Considering its expected uptake, future-proofing the TWDM-PON in terms of energy-efficiency is becoming an important deployment factor.

As discussed before, a TWDM-PON consists of multiple wavelengths channels, requiring tunable ONU transceivers (TRXs) that can tune to any of the wavelength channels present in the network. Due to this tunability, at low network loads, the network may be reconfigured by switching off idle wavelengths and thereby saving energy at the OLT (Valcarenghi *et al.*, 2014). In addition, conventional power-saving methods such as sleep and doze can be implemented at the ONUs to improve overall energy-efficiency (Dixit *et al.*, 2012). As a result, recent energy-efficient solutions for TWDM-PON including dynamic wavelength and bandwidth allocation (DWBA) algorithms incorporating wavelength reconfiguration and sleep/doze mode operations, have been proposed.

One of the most important aspects of network reconfiguration through wavelength reallocation is determining how many wavelengths to be switched off without affecting the system performance. Optimizing the number of active wavelengths, subjected to different parameters of the network such as wavelength tuning time (Luo *et al.*, 2012), energy consumption (Yang *et al.*, 2011), or distance from the OLT (Dixit *et al.*, 2013), have been proposed as a result. Energy-saving techniques such as wavelength

reconfiguration and sleep/doze mode operations on the other hand, result in an increase in the average delay experienced by transmission packets. If the tuning time is assumed to be negligible, unlimited wavelength switching could be practised in a network. Although unlimited wavelength switching may theoretically result in significant energy-savings, in reality, the tuning time of an ONU is significant. In fact, the tuning time of an ONU transmitter can range from nanosecond to sub-second order while the ONU receiver tuning time can range from nanosecond to second order based on implementation (Asaka, 2014). It has also been shown that the ratio between the tuning time and the polling cycle time (Kondepu *et al.*, 2014), and the OLT's awareness of the wavelength tuning time (Buttaboni *et al.*, 2014) are strongly correlated to performance matrices. As such, when considering a TWDM-PON with non-zero wavelength tuning time, the improved energy-savings through wavelength reconfiguration will be achieved at the expense of increased average delay of the network. To overcome this shortfall, using the average delay as a constraint in wavelength optimization (Xu *et al.*, 2015) and changing the polling sequences based on the wavelength tunning time (Wang *et al.*, 2015), have been reported in literature.

Introducing sleep/doze mode operations into a wavelength-reconfiguring algorithm increases the average delay further. In the energy-efficient DWBA algorithm reported in Dias *et al.* (2015), the Quality-of-Service degradation is minimized using Bayesian estimation and prediction techniques. The proposed DWBA algorithm considers a delay-constrained TWDM-PON and determines the maximum polling cycle time, $T_{poll\ max}$ that satisfies a given delay constraint. The $T_{poll\ max}$ is considered to achieve maximum energy-savings at the ONUs through sleep/doze mode operations. Based on this $T_{poll\ max}$, the number of active wavelengths is determined and the remaining idle wavelengths are switched off to achieve energy-savings at the OLT. Under the proposed framework, $T_{poll\ max}$ for both online and offline DWBA algorithms are determined. Under the online DWBA algorithm, the OLT uses average inter-arrival time of packets, estimated using Bayesian estimation at the ONUs, to predict the accumulated bandwidth at each ONU. The proposed traffic estimation and prediction prevents the ONUs from waiting extra cycles, before being allocated any bandwidth for transmission. **Fig. 9** illustrates the traffic flow of the proposed DWBA algorithm.

High Capacity TWDM-PONs

To accommodate the anticipated increase in customer bandwidth demand, NG-PON2 systems may support an optional extension for services in excess of 10 Gb/s upstream and/or downstream towards one PON client, as was specified in G989.1 amendment 1 (ITU, 2013). It was defined in the recommendation G.989 series that NG-PON2 system supports multiple wavelength channels and enables flexibility to add capacity as bandwidth demand is expected to grow to 100 Gbit/s and beyond (ITU, 2013). The cost-effective solution for capacity enhancement, such as realizing 25 Gb/s per channel is under hot discussion. Generally, 10 G-class optical devices have been preferred for cost control, thereby necessitating advanced modulation formats with higher spectral efficiency. Four-level pulse amplitude modulation (PAM-4) (Wei *et al.*, 2015) and duobinary (Li *et al.*, 2015) formats are the most promising format candidates, and based on these two modulation formats, many experimental demonstrations of 10 G-class devices enabled 25 Gb/s modulation have been reported. As the demodulation circuit for duobinary and PAM-4 formats are not yet commercially viable, signals are sampled by real-time oscilloscope and processed offline in most of these demonstrations. Several research groups have reported real-time demodulation for duobinary and PAM-4 formats, although further improvements are required for application in practical PON systems. To date, the reported real-time 25-Gb/s PON system demonstrations are mostly based on optical duobinary (ODB) format. Though the ODB format is robust to fiber dispersion, it necessitates costly Mach-Zender Modulators (MZMs) and wide-band receivers for modulation and detection.

Consequently, efforts in increasing the modulation bandwidth as well as decreasing the cost of both optical transmitters and receivers are underway. Related works are being reported, e.g., 30-GHz directly modulation DFB laser (Zhang *et al.*, 2015), 25-Gb/s

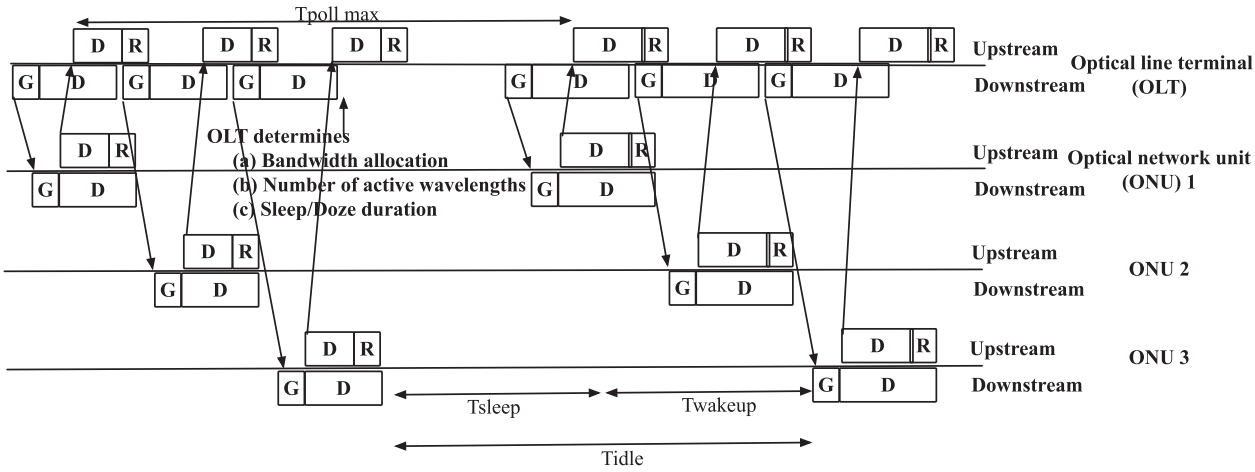


Fig. 9 Traffic flow of the OFF-DWBA algorithm. D-Data, R-REPORT and G-GATE. Dias, M.P.I., Van, D.P., Valcarenghi, L., Wong, E., 2015. An energy-efficient framework for wavelength and bandwidth allocation in TWDM PON. J. Opt. Commun. Netw. 7(6) 496–504.

Si-Ge APD (Huang *et al.*, 2016), and bandwidth improvement of CMOS-APD (Lee *et al.*, 2015), etc. Signal distortion from fiber chromatic dispersion is another issue to be handled, especially for high-speed signal, and some institutes are studying the proper dispersion compensation techniques for high data rate signal. With the development of wide-band transceivers and practical dispersion compensation techniques, NRZ format can still be a potential candidate for high capacity PON applications.

Future PONs

Wavelength Division Multiplexed PONs

The WDM-PON has been touted as a long-term access solution beyond TDM-PONs. The system operation of a WDM-PON is entirely different to the TDM-PON whereby in the former each ONU located in a home/office/building is assigned its own upstream and downstream wavelength channel. This is contrary to more a traditional PON architectures where one optical feed is shared among 32 or more users. In that case, each home operates at the same wavelength, and is allotted a 1/32nd time slot on the main fiber. In a WDM-PON, each home/office/building is assigned its own wavelength and has continual use of the fiber at that wavelength. However, virtually all PON technologies rely on some form of wavelength division multiplexing (WDM) to enable bi-directional communications. For example, in a typical GPON system, the upstream communication runs at 1310 nm wavelength, while the downstream traffic is transmitted at 1490 nm. A third wavelength at 1550 nm is used for video overlay. The utilization of WDM in PON systems is therefore already very commonplace.

Fig. 10 illustrates a typical WDM PON comprising a CO with an OLT, two cyclic AWGs, a trunk or feeder fiber, a series of distributions fibers, and ONUs at the subscriber premises (Wong, 2012). Downstream wavelengths and upstream wavelengths from the ONUs are multiplexed and demultiplexed in the AWG located at the CO. These wavelength channels also under (de) multiplexing at the AWG in the remote node located closer to the ONUs. The trunk fiber carries the multiplexed downstream and upstream wavelengths between the first and second AWG. The distribution fiber carries the demultiplexed wavelengths to and from the remote node to the ONUs. Both AWGs have the same free spectral range (FSR). Note that the downstream and upstream wavelengths allocated to each ONU are intentionally spaced at a multiple of the FSR, allowing both wavelengths to be directed in and out of the same AWG port that is connected to the destination ONU. For example, the downstream wavelength destined for ONU1, denoted λ_1, and the upstream wavelength from ONU1 denoted $\lambda_{1'}$, are spaced a multiple of the FSR. In a typical WDM PON, wavelength channels are spaced 100 GHz (0.8 nm) apart. In systems classified as dense WDM-PON (DWDM), a channel spacing of 50 GHz or less is deployed.

There are several advantages to the WDM-PON architecture over more traditional TDM-PON systems. Foremost is the dedicated bandwidth available to each ONU. Secondly, WDM-PONs can typically have improved security and scalability features, because each ONU only transmits and receives on assigned wavelength channels. Thirdly, the MAC layer in a WDM-PON is simplified, since the logical topology of a WDM-PON is a PtP topology between the OLT and each ONU, and thus does not require the use of PtMP media access controllers found in conventional TDM-PON networks. Finally, each wavelength in a WDM-PON network is effectively a PtP link, allowing each link to operate at a different speed and if required a different MAC protocol for maximum flexibility and pay-as-you-grow upgrades.

The main challenge with WDM-PON is the cost. Since each ONU is assigned his own wavelength, this suggests that the OLT must transmit on different wavelengths versus one shared wavelength as found in more traditional TDM-PON systems. Likewise, it requires that each of the ONUs on a link operate at a separate wavelength, suggesting that every ONU requires a wavelength-specific laser that can operate at its assigned wavelength. Since each ONU is assigned a unique upstream wavelength channel,

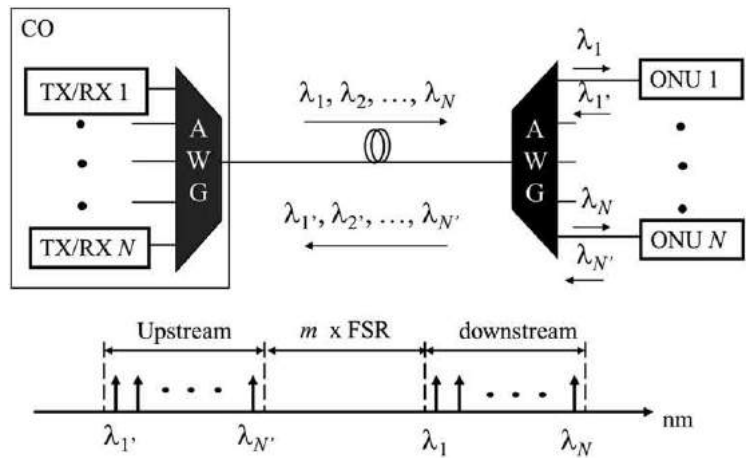


Fig. 10 WDM-PON. Inset: Allocation of upstream/downstream wavelength into two separate wavebands. Reproduced from Wong, E., 2012. Next-generation broadband access networks and technologies. IEEE/OSA J. Lightw. Technol. 30(4), 597–608.

distinct wavelength transmitters must be deployed at the subscriber premises. The simplest solution is to utilize fixed wavelength transmitters with wavelength tunability, e.g., tunable distributed feedback (DFB) laser (Suzuki *et al.*, 2007), tunable vertical cavity surface emitting lasers (VCSELs) (Wong *et al.*, 2006), and distributed Bragg reflector laser diodes (DBR-LDs), such as sampled grating DBR-LDs and super structure grating DBR-LDs (Kuwano *et al.*, 2001; Ishii *et al.*, 1996).

Alternatively, wavelength reuse schemes that eliminate optical sources at the ONUs have been proposed. Aside from carrying downstream signals, the downstream wavelength is used to wavelength seed a reflective semiconductor optical amplifier (RSOA) located at the designated ONU. The RSOA is intentionally operated in the gain saturation region whereby the RSOA amplitude squeezing effect can be exploited to erase the modulation on the seeding downstream wavelength (Katagiri *et al.*, 1999). Wavelength reuse schemes eliminate the need for seeding sources, is less costly than using tunable lasers, and allows direct modulation of the RSOA for upstream transmission. The drawback is severe performance degradation at the CO due to interference between the newly modulated upstream data and residual downstream data. A solution to improve transmission performance is through using orthogonal upstream and downstream modulation formats (Garces *et al.*, 2007) and line coding approaches (Kim *et al.*, 2007; Al-Qazwini and Kim, 2011).

In addressing the potential large inventory and cost of wavelength specific sources, researchers have been concentrating on developing cost-efficient and wavelength independent sources termed "colorless" sources. These sources include directly modulated light-emitting diode (LED) and super-luminescent diode (SLD) (Reeve *et al.*, 1988) which broadband spectrum from each ONU is spectrally sliced by the AWG in the upstream direction. These sources are limited in terms of system reach and modulation bit rate. In yet another category of colorless sources, optical light originating from the CO is fed into the ONUs to injection-lock Fabry-Perot laser diodes (F-P LDs) (Kim *et al.*, 2000) or to wavelength-seed RSOAs (Feuer *et al.*, 1996). The wavelength seeding scheme is identical to the injection-locking scheme except for the use of an RSOA which amplifies and modulates the incoming continuous wave (CW) light. The drawback of these schemes lie in the requirement of additional broadband light source(s), limited transmission bit-rate due to device limitation, and susceptibility to transmission performance impairment from fiber dispersion, Rayleigh backscattering noise, and broadband amplified spontaneously emission (ASE) noise from the broadband light source and the colorless source.

More recently, self-seeding of the RSOA was proposed in which each RSOA is self-seeded by its own spectrally-sliced CW light (Wong *et al.*, 2007). ASE light emitted from each RSOA is spectrally-sliced by the AWG located at the remote node and feedback to the RSOA. Self-seeding removes the need for active temperature tracking between the optical components within the remote node and between the remote node and each RSOA and identical RSOAs can be placed at all ONUs.

Extended-Reach PONs

ER-PONs can serve an increased number of subscribers, accommodate high bandwidth applications, and an increased network span that consolidates metro and access networks. This consolidation lowers capital expenditure (CAPEX) and operational expenditure (OPEX) due to a reduction in the number of active network interfaces and elements in the field. The idea of extending the reach of a PON is not new with many significant demonstrations and field trials carried out in the 1990s (Mestdagh and Martin, 1996). However, advances in optical amplification and WDM technologies in recent years along with the volume manufacturing of optical components such as SOAs, AWGs, and WDM filters, have lowered the cost of ER-PONs to a point that is competitive in the business and access market.

An illustrative example of an ER-PON is shown in **Fig. 11** whereby multiple COs which were previously connected to a metropolitan aggregation ring network (dash line) are now consolidated at the main central office (MCO) to support a large

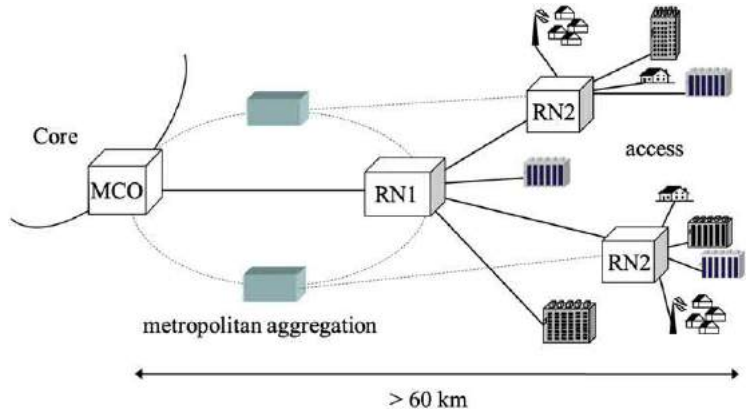


Fig. 11 Schematic diagram of a two stage extended-reach PON (ER-PON) showing consolidation of central offices (COs) previously connected to the metropolitan aggregation ring network at the main central office (MCO). Reproduced from Wong, E., 2012. Next-generation broadband access networks and technologies. IEEE/OSA J. Lightw. Technol. 30(4), 597–608.

number of customers via remote nodes. The ODN of a typical ER-PON comprises an extended-reach FF which connects the MCO to remote node 1 (RN1), DFs which connect RN1 to remote node 2 (RN2), and last mile fibers (LMFs) that connect RN2s to customers. Bandwidth-dedicated business customers and mobile customers can be connected directly to RN1 whereas residential users can be connected to RN2. By exploiting a combination of (a) TDM or WDM in the first stage, and (b) TDM in the second stage, system reach can be extended from the conventional 20 km to 60–100 km while maintaining a 1:32 or higher split ratio. In case where the first stage is based on TDM, a single upstream and downstream wavelength support all users with power splitting carried out in both RN1 and RN2s. When the first stage is based on WDM, wavelengths are multiplexed at the MCO and transmitted simultaneously through a feeder fiber to RN where optical amplification and demultiplexing take place. In turn, in the second stage of the ER-PON, each demultiplexed wavelength is then further power split at RN2 and then transmitted to multiple customers. A reach extender is typically housed in RN1 to compensate for power loss due to the long transmission distance and high split ratio.

A few reach extended technologies have been proposed to compensate for the insertion loss of ER-PONs. These include optical amplification at the MCO and/or RN1, using erbium doped fiber amplifier (EDFA) (Townsend *et al.*, 2008), SOA (Suzuki and Nakagawa, 2005), and distributed Raman amplification (DRA) (Kjaer *et al.*, 2006). EDFAs provide excellent gain power and noise performance in the C and L-bands. However, optical amplification is limited to these wavelength bands and unless gain control through optical gain clamping or pump power variation is implemented, the relatively slow speed in adjusting the EDFA gain makes this option disadvantageous to the bursty nature of upstream TDMA traffic. The SOA, on the other hand, benefits from the fact that it can operate at any wavelength of interest including the O band and with better gain dynamics than the EDFA. Other benefits include compactness and the ability to facilitate additional functionalities such as wavelength conversion and all-optical regeneration. However, the SOA operates on a per-channel basis. The inability to provide simultaneous amplification across multiple channels is its main drawback. The use of DRA in counter and co-propagation configurations enables amplification across a wide wavelength region over a bidirectional fiber link. Raman amplifiers can be tailored to provide a flat optical gain bandwidth that encompasses wavelengths exceeding those of common optical amplifiers. Using simultaneous launching of multiple pump wavelengths, DRA can achieve a flat Raman gain of up to 100 nm of bandwidth. Combining the benefits of EDFA and SOA, DRA allows simultaneous multichannel amplification, fast gain dynamics, and the ability to provide gain at any wavelength contingent on the pump wavelength. Using DRA does however require high pumping power, thereby adding to the power consumption at the RN1 and MCO.

In replacement of optical amplification, electronic repeater(s) can be implemented at RN1 to facilitate 1R or 2R regeneration of both upstream and downstream signals (Davey *et al.*, 2009). In using electronic repeater(s), benefits include the option of wavelength conversion, bit-error-rate monitoring, and optical power equalization of the burst mode signals in the upstream direction. However, a drawback of electronic repeaters is the need for bit-rate specific burst mode receivers that must also be capable of handling a wide dynamic range.

Orthogonal Frequency Division Multiplexing PON

Orthogonal frequency division multiplexing (OFDM) is a modulation technique which is now used in most new and emerging broadband wired and wireless communication systems (Cvijetic, 2012). In OFDM, the spectra of individual subcarriers overlap as depicted in **Fig. 12(a)**, but the signals on each subcarrier are mathematically orthogonal over one OFDM symbol period. As such, the subcarriers can be demodulated without interference. Dispersion-induced signal distortion can be eliminated by adding a cyclic prefix (CP) to the start of each time domain OFDM symbol before transmission, i.e., a number of samples from the end of the symbol is appended to the start of the symbol as shown in **Fig. 12(c)** (Armstrong, 2009), by which any distortion caused by a linear dispersive channel can be corrected simply using a 'single-tap' equalizer. Also, by employing bit-loading techniques, the modulation formats on each subcarrier can be different according to the channel response. Combined with the time- and frequency-domain multiplexing as shown in **Fig. 12(b)** (Qiu *et al.*, 2011), data can be assigned to end-users by different time slots or subcarriers with different modulation formats, thereby supporting a flexible 3-dimensional (3D) dynamic bandwidth allocation algorithm.

Despite advantages such as high spectral efficiency, high dispersion tolerance, and flexible bandwidth allocation, thus making OFDM a suitable modulation technique for future high-capacity PON systems, its drawbacks mainly due to high peak to average power ratio (PAPR) and high sensitivity to phase noise and frequency offset, easily degrades performance and limits its practical application. Digital signal processing (DSP) algorithms are being proposed to reduce PAPR (Popoola *et al.*, 2014) and to mitigate signal distortion caused by phase noise and frequency offset (Yi *et al.*, 2008). However, as high-speed ADC/DAC and complex DSP algorithms are required for the generation, demodulation and equalization of high-speed OFDM signal, most of the experimental demonstrations are realized off-line. Further, the capacity of the demonstrated real-time systems is generally lower than 10-Gb/s per channel. Another drawback of OFDM modulation is its low sensitivity. With the modulation format of 16-QAM or even higher, the sensitivity of directly-detected signal is much lower than the traditional NRZ format, thereby severely limiting the loss budget of the PON system. To solve this problem, coherent detection is introduced in OFDM-PONs to improve receiver sensitivity. As a result, the loss budget is increased at the cost of a higher system complexity.

So in the short term, limited by the cost of high-speed ADC/DAC and the complex DSP modules, OFDM-PONs are still considered an expensive option. However, with the emergence of novel applications, low spectral-efficiency formats such as NRZ

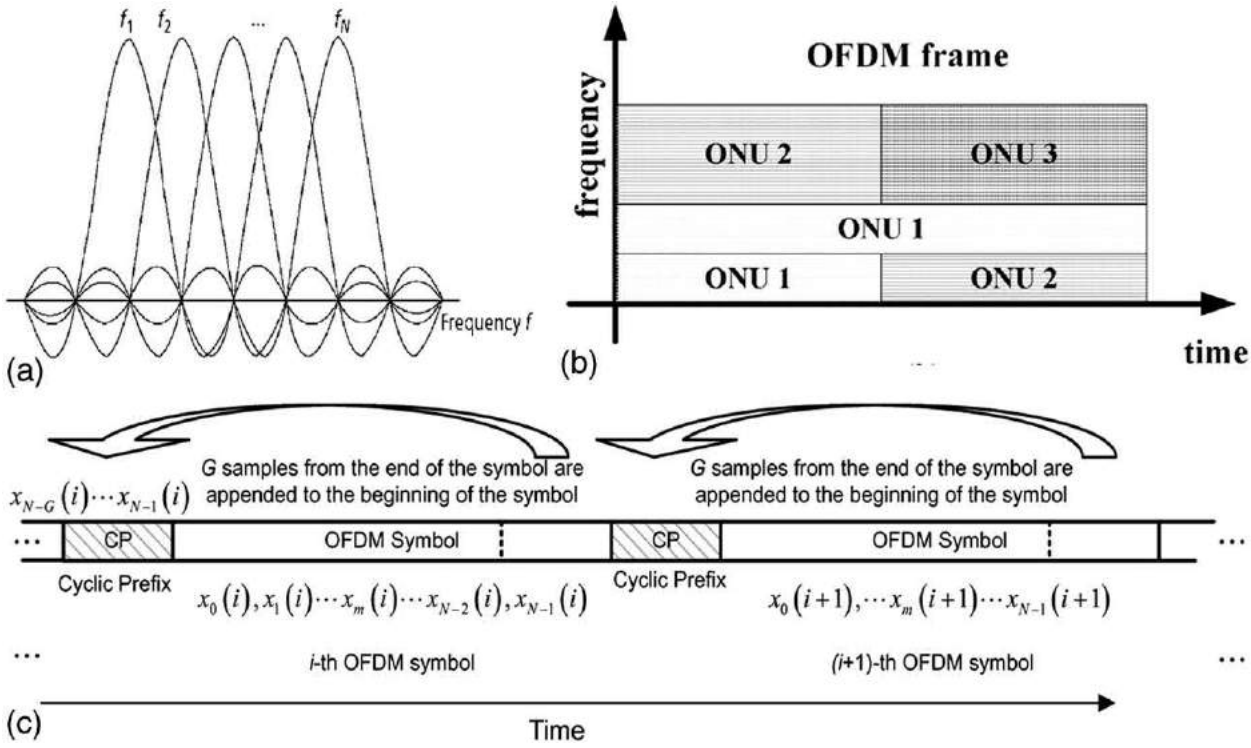


Fig. 12 (a) OFDM spectrum for N orthogonal frequency domain subcarriers; (b) frequency and time domain partitioning of an OFDM frame. (c) Time domain sequence of OFDM symbols showing the cyclic prefix. Adapted from Armstrong, J., 2009. OFDM for optical communications. J. Lightw. Technol. 27(3), 189–204 and Qiu, K., Yi, X., Zhang, J., *et al.*, 2011. OFDM-PON optical fiber access technologies. In: Asia Communications and Photonics Conference Optical Society of America, 1–9.

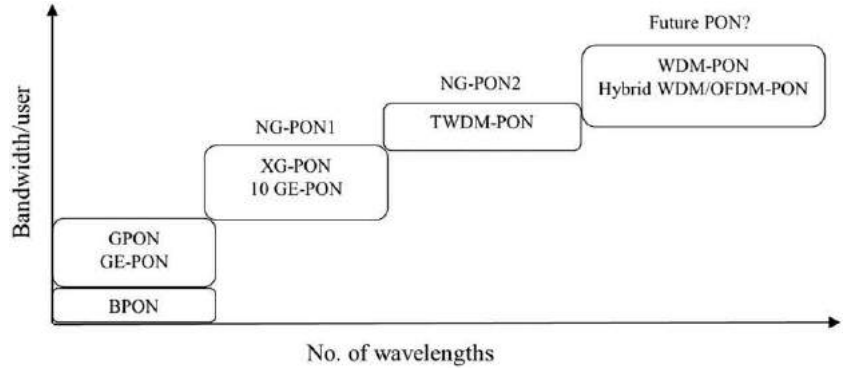


Fig. 13 Evolution of broadband passive optical access networks.

and RZ, will not be able to satisfy the increase in bandwidth demand and ultimately, and the spectral-effective OFDM-PON will be a solution by then.

Summary and Conclusion

A brief overview of the evolution of passive optical networks with particular focus on standardized systems mentioned, and the emerging trends to address future bandwidth needs, was discussed in this article. This summary of the evolution of PONs is illustrated in Fig. 13. Driven by the increasing user bandwidth demand for emerging applications, PONs with increasing capacity and cost-efficiency have been deeply discussed and studied around the world. BPON, GPON, XG-PON have been standardized by ITU, while GE-PON and 10GE-PON have been standardized by IEEE. These technologies are all developed by gradually increasing the data rate. The main reasons behind the push for these TDM/TDMA PON systems are to extend the longevity of existing ODNs and to allow co-existence with the current generation PONs such that the operational impact on existing users will be minimized.

Due to the chromatic dispersion and cost issues, only increasing the date rate per wavelength is more difficult and encounters some bottlenecks especially when the bit rate is beyond 10 Gb/s. In order to increase the system capacity, NG-PON2 employs the technique of stacking wavelength, called as TWDM-PON, which has been standardized by ITU-T in 2015 and is supposed to be commercially available in around 2018. From 2016, IEEE next generation Ethernet passive optical network (NG-EPON) has also been discussed in-depth to provide a capacity of 100-Gb/s with 25-Gb/s per wavelength, which has attracted extensive attention from the industry. In the future, in order to satisfy the bandwidth requirement, more advanced technologies will be proposed to upgrade the system capacity such as OFDM-PON and WDM-PON. High-speed PON with higher data rate per wavelength is also an effective solution to increasing the system capacity. So the combination of stacking more wavelengths and increasing the date rate of per wavelength or using complicated technologies such as coherent detection and digital signal processing will be effective solutions to overcome the problem of system capacity. In order to successfully deploy these technologies, the implementation complexity must be minimized to a level that is comparable to existing commercialized systems and with a cost that is sufficiently low to meet the cost constraints of the access market. In all PON with higher capacity and lower cost will be the ultimate goal from both industry and academia.

See also: Passive Optical Components

References

CommScope, 2013. EPON-GPON comparison,CommScope Solutions Marketing.

Al-Qazwini, Z., Kim, H., 2011. Line coding for downlink DML modulation in lambda-shared, RSOA-based asymmetric bidirectional WDM PONs. In: Proceedings Optical Fiber Communication Conference National Fiber Optic Engineers Conference (OFC/NFOEC), paper OMP5.

Armstrong, J., 2009. OFDM for optical communications. J. Lightw. Technol. 27 (3), 189–204.

Asaka, K., 2014. What will be killer devices and components for NG-PON2. In: Proceedings of the 40th European Conference and Exhibition on Optical Communication (ECOC).

Buttaboni, A., Andrade, M.D., Tornatore, M., 2014. Dynamic bandwidth and wavelength allocation with coexistence of transmission technologies in TWDM PONs. In: Proceedings of the 16th International Telecommunications Network Strategy and Planning Symposium (Networks).

Cvijetic, N., 2012. OFDM for next-generation optical access networks. J. Lightw. Technol. 30 (4), 384–398.

Davey, R., Grossman, D.B., Rasztovits-Wiech, M., *et al.*, 2009. Long-reach passive optical networks. J. Lightw. Technol. 27, 273–291.

Dias, M.P.I., Van, D.P., Valcarenghi, L., Wong, E., 2015. An energy-efficient framework for wavelength and bandwidth allocation in TWDM PON. J. Opt. Commun. Netw. 7 (6), 496–504.

Dixit, A., Lannoo, B., Colle, D., Pickavet, M., Demeester, P., 2012. ONU power saving modes in next generation optical access networks: Progress, efficiency, and challenges. Opt. Express 20, B52–B63.

Dixit, A., Lannoo, B., Colle, D., Pickavet, M., Demeester, P., 2013. Flexible TDMA/WDMA passive optical network: Energy-efficient next-generation optical access solution. J. Opt. Switch.Netw. 10, 491–506.

Feuer, M.D., Wiesenfeld, J.M., Perino, J.S., *et al.*, 1996. Single-port laser-amplifier modulators for local access. IEEE Photon. Technol. Lett. 8, 1175–1177.

Förster, M., 2006. Worldwide Development of FTTH. Available at: https://www.hft-leipzig.de/fileadmin/image_hftl/presse/Institut_HF/FTTH_Foerster.pdf.

Garces, I., Aguado, J.C., Martinez, J.J., *et al.*, 2007. Analysis of narrow-FSK downstream modulation in colorless WDM PONs. Electron. Lett. 43 (8), 471–472.

Huang, Z., Li, C., Yu, K., *et al.*, 2016. A 25Gbps low-voltage waveguide Si-Ge avalanche photodiode. In: CLEO: Science and innovations, Optical Society of America, STh4E. 6.

IEEE:802.3ah, Ethernet in the first mile task force.

IEEE, 2004. Standard 802.3ah – IEEE Standard for Local and Metropolitan Area Networks - Specific requirements - Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) access method and physical layer specifications Amendment: Media Access Control Parameters, Physical Layers, and Management Parameters for Subscriber Access Networks.

IEEE, 2009. Standard 802.3av – IEEE Standard for Information technology - Telecommunications and information exchange between systems – Local and metropolitan area networks - Specific requirements Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications Amendment 1: Physical Layer Specifications and Management Parameters for 10 Gb/s Passive Optical Networks.

Ishii, H., Tanobe, H., Kano, F., *et al.*, 1996. Quasicontinuous wavelength tuning in super-structure-grating (SSG) DBR lasers. IEEE J. Quantum Electron. 32 (3), 433–441.

ITU, 2013. 40-Gigabit-capable passive optical networks (NG-PON2): General requirements, ITU-T Recommendation G.989.1.

ITU-T, 2005. G.983.1 Recommendation: Broadband optical access systems based on passive optical networks (PON).

ITU-T, 2008a. G.984 – Series recommendation, Gigabit-capable Passive Optical Networks (G-PON).

ITU-T, 2008b. G.984 x-series recommendations: Gigabit-capable passive optical networks (G-PON): Transmission convergence layer specification (G.984.3).

ITU-T, 2010. G.987.x – Series recommendation: 10-Gigabit-capable passive optical network (XG-PON).

Katagiri, Y., Suzuki, K., Aida, K., 1999. Intensity stabilisation of spectrum-sliced Gaussian radiation based on amplitude squeezing using semiconductor optical amplifiers with gain saturation. Electron. Lett. 35 (16), 1362–1364.

Kim, H.D., Kang, S.-G., Lee, C.-H., 2000. A low-cost WDM source with an ASE injected Fabry-Perot semiconductor laser. IEEE. Photon. Technol. Lett. 12, 1067–1069.

Kim, S.Y., Jun, S.B., Takushima, Y., Son, E.S., Chung, Y.C., 2007. Enhanced performance of RSOA based WDM PON by using Manchester coding. OSA J.Opt. Netw. 6, 624.

Kjaer, R., Monroy, I.T., Oxenlowe, L.K., Jeppesen, P., Palsdottir, B., *et. al.*, 2006. Bi-directoinal 120 km long reach PON link based on distributed Raman amplification. In: Proceedings IEEE LEOS Annual Meeting, paper WEE3, Oct.

Kondepu, K., Valcarenghi, L., Van, D.P., Castoldi, P., 2014. Impact of ONU tuning time in TWDM-PON with dynamic wavelength and bandwidth allocation: An FPGA-based evaluation. In: Proceedings of the 40th European Conference and Exhibition on Optical Communication (ECOC).

Kramer, G., 2005. Ethernet Passive Optical Networks. New York, NY: McGraw-Hill.

Kuwano, S., Teshima, M., Uematsu, H., Iwatsuki, K., 2001. WDM optical packet transmission experiment over 235 km of installed fibers. In: Proceedings Optical Fiber Communication Conference National Fiber Optic Engineers Conference (OFC/NFOEC), TuK4, Mar.

Lee, M.-J., Lee, J.-M., Rucker, H., Choi, W.-Y., 2015. Bandwidth improvement of CMOS-APD with carrier-acceleration technique. IEEE Photon. Technol. Lett. 27 (13), 1387–1390.

Li, Z., Yi, L., Wang, X., *et al.*, 2015. 28 Gb/s duobinary signal transmission over 40 km based on 10 GHz DML and PIN for 100 Gb/s PON. Opt. Express 23 (16), 20249–20256.

Luo, Y., Sui, M., Effenberger, F., 2012. Wavelength management in time and wavelength division multiplexed passive optical networks (TWDMPONs). In: Proceedings of IEEE Global Telecommunications Conference (GLOBECOM).
Luo, Y., Zhou, X., Effenberger, F., *et al*., 2013. Time-and wavelength-division multiplexed passive optical network (TWDM-PON) for next-generation PON stage 2 (NG-PON2). J. Lightw. Technol 31 (4), 587–593.
Mestdagh, D.G., Martin, C.M., 1996. The super-PON concept and its technical challenges. Broadband Commun.
Oakley, K.A., 1988. An economic way to see in the broadband dawn (passive optical network). In: Proceedings of IEEE Global Telecommunications Conference and Exhibition vol. 3, pp. 1574–1578.
Popoola, W.O., Ghassemlooy, Z., Stewart, B.G., 2014. Pilot-assisted PAPR reduction technique for optical OFDM communication systems. J. Lightw. Technol. 32 (7), 1374–1382.
Qiu, K., Yi, X., Zhang, J., *et al*., 2011. OFDM-PON optical fiber access technologies. In: Asia Communications and Photonics Conference Optical Society of America, 1–9.
Reeve, M.H., Hunwicks, A.R., Zhao, W., *et al*., 1988. LED spectral slicing for single-mode local loop applications. Electron. Lett. 24 (7), 389–390.
Suzuki, H., Fujiwara, M., Suzuki, T., *et al*., 2007. Wavelength-tunable DWDM-SFP transceiver with a signal monitoring interface and its application to coexistence-type colorless WDM-PON. In: Proceedings European Conference on Optical Communication, PD3.4.
Suzuki, N., Nakagawa, J., 2005. First demonstration of full burst optical amplfied GEPON uplink with extended system budget of up to 128 ONU and 58 km reach. In: Proceedings European Conference Optical Communication, Sept., Paper Tu 1.3.3.
Townsend, P.D., Talli, G., MacHale, E.K., Antony, C., 2008. Long reach PONs. In: Technical Digest of Conference Optical Internet, pp. 1–2.
Valcarenghi, L., Yoshida, Y., Maruta, A., Castodi, P., Kitayama, K., 2014. Energy savings in TWDM(A) PONs: Challenges and opportunities. In: Proceedings of 15th IEEE Transparent Optical Networks (ICTON), p. We.A4.4.
Wang, H., Su, S., Gu, R., Ji, Y., 2015. A minimum wavelength tuning scheme for dynamic wavelength assignment in TWDM-PON. In: Proceedings of the 14th International Conference on Optical Communications and Networks (ICOCN).
Wei, J., Eiselt, N., Griesser, H., *et al*., 2015. First demonstration of real-time end-to-end 40 Gb/s PAM-4 system using 10-G transmitter for next generation access applications. In: European Conference on Optical Communication, 32–33.
Wong, E., 2012. Next-generation broadband access networks and technologies. IEEE/OSA J. Lightw. Technol. 30 (4), 597–608.
Wong, E., Lee, K.L., Anderson, T.B., 2007. Directly modulated self-seeding reflective semiconductor optical amplifiers as colorless transmitters in wavelength division multiplexed passive optical networks. J. Lightw. Technol. 25 (1), 67–74.
Wong, E., Zhao, X., Chang-Hasnain, C.J., Hofmann, W., Amann, M.C., 2006. Optically injection-locked 1.55 μm VCSEL as upstream transmitters in WDM-PONs. IEEE Photonics Technol. Lett. 18 (22), 2371–2373.
Xu, W., Fu, M., Le, Z., 2015. Energy efficiency scheme for delay aware TWDMPON. In: Proceedings of the 14th International Conference on Optical Communications and Networks (ICOCN), pp. 14–16.
Yang, H., Sun, W., Hu, W., Li, J., 2011. ONU migration in dynamic time and wavelength division multiplexed passive optical network (TWDM-PON). Opt. Express 21, 285–295.
Yi, X., Shieh, W., Ma, Y., 2008. Phase noise effects on high spectral efficiency coherent optical OFDM transmission. J. Lightw. Technol. 26 (10), 1309–1316.
Zhang, Z., Liu, Y., Guo, J., *et al*., 2015. 30-GHz directly modulation DFB laser with narrow linewidth. In:Asia Communications and Photonics Conference. Optical Society of America, AM1B. 3.

Relevant Websites

https://www.corecess.com/eng/solution/wdmpon.asp
Corecess.
https://www.lg-nortel.com/index.html
LG-Nortel.

Holographic Recording Media and Devices

Pierre-Alexandre Blanche, The University of Arizona, Tucson, AZ, United States

Holography Terminology

Holographic recording materials are able to change their refractive index and/or their absorption according to the intensity pattern created by two interfering laser beams. Then, when the material is re-illuminated by the appropriate light, the material can diffract the incoming beam and display the hologram. Thus, there are two functions that the material must perform: sensitivity to the recording light, and physical transformation of the optical properties.

The distinction between an image recording material and a holographic recording material is their necessary spatial resolution. For imaging, the material only needs to resolve details down to tens of microns, so the image is detailed and sharp to the human eye. For holography, we need to reproduce the modulation obtained by the interference between two laser beams. In this case, the separation between bright and dark regions can be sub-micron scale. This is the reason why digital holography had to wait for the development of high-pitch focal-plane-array sensors to become practical.

The goal of this entry is not to explain holography. However, since it heavily relies on the terminology associated with this method, it is necessary to give a brief introduction on the most important techniques.

When the material records the hologram as an absorption modulation, one talks of an amplitude hologram, since it is the amplitude term of the light wave that is affected when interacting with the media. When it is the index of refraction of the material that is modulated by the recording, one talks of a phase hologram. In this later case, there is no absorption of the incident light, and the efficiency of the hologram can be larger than for an amplitude hologram.

Holograms can also be recorded by modifying the surface of the material, such as by scribing or by lithography. In this case, the hologram is also a phase modulator, due to the difference in light path.

Another distinction with holograms, is they are either transmission or reflection holograms. With transmission holograms, the incident light goes through the medium and is diffracted on the other side of the material. The Bragg planes in this case are oriented more or less perpendicular to the surface of the material. For reflection holograms, the light is diffracted back toward the same side of the material as the incident beam. In this case, the Bragg planes are roughly parallel to the surface of the material.

There is an important difference between the spectral and angular properties of transmission and reflection holograms. Transmission holograms are angularly selective but spectrally dispersive, which means that when illuminated with a white light (broad band source) they diffract the incident light into a rainbow (the spectrum is dispersed). However, when transmission holograms are illuminated with a monochromatic source, they will only diffract at a very specific angle of incidence, the Bragg angle (angular selection).

The behavior of reflection holograms is quite the opposite: they are spectrally selective and angularly tolerant. Illuminated with a white light, they only diffract back a specific color, acting like a filter (spectral selection). However, reflection holograms can diffract the incoming light over a large angle of incidence.

Beware that a reflection coating can be applied to a transmission hologram to make it look like a reflection hologram: the light will be diffracted back. However, its spectral and angular characteristics will still be that of a transmission hologram.

There exist two regimes of diffraction: Raman-Nath and Bragg. The Raman-Nath regime is characterized by having a significant amount of energy not diffracted by the hologram (strong zero order), and diffracted into orders higher than 1: ± 2, ± 3, etc. This happens when there is only a short distance of interaction between the light and the hologram. Holograms operating in that mode are also called "thin", even though the actual thickness of the material is only one of several parameters defining the regime. On the other hand, when in the Bragg regime, holograms diffract most of the energy into the first order. For this to happen, the interaction length between the light and the grating should be relatively long, so these holograms are called "thick". The specific criteria between these two regimes is not just the physical thickness (d) but involves the light wavelength (λ_0), the material index of refraction (n), the grating spacing (Λ), and the angle of incidence (θ) in the Klein and Cook relationship:

$$Q' = \frac{2\pi\lambda_0 d}{n\Lambda^2\cos\theta} \tag{1}$$

where $Q' < 1$ for the Raman-Nath regime, and $Q' > 1$ for the Bragg regime.

Another criteria for the energy distribution in higher orders is the Moharam and Young equation involving the index modulation (Δn):

$$\rho = \frac{\lambda_0^2}{n\Delta n\Lambda^2\cos\theta} \tag{2}$$

where $\rho < 1$ for the Raman-Nath regime, and $\rho \geq 1$ for the Bragg regime.

We now need to introduce the different figures of merit characterizing a holographic material:

Sensitivity: sensitivity is the most important criteria defining the amount of energy density in J cm^{-2} units required to achieve a certain amount of efficiency. The sensitivity figure is generally not divided by the efficiency (J cm^{-2}) but given for a specific

efficiency (e.g., J cm^{-2} @ 90%). The sensitivity can also be given as the modulation (either index of refraction (Δn) or absorption ($\Delta\alpha$)) according to the energy density. The sensitivity generally depends of the wavelength used to record the hologram, which brings us to the next figure of merit the spectral sensitivity.

Spectral sensitivity: spectral sensitivity is the range of wavelengths to which the material is responsive, and holograms can be recorded. Ideally, it should be a curve of the sensitivity according to the wavelength, but most of the time there is only indication for the main laser lines: krypton red 647 nm, HeNe red 633 nm, argon green 514 nm, doubled YAG green 532 nm, and cadmium blue 441 nm. When a material has a very limited spectral sensitivity confined in only one color, it is called monochromatic. When the material is sensitive to red, green, and blue, even if only at very specific laser lines, it is called panchromatic.

Maximum efficiency (η_{max}): maximum efficiency (η_{max}) is the maximum diffraction efficiency expressed either as the ratio between the first order diffracted intensity divided by the incident intensity (external efficiency), or the ratio between the first order diffracted intensity and transmitted intensity without hologram (internal efficiency). Note that since the transmitted intensity is already reduced by the absorption, scattering, and Fresnel reflections from the material, the internal efficiency is always larger than the external efficiency.

Maximum modulation: maximum modulation is the maximum index (Δn) or absorption modulation (Δa) of the material. The maximum modulation is important to determine the diffraction efficiency. Recording geometry and wavelength are also important in that regard, but as a rule of thumb, longer wavelengths require larger modulation. Also the maximum modulation is important to consider when multiplexing several holograms into the same material. In this case, each of the multiplexed holograms takes a fraction of the maximum modulation range.

Spatial resolution: spatial resolution is the range of lateral frequencies the material is able to record and reproduce. When plan waves interfere, they produce intensity fringes spaced according to the grating equation:

$$\Lambda = m\frac{\lambda_0/n}{\sin(\theta_d) - \sin(-\theta_i)} \tag{3}$$

Where Λ is the fringes spacing, m the diffraction order, λ_0 the wavelength of the light, θ_d the diffraction angle, and θ_i the incidence angle, with both angles being defined by the recording geometry. This spacing ranges from a few hundred line pairs per millimeter (lp mm^{-1}) for low angle transmission hologram, to several thousand for reflection holograms at small wavelength (blue). An image hologram can be decomposed as the superposition of a multitude of these plane waves (Fourier decomposition). The material should be able to reproduce these small features in order to reproduce the hologram with maximum angular resolution.

Illumination Dynamic Range

The interference pattern is composed of an intensity modulated structure. So the material must be able to react not only to a certain average level of light (sensitivity), but reproduce an entire range of intensity. The dynamic range is given by the maximum and minimum intensity the material can react to. If the minimum intensity is not zero, the intensity ratio between object and reference beams can be detuned to accommodate the minimum.

Illumination Linearity Response

Within the dynamic range, it is expected that the material reacts linearly to the intensity. Thus, if the fringe pattern is sinusoidal, the modulation is also sinusoidal. If this is not the case, the phase or intensity modulation inside the material introduces higher spatial frequencies which translate into more energy directed into higher orders and a loss of intensity in the ± 1 order.

Stability:stability can be understood as the shelf life of the material before exposure, or the lifetime of the hologram once it has been processed, which are different and has to be specified. In any case. the stability of the material depends on environmental conditions such as temperature, humidity, and UV exposure.

Absorption spectrum: absorption spectrum is the absorption coefficient (cm^{-1}) as a function of wavelength. Unexposed material should have some absorption in order to interact with the recording light (sensitivity spectrum). However, it is important that once processed, the material be as transparent as possible for the waveband of interest. Absorption reduces the external efficiency.

Scattering: scattering in addition to the absorption some loss can be due to scattering. The dispersion is mostly due to inclusions in the material whose size is comparable to the wavelength of light, resulting some Mie scattering. Scattering can be due to micro-bubble formation such as in gelatin emulsion material, or polymer crystal formation such as in photopolymer. These problems will be detailed in the specific material sections.

Thickness change: When the material is processed after exposure, it can experience some thickness change, either shrinkage or swelling. This volume modification reshapes the Bragg's planes by tilting them and imposing a different grating spacing. The tilt changes the diffraction angle and the modification of the spacing affects the diffracted wavelength. Both changes can be computed according to the grating equation (Eq. (3)).

Pulse response: The energy delivered for recording the hologram can be such as a long exposure at relatively low intensity, or as a short pulse with high peak power. Laser pulses can be formed to last several hours (CW), or as short as picosecond, femtosecond, or even attosecond pulses. For a given energy, when the duration of the pulse decreases, the peak intensity increases proportionally. Depending on the mechanism involved to record the hologram in the material, there is a peak power, or a pulse

duration, for which this mechanism is less efficient or changes entirely. At that moment, there is a reciprocity failure in the sensitivity of the material.

Permanent Materials

One can claim that any material can be used to permanently record a hologram; it all depends of the recording energy. In this section we will focus our attention on the most sensitive and widely-used materials today.

Between the recording and display steps, permanent holographic materials need to be processed to enhance the phase or amplitude modulation that have been initiated by the exposure. This post-processing changes the nature of the material and fixes the hologram permanently. This means that new holograms cannot be recorded with further exposure of the material, and the hologram(s) already present cannot be modified.

Permanent holographic recording materials are the first to have been explored, and their development is closely related to the photographic recording medium. They are still developed today due to the large demand for long lasting holographic images, either for the security tag industry (bank notes and anti-counterfeit product), art work, holographic optical elements such as notch filters, dispersion gratings and null testers, or, to a lesser extent, holographic data storage.

Silver Halide

The chemistry of silver halide material is similar to the one used in analog black and white photography. Both of them use a colloidal suspension of silver halide crystals such as AgBr, AgCl, and AgI in gelatin, and require a post-exposure wet processing to make the image/hologram appear. This wet process presented in Fig. 1, uses a developing agent that enhances the latent image by transforming the entire silver halide crystal that has been exposed into an opaque metallic silver particle. The developer bath is followed by a stop bath to halt the reaction, then a fixer bath that removes the remaining unaltered silver halide crystals. At this point the picture/hologram is a gray scale (amplitude modulation) reproduction of the intensity distribution that illuminated the material).

There are two differences between the photographic and holographic material: the size of the crystals, which influences both the spatial resolution and the sensitivity; and the bleaching process, which is used in holography to transform the amplitude modulation into phase modulation and enhance the efficiency.

The size of the features to resolve in the case of a hologram (tens of nanometers) is orders of magnitude smaller than in a case of visual photography (tens of microns). Therefore, the grain size of the silver halide crystal used for hologram recording should be much smaller than for visible photography. In this regard, the holographic plates are similar to the ones used for X-ray medical applications. The sensitivity of silver halide is directly proportional to the size of the grain. So, the holographic emulsions are much less sensitive than the photographic one. For example, an emulsion able to resolve 5000 lp cm^{-1} has a grain size of 10 nm and a sensitivity of a few mJ cm^{-2}.

When the emulsion is processed to generate an amplitude modulation, its diffraction efficiency is limited to a maximum of 3.7% due to the resulting average absorption. It is possible to enhance the efficiency up to 100% by transforming this amplitude modulation into a phase modulation. This transformation is done by another wet post-processing called the bleaching process that is applied in addition to the traditional developer and fixer baths. In this bleaching process, the remaining silver particles are dissolved and leave voids with lower index of refraction than the surrounding gelatin (Fig. 2).

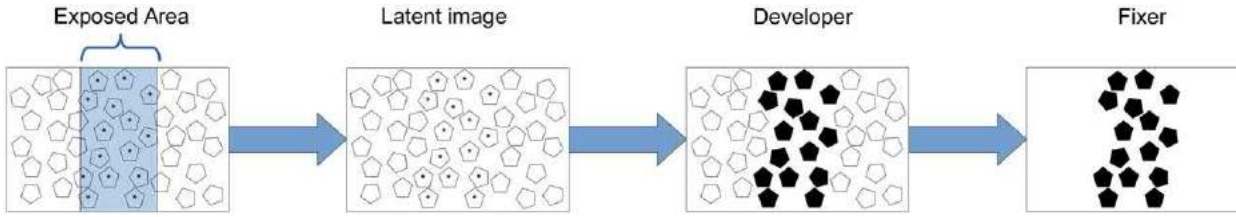


Fig. 1 Exposure and post processing of silver halide emulsion for amplitude modulation.

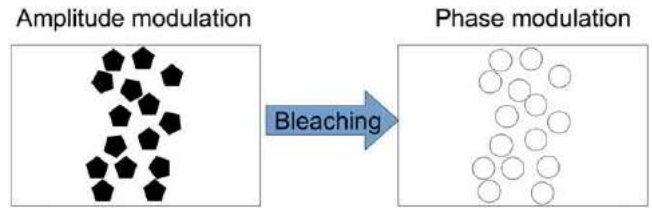


Fig. 2 Transformation of the amplitude modulation (absorption) into phase modulation (refractive index) by bleaching.

The advantage of silver halide emulsion is: a very large spectral sensitivity that covers the entire visible region (panchromatic) and even part of the infrared. This spectral sensitivity allows the recording of full color reflection holograms. Silver halide emulsions also have a very good sensitivity in the order or better than mJ cm^{-2}. This sensitivity is especially valuable when compared to other holographic recording materials that require much higher energy density. The spatial resolution of silver halide could be made such that it allows the recording of both transmission and reflection holograms. However, there is a trade-off between the spatial resolution and sensitivity. So, a careful selection of the emulsion according to the application is advised.

The disadvantages of silver halide material are the relative complexity of manufacturing and processing. Indeed, the chemistry, the application of the emulsion, and the careful handling in dark room, make silver halide a complex material to fully master. Other disadvantages are the limited shelf life, the material shrinkage during processing, and the possible scattering that occurs due to the bleaching process. This scattering is caused by the relatively large size of the voids left when the silver grains are removed, leaving some haze when the holographic image is rendered.

The shrinkage of the emulsion is particularly inconvenient for the production of reflection holograms. Any material thickness change between recording and display induces chromaticity change, and to a lesser extent, optical aberrations. It is possible to control, and even take advantage of the shrinkage by both pre and post processing, but this requires fine-tuning of the procedure.

For further reading about silver halide emulsion for holography see Bjelkhagen (1993).

Dichromated Gelatin

Dichromated gelatin (DCG) is quite similar to silver halide. The medium in both cases is gelatin, which is mainly composed of collagen, a natural protein, and both materials can produce very efficient phase holograms. However, in the case of DCG, the sensitizer is ammonium dichromate $(NH_4)_2Cr_2O_7$ (potassium dichromate is sometime used $K_2Cr_2O_7$), which is easier to obtain than fine silver halide crystals. The trade-off for this easier manufacturing is that both sensitivity and spectral response of the DCG are more limited than silver halide. DCG usually requires exposure in the order of several mJ cm^{-2} and is only sensitive to the blue and green region of the spectrum with a response of about 1–5 for these different wavelengths (1:488 v 5:532 nm). Sensitization of gelatin in the red has been demonstrated using other colorants, such as methyl blue, but this requires even higher energy exposure.

This requirement for large energy means that DCG needs high power lasers for the recording, or longer exposure time. Longer exposure time puts high constraints on the stability of the setup, and eventually a phase stabilization system could be required.

Once the material has been exposed, the dichromate is removed from the emulsion by a fixer bath and rinsed with water. The water also swells the gelatin, which creates larger voids where the gelatin was not exposed. The exposure has cross-linked the collagen molecular chains of the gelatin making it harder. The water is then removed by successive baths of increased concentration of isopropyl alcohol, to end up with a 100% concentration alcohol final bath. After processing, the gelatin film is fully transparent through all of the visible spectrum, as it should no longer contain any dichromate.

The index modulation of the DCG hologram can be finely tuned by:

- The aging of the material before or after exposure, old material is harder and has lower index modulation.
- The exposure energy, very low or high exposure leads to lower index modulation.
- The development process. In this final case, the temperature of the water bath is particularly critical. Higher temperatures produce a higher index modulation.

The modulation can be increased to up to 10% of the average index, which is one of the highest index modulation for any holographic recording material. Past that point, the size of the voids causes Mie scattering and the gelatin film becomes hazy.

In the case when the developing process did not produce the desired index modulation, the gelatin can be reprocessed starting at the water bath, reducing or increasing the modulation by changing the temperature of the water. This is particularly useful when manufacturing dispersing elements where the blaze and superblaze curves (dispersion spectrum) need to be finely tuned.

Yet another advantage of the DCG is its spatial resolution. For silver halide the resolution is determined by the size of the crystals. For DCG, the resolution is due to the photo-reduction of the chromium ions, so resolution up to 10,000 lp mm^{-1} can be achieved.

To avoid water re-absorption by the gelatin once it has been processed, it can be encapsulated between glass plates using optical glue. The gelatin media is particularly resistant to UV and thermal degradation.

For more information about DCG, see Stojanoff (2011).

Photopolymers

This group of materials are organic synthetic media that undergo a polymerization initiated by the illumination. In addition to this polymerization, the photopolymers also experience a change of refractive index, so the hologram is recorded as a phase modulation. This change of refractive index is not directly due to the polymerization, but rather to the diffusion of the remaining monomers to the polymerized regions. This diffusion by osmosis locally increases the density of the material in these regions and increases the index. When the modulation reaches the desired value the diffusion process can be halted by fully polymerizing the material, which can be achieved by UV illumination or thermal treatment (heating) depending of the material formulation.

The holographic recording process in photopolymer is presented in **Fig. 3** where the three steps: photo-polymerization, diffusion, and final polymerization are illustrated.

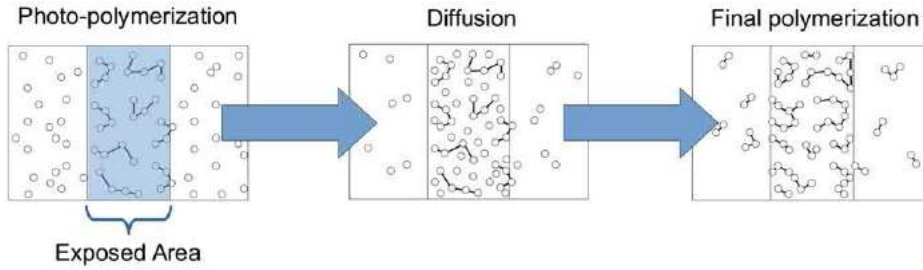


Fig. 3 Holographic recording process in photopolymer.

The sensitivity of photopolymers is on the order of 10 mJ cm^{-2} and they can be sensitized to respond to the entire visible spectrum. The spatial resolution can reach up to 4000 lp mm^{-1}, and the amount of shrinkage is minimal between recording and full polymerization, which is extremely useful for accurate color reproduction. Photopolymer also benefit from a prolonged shelf life and most formulations can be kept in the dark for years without any alteration. Very old samples (5+ years) have shown some crystallization though.

Commercially available photopolymers are the simplest holographic recording material to use. The post processing by UV illumination can be accomplished by direct sunlight exposure, and does not require a dedicated source.

The disadvantage of using commercial photopolymers is the inability to control the thickness of the material, which can be an important factor for dispersive elements. Also the backing medium that encapsulates the material is not under the control of the user, and in some cases has been reported to induce birefringence.

For more information about photopolymers, see Guo *et al.* (2012), as well as Berneth *et al.* (2011).

Photoresists and Embossed Holograms

The photoresists used for holography recording are the same materials as the one used by the semiconductor industry for lithography. Once exposed to light, the photoresist changes its ability to be dissolved by a solvent. One can distinguish between the positive resist where the parts illuminated become soluble, and the negative resist where the parts illuminated becomes insoluble. For holography this means that in both cases the modulation recorded is a surface relief.

The surface relief modulation can be used as a holographic pattern without any further modification. In this case, the hologram is used in transmission and the optical path length difference is given by the thickness modulation times the difference between the index modulation of the material and the air. The surface relief can also be coated with a reflective material such as metal. In this case the hologram is used in reflection and the path length difference is twice the thickness modulation.

The most common use for photoresist is to be a master copy for embossed hologram recording. The process is presented in **Fig. 4**, where the surface relief structure of the photoresist is copied by electroplating. During electroplating, a thick layer of metal is grown on the top of the structure. This metal is used as a stamp to reproduce the relief in a heated thermoplastic material by embossing. The thermoplastic material is then coated with metal and further encapsulated by a protective layer. This process is used for mass producing holograms such as security tags for credit card or as an anti-counterfeiting measure for luxury items.

Because most photoresists have been developed for the photo-lithographic process, they have been optimized for short wavelengths such as UV light where smaller details can be imaged. Photoresist spectral sensitivity decreases dramatically at longer wavelength, such as in the visible. The holographic structure can be recorded by the interference of two UV laser beams, but more frequently it will transfered to the photoresist by a lithographic process, either UV mask exposure, or by direct beam writing.

For more information about photoresist, see del Campo and Greiner (2007).

Photo-Thermo-Refractive Glasses

Photo-thermo-refractive Glasses (PTRG) are inorganic glasses such as $Na_2O - ZnO - Al_2O_3 - SiO_2$ doped with silver (Ag), cerium (Ce), and/or fluorine (F) atoms. These glasses are highly transparent in the visible and infrared (from 350 nm to 2700 nm), but UV exposure below 350 nm followed by thermal precipitation of crystalline phase produces a decrease of the refractive index. This phenomenon can be used to record phase volume holograms.

The maximum value of the refractive index modulation for PTRG is quite small (10^{-3}) compared to other holographic materials, but this is compensated by the very large material thickness and high transparency. Modulation times thickness ($\Delta n \cdot d$) in PTRG is enough to create very efficient volume hologram with diffraction efficiency exceeding 99%.

The photosensitivity of PTRG for 325 nm irradiation followed by 3 h of development at 520°C is about 1.5×10^{-3} cm^2 J^{-1} which means that a standard exposure for high efficiency hologram recording is in the order of hundreds of mJ cm^{-2}. The spatial resolution of PTRG materials is particularly large ranging from 0 (continuous) up to 10,000 lp mm^{-1}.

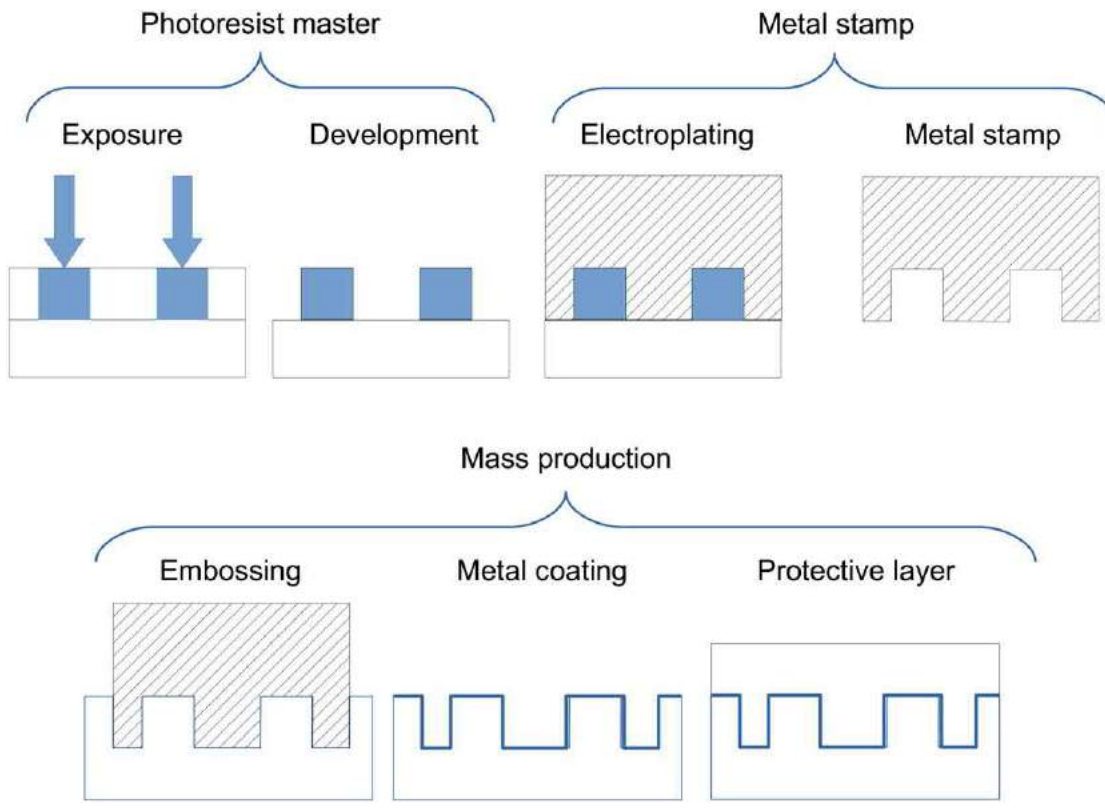


Fig. 4 Production of embossed hologram from a photopolymer surface relief structure.

PTRG materials are particularly important because they are useful for manufacturing the fiber Bragg gratings that can be found in telecommunication fibers and fiber lasers. The fiber Bragg grating is used as a notch filter to select a particular bandwidth and reject all the others. Based on this principle, optical add-drop multiplexers can be built.

Another major advantage of the PTRG materials is their very high damage threshold. Being particularly transparent, they can withstand nanosecond pulses with an energy exceeding tens of mJ cm^{-1} (measured at 1064 nm, and 10 Hz repetition rate), and an average power exceeding 100 kW cm^{-2} of CW irradiation (measured at 1085 nm). This allows for creation of holographic beam combiners for high energy lasers.

For more information about photo-thermo-refractive glasses, see Glebov (2002).

Holographic Sensors

The spectral characteristic of the light diffracted by an hologram is extremely sensitive to the index modulation (Δn), and the spacing between the Bragg planes (Λ). This can be seen from the Bragg's Eq. (3) that can be further differentiated to find the spectral dispersion:

$$\frac{\Delta\lambda}{\lambda} = \frac{\Delta n}{n} + \frac{\Delta\Lambda}{\Lambda} + \cot\theta\Delta\theta \tag{4}$$

Therefore, if a recorded hologram is altered by its environment by swelling, or shrinking, the characteristics of the diffracted spectrum will change. This spectral, or color, change can be used as a sensor to detect the component responsible for the thickness change.

Gelatin emulsion, for example, is very well known to absorb the air humidity and swell. Swelling separate the Bragg planes already recorded and makes the hologram color shifts toward the red part of the spectrum, and eventually disappear in the IR. It is possible to then restore the original color of the hologram by dehydrating the gelatin by heating it up.

Based on this example of the gelatin, several other polymer matrices have been used to make the hologram sensitive to a wide variety of components. On can cite, poly(2-hydroxyethyl methacrylate) (pHEMA), poly(acrylamide) (paam), or poly(vinyl alcohol) (PVA), and poly(dimethylsiloxane) (PDMS). These matrices are porous to a selection a solvents, and their absorption will make them swell.

It is also possible to dope the porous polymer matrix with other molecules such as crown ethers that will bind with specific metal ions present in the solution and make the hologram sensitive to that specific ion.

Another example is the addition of 3-(acrylamido)phenylboronic acid (3-APB) into pAAm matrix. The 3-APB molecule can bound with glucose which makes the hologram a potential sensor for monitoring bodily fluid sugar.

This technique can be generalized to any chemical reagent sensitive to specific molecules. Once incorporated into the porous matrix that has been cross linked to form an hologram (see the case for the gelatin in Section Dichromated Gelatin), the spectral dispersion is affected by the presence of that particular molecules, and the hologram can be used as a sensor.

For more information about holographic sensors, see Yetisen *et al.* (2014).

Refreshable Materials

In this section we will introduce the holographic recording materials where the hologram can be erased, and the material reused to record a new diffraction pattern. This class of material is not used to permanently display the hologram, but rather, to present it momentarily, analyze the diffraction, change the recording setup, and write a new interference figure.

These refreshable materials can further be distinguished between those which are dynamic and do not need post processing for the hologram to appear, and those which need post processing. The advantage of the dynamic materials is that they do not need to be moved away from their original position in the recording setup before being read. This allows for a greater stability and reproducibility in the experiment, especially in holographic interferometry. The advantage of the non-dynamic material is that the hologram is more stable and will last longer even during the exposure by the reading beam.

Photochromic Materials

The word photochromism describes a material that changes its absorption coefficient due to illumination. This change of absorption can be used to record amplitude holograms. However, the diffraction efficiency of phase holograms is much larger, and the Kramer–Kronig relationship specifies that a change of absorption always goes along with a refractive index change, eventually at a shifted wavelength. So, it is better to use the photochromic material for its index modulation than its absorption modulation. To do so, the writing wavelength is tuned near the absorption peak, but the reading of the hologram is performed in a more transparent region of the spectrum, where the index modulation is enhanced.

The photochromic effect is a macroscopic, observable modification of the material property, and can be generated by many different microscopic alterations. It can be permanent: such as photobleaching; or reversible: like in photo-isomerization. Photochromism can happen in inorganic material, such as doped glass, or inorganic molecules, such as azoic dyes.

In photochromic glasses, the SiO_2 matrix is doped with a metallic compound such as silver halide crystals. Under light excitation, the transparent silver halide molecules undergo a decomposition into metallic silver particles, which are opaque. Because the halogen molecules are trapped inside the glass and cannot escape, they can recombine with the silver after the illumination is gone, and the glass retrieves its initial absorption profile.

Photochromic glasses are not often used for holography due to their slow response time and low sensitivity.

On the other hand, there exist several types of organic molecules that react to illumination by reversibly changing their absorption spectrum and refractive index. The most commonly used in optics and more specifically for holography recording are spiropyrans, diarylethenes, and azobenzenes. When a photon is absorbed by one of these molecules, it undergoes a conformation change such as isomerisation.

Some molecules such as bacteriorhodopsin have several metastable excitation states that can be addressed with different wavelengths. By taking advantage of this particularity, the hologram can be recorded at one wavelength, read non-destructively with another, and erased by a third. The selection of the different wavelengths is such that the recording is done in a strong absorption region of the molecule, the reading away from it in a transparent part of the spectrum, and the erasing excites the molecule into an unstable conformation that decays back into the stable form. This process is illustrated in **Fig. 5**.

For more information about photochromic glasses, see Glebov (2002). For more information about photochromic organic molecules, see Dürr and Bouas-Laurent (2003).

Persistent Spectral Hole Burning

When cooled down at cryogenic temperature (liquid helium ≈ 4 K), photochromic materials can experience an inhomogeneous broadening of their absorption spectra. This means that each of the absorption centers (molecules or ions) acquire a narrow spectrum that is shifted in frequency due to the interaction with the host matrix (either glass or polymer). The material spectrum still spans a large bandwidth, but individual centers possess a narrow line as shown in **Fig. 6**.

The observation of the inhomogeneous broadening requires very low temperature. This is because at room temperature the spectrum broadening of individual centers is due to the interaction with phonons and other excitation. At cryogenic temperature, these interactions are minimized, revealing the inhomogeneous broadening of the material.

If the host matrix was perfect at cryogenic temperature each absorption center would be in the exact same state, and the spectrum would be a single narrow line. In these conditions, no inhomogeneous broadening happens and no spectral hole burning would be possible. There will be only one spectral line.

The hole burning itself happens when the inhomogeneous broadened photochromic material is illuminated with a narrow bandwidth source such as a laser. The source will only excite the centers with a spectrum overlapping the light frequency. Once

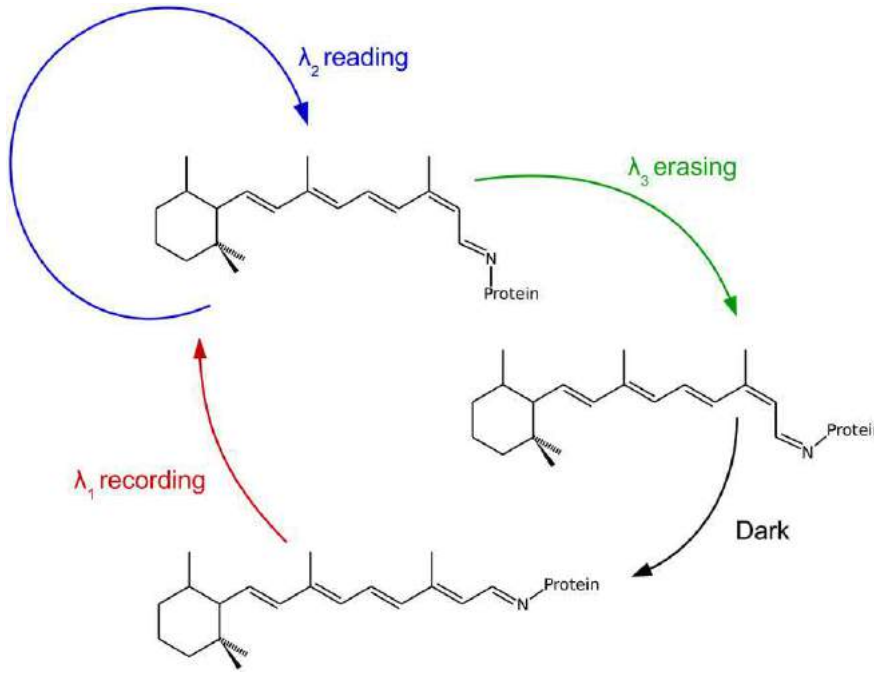


Fig. 5 Molecular structures of the various isomeric forms of bacteriorhodpsin, and the different wavelengths used for writing, reading and erasing an hologram.

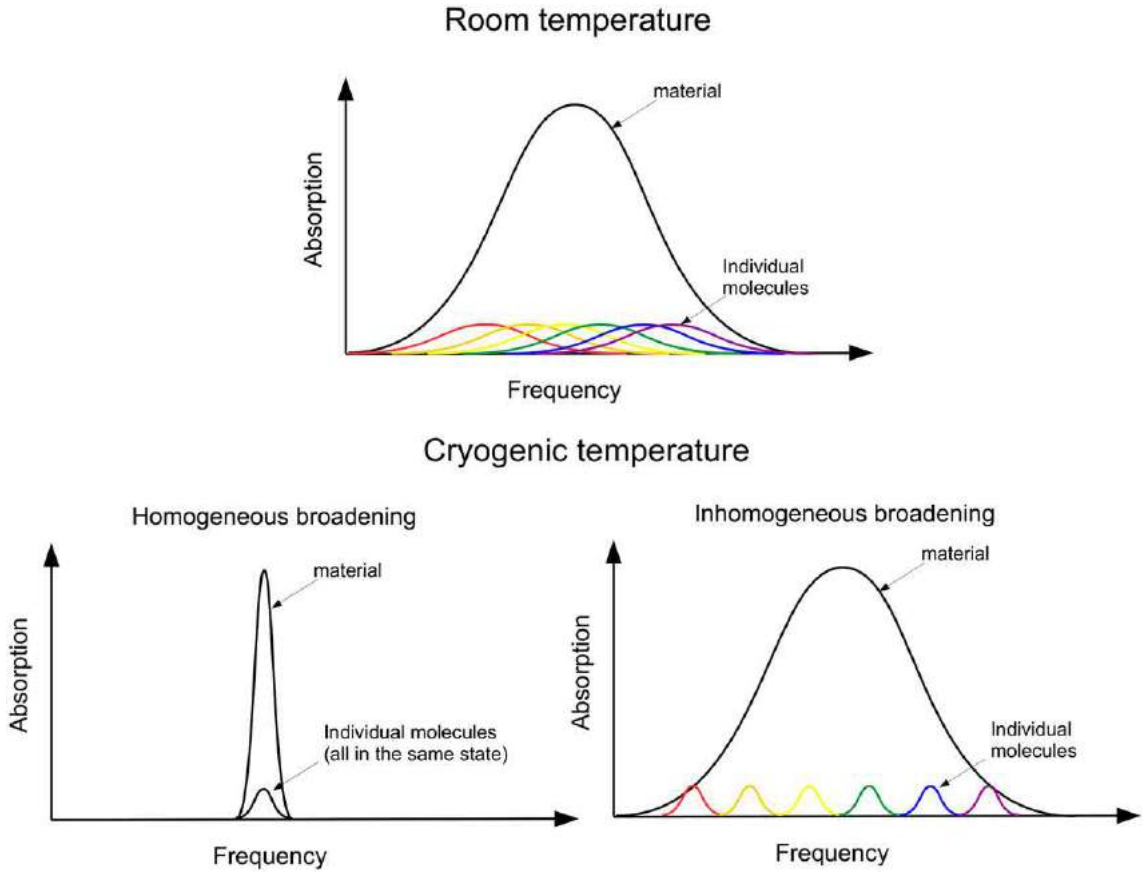


Fig. 6 Spectral broadenings. Top: at room temperature, the broadening is due to phonon interaction and the spectrum of individual center is large. Bottom left: at cryogenic temperature, in a perfect lattice, all the centers are in the same state and the spectrum of the material would be narrow. Bottom right: the spectrum of individual center is narrow but due to the interaction with the lattice each spectrum is shifted in frequency.

photo-excited, the molecule or ion composing the center will have a different absorption spectrum and leave a notch in the original absorption band as presented in Fig. 7.

An important parameter for spectral hole burning material is the ratio between the width of the notch that can be formed, Γ_{ZPL} (or zero phonon line), and the width of the material spectrum due to the inhomogeneous broadening, Γ_{inh}. This ratio indicates the number of individual frequencies that can addressed in the material to record information, and can be as large as 10^6.

Spectral hole burning materials can either be permanent or not. It is reversible if the photo-excited state of the absorption centers can decay back to the original state. The temporal behavior is strongly dependent on the material, the temperature, as well as the mechanisms of relaxation that bring back the excited state to its original state. Hours of dark decay time have been demonstrated.

The interest of spectral hole burning for holography come from the possibility to use massive wavelength multiplexing. Each individual spectral notch that can be created is an opportunity to write a different and independent hologram. The use of the frequency domain to record the hologram is a new dimension for encoding the information in addition of time and space, which increases dramatically the quantity of information that can be stored in the material. Spectral hole burning give the opportunity to increase the capacity of holographic data storage, and can also be used to design very high density correlation filters.

For more information about persistent spectral hole burning and its application for holography, see Moerner (1988).

Polarization Sensitive Materials

Some of the photochromic materials are sensitive to the light polarization and can be used to record not only amplitude-modulation holograms but also polarization-modulation holograms. This mechanism is also refer as orientational hole burning in comparison to spectral hole burning.

Polarization holograms are formed when two coherent beams with orthogonal polarization interfere. In that case, the intensity is constant but the polarization vector changes along the period of the grating. For left and right circular polarized recording beams, the modulation is a linear polarized vector whose direction is rotating around the bisector of the beams propagation. For linear s- and p-polarized recording beams, the modulation is elliptical polarizations changing direction. These cases are presented in Fig. 8.

Sensitivity of photochromic molecules to polarization can be explained by the preferential absorption when the electric field is aligned to the axis of delocalized electronic orbital. This is the case of the azobenzene dyes that become oriented perpendicular to the light electric vector after multiple trans-cis photoisomerization and cis-trans relaxation cycles. During each cycle, the molecule

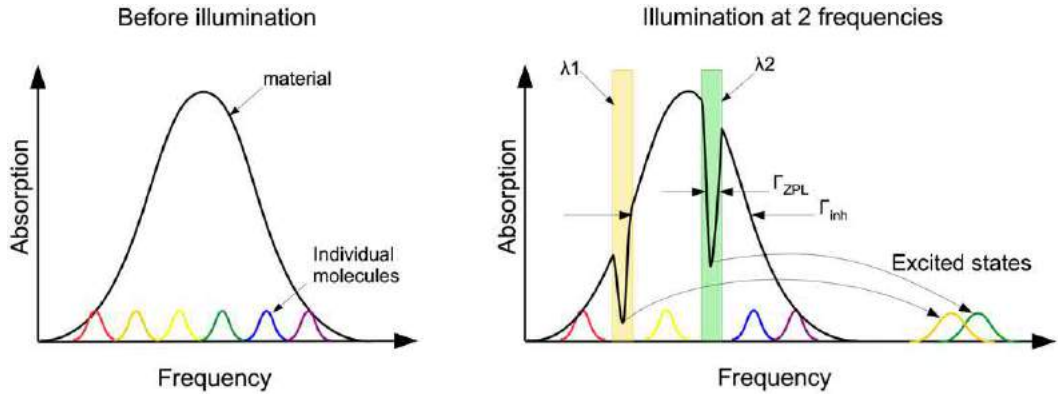


Fig. 7 Spectral hole burning principle: a narrow band source only excites the centers that have a spectrum overlapping the source frequency. The center excited states have a different spectrum which leave a notch in the material initial spectrum at the same frequency as the source.

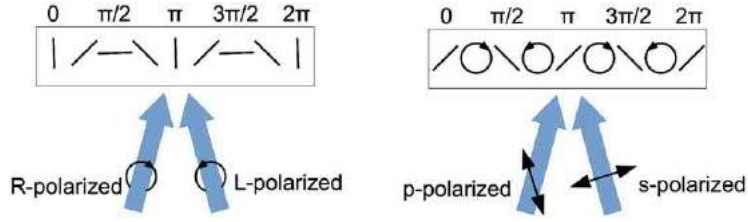


Fig. 8 Polarization gratings formed by the interference of, left: left and right circular polarized coherent beam. Right: s- and p- polarized coherent beams.

can rotate by a finite and random amount. Over a large number of excitation-relaxation cycles, the molecule distribution becomes predominantly organized with the main axes orthogonal to the polarization. The orientation is a statistical process and is not driven by any torque, which means that it is rather inefficient, explaining the slow dynamic time of the material in comparison to the very fast isomerization process.

In some materials, the multiple transitions also induce a molecular migration that can leave a surface relief modulation in addition to the volume phase hologram.

Even though the trans- and cis- forms of the azobenzene have different absorption spectra, the photo-orientation process leaves the molecules in their relaxed form (trans). The hologram formed in this case is a phase modulation due to the birefringence induced by the anisotropic molecular distribution.

Sensitivity of polarization sensitive azobenzene is on the order of mJ cm^{-2}, and the index modulation achievable is rather large, on the order of 10^{-2}.

For more information on polarization holography and polarization sensitive materials, see Nikolova and Ramanujam (2009).

Photorefractive Materials

The literal meaning of a photorefractive material is one that changes its refractive index under illumination. However, the scientific community has come to define the photorefractive process very specifically as the reversible and dynamic change of the index due to an electronic process. So, photorefractive materials should not be confused with photochromic even though the observable macroscopic effect is similar.

The photorefractive effect is a multiple step process that starts with the absorption of photons and the generation of electric charges inside the material. This aspect is similar to the photovoltaic process. Both electron and hole charges are created since there is conservation of the material electrical neutrality. However, in photorefractive materials, the mobility of the charge carrier depends on its sign. One type of charge carriers migrate inside the material while the other type stays in the localization where it was created. There are several charge transport mechanisms such as diffusion, drift (under external electric field), and photovoltaic that explain this migration. The mobile charge carriers are eventually trapped in the dark regions of the material, and the local charge distribution creates a space-charge electric field between the illuminated and dark regions of the material. This space-charge field modulates the index of refraction by either nonlinear electro-optic effect (in inorganic crystals) or molecular orientation (in organic materials).

The different steps leading to the photorefractive process is presented in **Fig. 9**.

The photorefractive effect is characterized by a phase shift (φ) between the intensity pattern and the index modulation. This phase shift is responsible for the self-diffraction of the recording beams in a two beam coupling experiment.

Photorefractive materials can be organized into two different categories: inorganic crystals and organic compounds. Inorganic crystals are grown at high temperature and their composition is imposed by their stoichiometry and crystalline structure. Some inorganic dopants, such as metallic atoms, can be incorporated to modify the electro-optical properties like absorption band or carrier mobility. Examples of inorganic crystals include the silenites ($Bi_{12}SiO_{20}$, $Bi_{12}GeO_{20}$), strontium-barium niobate ($Sr_xBa_{(1-x)}Nb_2O_6$), barium titanate ($BaTiO_3$), lithium niobate ($LiNbO_3$), as well as semiconductors such as GaAs, InP, and CdTe.

Photorefractive organic compounds are mixtures of several organic molecules exhibiting specific function to achieve the photorefractive effect. They are mainly composed of a photoconductive polymer matrix that allows the charge transport. Such matrices are PVK: poly (N-vinylcarbazole), or PATPD: poly (acrylic tetraphenyldiaminobiphenol). To enhance the sensitivity in the

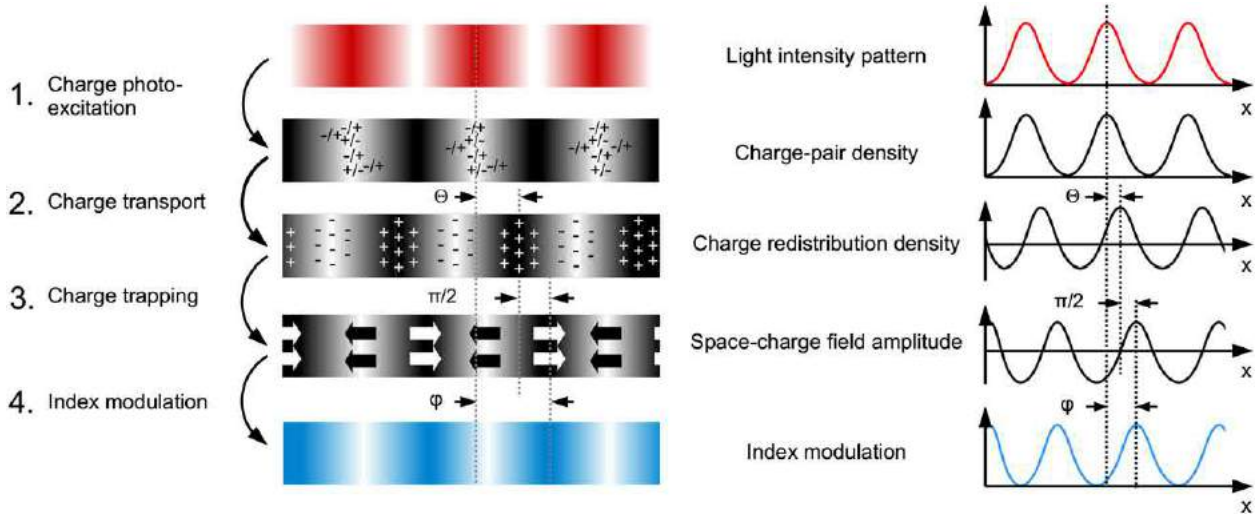


Fig. 9 Photorefractive process starting with the intensity distribution of the interference pattern, charge photo-generation, charge transport and distribution after migration and trapping, space-charge field, and final index modulation.

visible the matrix is doped with a sensitizer that is responsible for the charge photogeneration. The most used sensitizer that covers a large band in the visible region is a derivative of the C_{60} fullerene molecule, PCBM: phenyl-C61-butiric acid methyl ester. The index modulation is provided by chromophores, rod-like molecules with large dipolar moment and polarizability, such as DMNPAA: 2,5-dimethyl-4-nitrophenylazoanisole, or 7-DCST: 4-azacycloheptylbenzylidene-malonitrile. To increase the mobility of the chromophores that need to orient in the space-charge field, the glass transition of the polymer matrix is lowered with plasticizers, low molecular weight molecules such as ECZ: N-ethylcarbazole, or BBP: n-butyl benzyl phtalate.

The optical properties of the photorefractive crystals and organics are strongly dependent on their nature and cannot be generalized. Their sensitivity ranges from μJ cm^{-2} to mJ cm^{-2}, and their spectral response varies through all the visible up to the infrared.

Due to their very specific properties, photorefractive materials are used in numerous applications, such as phase conjugation, coherent beam amplification, imaging through turbid or scattering media, dynamic holographic imaging, and holographic interferometry.

For more information on photorefractive inorganic crystals, see Günter and Huignard (2007). For more information on photorefractive organic materials, see Blanche (2016).

Photo Thermoplastic Process

The term Photo Thermoplastic does not describe a material, but a material process. Thermoplastic polymers have their structural rigidity altered by temperature. By heating a sheet of material it becomes supple and can be deformed. This property can be exploited to record a surface relief hologram which is the encoded in the deformation of the surface of the sheet. To do so, the thermoplastic material is coated on the top of a photoconductor. The surface of the photoconductor is charged by corona discharge which attracts opposite charge on the top of the thermoplastic. When illuminated, the charges that were on the top of the photoconductor migrate through the material and came in contact with the thermoplastic. After illumination, the photoconductor is charged a second time. Because of this second charging, there is now more charge in the region that was illuminated than where the material was not exposed to light. At that moment, the thermoplastic material is heated up and the electrostatic attraction between the charges squeeze the film, modulating its thickness. To erase the surface relief, the material is evenly charged and heated. These processes are illustrated in **Fig. 10**.

For more information about the photo thermoplastic process, see Lin and Beauchamp (1970).

Electronic Devices

Twenty years ago, there was a technological transition from analog to digital photography. This transition was possible thanks to the development of the focal plan array detector and the personal printer. Today, we are assisting to the same kind of transformation for holography. More and more often, the analog recording media is substituted for an electronic device. This transformation is made possible by the high resolution of both the detectors and the micro display devices. Once again, when photography requires micrometer resolution, holography, which records the wavefront features, requires nanometer precision. Needless to say, this is much more difficult to achieve.

The advantages of the electronic devices are numerous. They are dynamic and do not require post processing to capture the hologram, they can be used over a very large wavelength bandwidth, they allow computer manipulation of the holographic pattern for direct interpretation or eventual transformation, they are easy to use, and they are reusable with no pre-processing.

The figure of merit for the recording and reproduction of holograms by electronic devices is the space-bandwidth product (SBP), which is a measure of the rendering capacity of an optical system. It is defined as the product of the spatial frequency bandwidth and the spatial extent of the image. In other words, at constant SBP you can either capture/reproduce a small image at high resolution, or a large image at low resolution. The transformation from one to another can simply be performed by a magnifying lens.

Although very useful and convenient, electronic devices are still suffering from a coarse resolution and pixel count compared to the recoding materials (small SBP), but this might be solved in the very near future. A more fundamental problem is their inability to record and reproduce volume holograms. Detectors only record the interference along a plane, and the current micro-

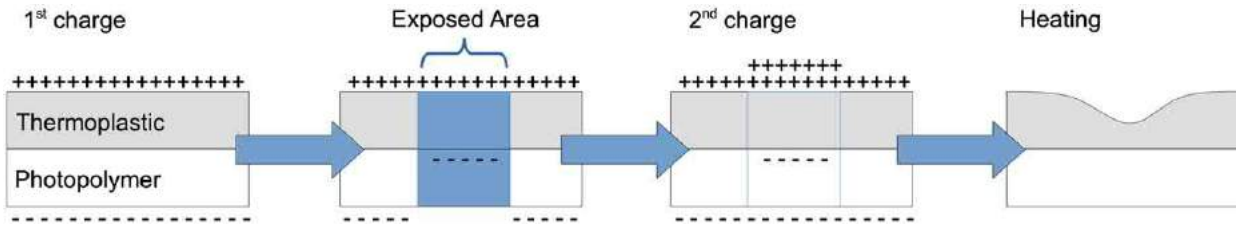


Fig. 10 The photo thermoplastic process: charging the material, exposure drives the charges inside the photopolymer, the second charging increases the electrostatic attraction where the material has been exposed, heating the thermoplastic modulates the surface relief.

displays only produce surface relief modulation. For a volume hologram to be recorded, the phase needs to be sensed over a non-negligible thickness, and reproduced likewise. This excludes both the sensing and reproduction of reflection hologram by electronic devices.

Even with these limitations, electronic devices are used in a large variety of holographic setups, from 3D display to microscopy and adaptive optics. They are promising the same bright future for holography as the one they created for photography.

Focal Plane Array Detector

The function of detecting the intensity modulation generated by the interference of the coherent beams is provided by the focal plane array detector (FPAD). Two technologies can be used: either charge-coupled devices (CCDs), or complementary metal-oxide-semiconductors (CMOSs).

These devices have been popularized by digital photography and are very well known, even outside the scientific community. It is only recently, however, that the pixel pitch and count made them interesting for holography. A commercial grade FPAD can now have 20 million pixels with a 4-micron square size. This size of pixel correspond to a resolution of 125 lp mm^{-1} which is far from the several thousand that analog materials offer, but is enough for detecting interference fringes for in-line holography, or low angular separation beams at long wavelength.

Scientific grade FPAD can have a pixel size down to 2.2 mm for a total of approximately 120 million pixels on a single sensor. The problem with decreasing the pixel size is the same as with the silver halide emulsion: when the size shrink, there are fewer and fewer photons interacting with the pixel (crystal), and the sensitivity decreases dramatically. This is particularly true for the CMOS technology where the pixel surface is shared between the sensing area and the electronic amplification and logic.

Once the interference pattern has been detected by the FPGA, it can be used to reconstruct the 3D light field (phase and amplitude) using a computer, or displayed by another electronic device to reproduce the hologram.

For more information about the use of focal plan array detector for holography, see Schnars and Jüptner (2002).

Acousto-Optic Modulator

A sound wave is a compression wave propagating inside a material. This compression (and dilatation) modulates the index of refraction and if the frequency is carefully selected, can be used to diffract the light. Acousto-optic modulators (AOM), use this principle to generate diffraction gratings. They are composed of a piezo-electric transducer (PZT) that generates the wave, which is coupled to a transparent medium, generally glass, quartz, or other crystal. Two configurations are presented in **Fig. 11**. The first one is the commonly used Bragg cell, where the sound wave generated by the transducer travels inside the material. The second is one where the transducer is put on top of a waveguide and generates a surface acoustic wave. The wave diffracts the light out of the waveguide so this configuration is called leaky-mode coupling.

There are two modes of operation for the Bragg cell: static and traveling wave. When the acoustic wave is traveling, it shifts the diffracted beam frequency due to the Doppler effect. This shift is about 1 GHz due to the speed of sound in the material. To avoid this shift, a sound reflector can be used at the other end of the material. This reflection generates a standing wave like in musical wind instruments.

By changing the frequency and the amplitude of the vibration at the PZT, it is possible to change the diffraction angle and amplitude. AOM can achieve a diffraction efficiency up to 99%.

The sound wave can only generate diffraction gratings that redirect the incident beam but do not change its wave front. Used simply as is, AOMs would not be able to form a 3D image. However, if instead of using a fixed frequency, the AOM is driven with a sum of many frequencies, it becomes similar to a one-dimensional arbitrary holographic pattern: any waveform can be decomposed as a sum of sinusoids.

The advantage of the AOMs is their very large space-bandwidth product compared to the spatial light modulator. An AOM can produce larger images size, and larger view angle than the other technologies.

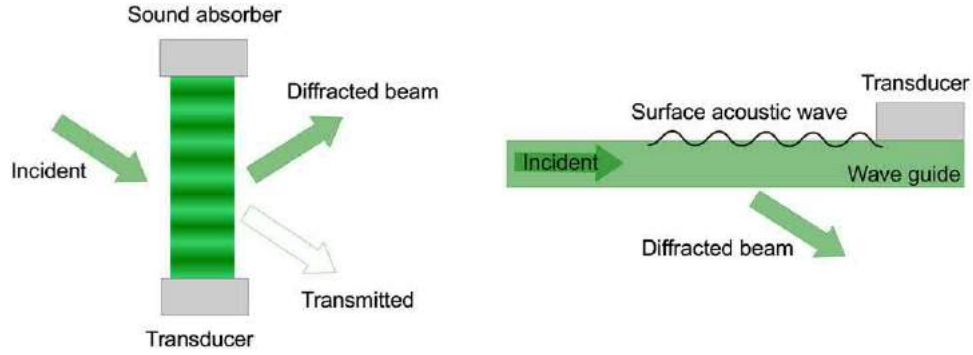


Fig. 11 OAM configurations. Left: Bragg cell where the sound wave generated by the transducer modulates the index inside the material. Right: leaky-mode mode coupling where the transducer produces a surface acoustic wave on top of a waveguide.

Due to the very fast reaction time, AOM are driven by a signal ranging from tens to hundreds of MHz. AOMs are used for Q-switching pulsed lasers, for pulse shaping, in telecommunication to modulate the optical signal so it carries the information, and in spectroscopy for frequency control.

For more information about acousto-optic modulator, see Efron (1994).

Spatial Light Modulator

Spatial Light Modulators (SLM) are dynamic pixelated electronic devices where each pixel can be individually controlled to change the amplitude or the phase of an incoming light beam. Initially developed for the display industry, SLMs are also called micro displays because of the small size of the pixels. Thanks to that small size, they can also be used to display holographic patterns, which diffract the incident visible light over an appreciable angle. Larger pixels reduce the angle of diffraction according to the Bragg equation and need to be used at larger wavelengths.

Micro displays where the light is directly emitted from the pixels such as in light emitting diodes (LEDs) and thin film transistors (TFTs), cannot be used for holography since the phase of the different sources cannot be controlled individually. A micro display system where the phase of the light emitter could be controlled would be the equivalent of a phase array radar for the visible and is currently the object of active research.

Because SLMs are dynamic and refreshable at will, they are extremely attractive for dynamic holography applications such as optical tweezers, optical switching, non-mechanical scanner, wave front correction, and holographic interferometry.

Of course, 3D display is also a potential application, but the space-bandwidth product of commercially available SLMs (number of pixels over the pitch) is still too small to generate a comfortable 3D image, even with a 4 K UHD resolution: 4096×2160 pixels. Systems tiling several SLMs together to increase the SBP have been demonstrated though.

Two types of SLMs can be distinguished according to their mode of operation: liquid crystal on silicon (LCoS) and micro-opto-electro-mechanical systems (MOEMS).

Liquid crystal on silicon

Liquid crystals (LCs) are liquid form materials composed of birefringent molecules that self-align. Due to their large dipole moment, the LC molecules are also sensitive to externally applied electric field. When an electric field is applied the molecules rotate, which changes the optical properties of the material: birefringence amplitude and axis, as well as the refractive index.

In LCoS SLMs, a layer of LC is deposited on top of a complementary metal-oxide-semiconductor (CMOS) structure forming individual cells where the voltage can be applied independently. This is the same technology used in liquid crystal display (LCD) and LC television. When a voltage is applied, the LCs induce phase retardation in the polarized incident light as large as a few wavelengths, which is enough to produce a fully modulated (2π) holographic phase pattern. This phase modulation can take multiple values (usually 8 bits $=$ 256 levels), which allow the LCoS to reproduce continuous phase holograms.

It has to be noted that at maximum spatial frequency, the pattern is only represented by 2 pixels per period, and only binary phase can be displayed. To reproduce multilevel functions, several pixels need to be used and the frequency is reduced.

To increase the reflectivity of the device, a layer of aluminum or dielectric mirror is coated between the LC and the CMOS electronic. The mirror can be ordered according to the wavelength at which the device is supposed to be used (IR, visible, or specific laser line) which give a reflectivity larger than 90%.

Amplitude modulation LCoS displays use a polarizer in front of the SLM so the phase retardation is converted into amplitude modulation following Malus's law: $I = I_0 \cos^2 \rho$, where ρ is the angle between the polarization vector and polarizer direction. This mode of operation is not interesting for holography since amplitude holograms only reach 10% in the first order.

A phase only LCoS is more interesting for holography due to the higher diffraction efficiency (up to 100%). In this case, the device modulates the index of refraction directly, according to the main axes of the birefringence.

Typical LCoS pixel pitch is a few microns, which offers a maximum diffraction angle of a few degrees in the visible (4 mm pitch $\approx 4^\circ$ in the visible). The pitch is limited by the field bleeding from one cell to another and it would be hard to reduce it further in the future. An LCoS benefits from a fill factor (active pixel area over pixel size) as large as 90%. The refresh rate is limited by the viscoelastic relaxation of the molecules but can increase up to a few hundred Hz. Unfortunately, this refresh rate does not allow for the time multiplexing or very fast reconfiguration sought in some holographic applications.

For more information about LCoS SLM and holographic application, see Osten and Reingand (2012).

Micro-opto-electro-mechanical systems

Micro-electro-mechanical systems (MEMSs) are devices where a micron scale mechanical feature can be activated by an applied voltage. When this device is also used to interact with light, it is named MOEMS with the introduction of the term "opto".

The best known of these MOEMSs, due to its commercial success, is the Texas Instruments DMD (digital micromirror device) which is also known as the DLP (digital light processor). The DMD is used in display applications such as televisions and projectors. It is composed of an array of micron size mirrors: 13 μm for the 0.7″ XGA, down to 5.4 μm for the 1080p "pico". The size of the DMD follows the resolution standard for display: XGA $= 1024 \times 768$, 1080p $= 1920 \times 1080$, WQXGA $= 2560 \times 1600$.

The DMD is a binary device and the mirrors can only take two orientations: tilted left or right by 12° for the large DMD, or 17° for the pico, according to the surface normal. Accordingly, the DMD can be used to display binary amplitude holograms. The incident light is reflected by the mirrors, and since the pattern is composed of left and right tilted mirrors there are two reflection

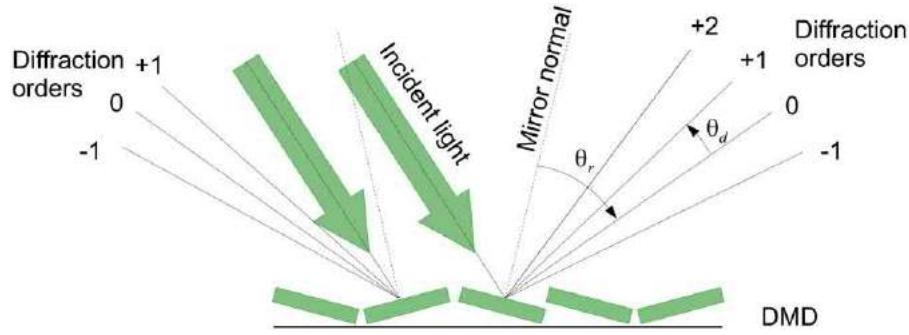


Fig. 12 Geometry of diffraction from the DMD.

directions θ_r (or zero orders). In each of these directions, there are multiple diffraction orders ($\pm 1, \pm 2, \ldots$) due to the diffraction by the structure created by the mirror orientation (see **Fig. 12**). The maximum angle of diffraction (θ_d) around the 0 orders is given by the pixel pitch and is about 2° for 13 μm.

The maximum diffraction efficiency for the binary amplitude hologram displayed by the DMD is 10.1% as predicted by the Fourier decomposition theory. However, the big advantage of the DMD is its refreshing rate, which can be up to 20 kHz (100 times faster than LCoS). This enables applications that are not accessible with LCoS SLMs.

That fast refresh rate is what allows the DMD to display numerous intensity levels (10 bit gray scale), although it is a binary device that can only display black (light reflected away from viewer) or white (light reflected toward viewer). By oscillating the mirror at different frequencies, the intensity directed to the viewer is modulated. This only works because the eye integration time is rather slow ($\approx$24 Hz) compared to the 20 kHz of the device. DMDs cannot be used for applications requiring smaller integration time.

Another diffractive MOEMS worth mentioning is the grating light valve (GLV) which is composed of parallel thin ribbons (few micron pitch) that can take two positions: up or down. The phase difference generated by these two positions can be used to diffract the incident light with 40% efficiency (binary phase hologram). The GLV has been shown to have a refresh rate of up to 50 GHz (2000 times faster than the DMD), but also has some severe limitations for holography: it is a 1-dimensional array and only 1 or 2 ribbons can be activated at a time.

Other types of MOEMS are under development specifically for holography. One such device is an analog piston MOEMS where the micro-mirrors are moved up or down according to the voltage applied. Like in the GLV, this movement creates a phase difference between the mirrors which diffracts the light. The difference from the GLV is that the mirrors can be positioned at multiple levels instead of just two. This allows users to display a continuous phase hologram which can have a diffraction efficiency as high as 100%. Such a piston MOEMS will combine both the efficiency of the LCoS, and the speed of the DMD. If a piston MOEMS can be manufactured with a very large number of mirrors (hundreds of trillions), it could enable true holographic television along with other diffractive applications.

For more information about MOEMS SLM and holographic application, see Blanche *et al.* (2015) as well as Blanche *et al.* (2017).

Acknowledgements

The author would like to thank Mr. Colton Bigler for the careful revision of the manuscript.

See also: Overview: Holography

References

Berneth, H., Bruder, F.-K., Fäcke, T., *et al.*, 2011. Holographic recording aspects of high-resolution bayfol® hx photopolymer. Proceedings of SPIE 7957, 79570H.
Bjelkhagen, H.I., 1993. Silver-Halide Recording Materials: For Holography and Their Processing. Springer.
Blanche, P.-A. (Ed.), 2016. Photorefractive Organic Materials and Applications. Springer.
Blanche, P.-A., Banerjee, P., Moser, C., Kim, M.K., 2015. Special section guest editorial: Special section on the interface of holography and mems. Journal of Micro/Nanolithography, MEMS, and MOEMS 14 (4), 041301.
Blanche, P.-A., LaComb, L., Wang, Y., Wu, M.C., 2017. Diffraction-based optical switching with MEMS. MDPI Applied Sciences 7 (4), 411.
del Campo, A., Greiner, C., 2007. Su-8: A photoresist for high-aspect-ratio and 3d submicron lithography. Journal of Micromechanics and Microengineering 17 (6), R81.
Dürr, H., Bouas-Laurent, H., 2003. Photochromism: Molecules and Systems. Gulf Professional Publishing.
Efron, U., 1994. Spatial Light Modulator Technology: Materials, Devices, and Applications, 47. CRC Press.
Glebov, L., 2002. Photochromic and Photo-Thermo-Refractive Glasses. John Wiley & Sons, Inc.

Günter, P., Huignard, J.-P., 2007. Photorefractive Materials and Their Applications. Springer.
Guo, J., Gleeson, M.R., Sheridan, J.T., 2012. A review of the optimisation of photopolymer materials for holographic data storage. Physics Research International 2012.
Lin, L.H., Beauchamp, H., 1970. Write-read-erase in situ optical memory using thermoplastic holograms. Applied Optics 9 (9), 2088–2092.
Moerner, W.E., 1988. Persistent Spectral Hole-Burning: Science and Applications. Berlin, Heidelberg: Springer.
Nikolova, L., Ramanujam, P.S., 2009. Polarization Holography. Cambridge University Press.
Osten, W., Reingand, N., 2012. Optical Imaging and Metrology: Advanced Technologies. John Wiley & Sons.
Schnars, U., Jüptner, W.P.O., 2002. Digital recording and numerical reconstruction of holograms. Measurement Science and Technology 13 (9), R85.
Stojanoff, C.G., 2011. A review of selected technological applications of dcg holograms. Proceedings of SPIE 7957, 79570L. (-79570L-15).
Yetisen, A.K., Naydenova, I., Da Cruz Vasconcellos, F., Blyth, J., Lowe, C.R., 2014. Holographic sensors: Three-dimensional analyte-sensitive nanostructures and their applications. Chemical Reviews 114 (20), 1065410696.

Colour Holography: Perception Versus Technical Reality

Andrew Pepper, Nottingham Trent University, Nottingham, United Kingdom

The opportunity for optical holography to record and display three-dimensional objects (and the volume surrounding them), at extremely high fidelity, has prompted a distorted public perception. Almost since we became aware of the process, and its capabilities, there has been an assumption that the remarkable dimensional images would be in full (natural) colour. This is not a surprising supposition due to the clarity and illusionistic impact of holographic images displaying full parallax (Benton, 1977). Our familiarity with looking into display spaces, very much like looking through a window into a room in a building (Pepper, 2000), suggests that what we appear to 'see' must surely be in full colour.

Most of the first holograms produced were monochromatic; a limitation of the technical processes which created them. When Dennis Gabor first published the announcement about his work with optical holograms in 1947 (Gabor, 1948), he used a heavily filtered mercury arc lamp to record and illuminate his photographic plates, which stored the microscopic interference patterns on which holograms rely. This was the only suitable and accessible source of light at the time, which fit the specific requirements of his optical recording configuration, but it did limit the size of the object he could record to a few millimetres.

These early holograms used Gabor's 'in-line' recording process, which placed the 'object' to be recorded directly between his chosen source of light and the photographic plate. Light waves, directly from the illuminating source, and those modulated through diffraction by the object, were combined on the photographic plate to record the resulting, microscopic, interference wave pattern.

The reference to an 'object' here is intentionally placed in inverted commas as Gabor used a flat photographic transparency for his recording. Although the transparency is physically three-dimensional, and millimetres thick, there is a tendency to describe it as flat because of our learned visual, perceptual and verbal framework used to view and describe 3-D. This aspect of these early recordings, and the demonstration of the optical and physical viability of this new method for recording three-dimensional objects caused a slight perceptual and cultural friction. If it is three-dimensional, it is likely to be incredibly accurate, and if it is so accurate, it is reasonable to assume that it is correctly coloured. Feedback from visitors to exhibitions of optical holography includes incorrect assumptions that some of the monochromatic images they have encountered are, when recalled and discussed, full colour. We have a natural tendency to 'fill in the gaps', and this is incredibly easy to do when confronted with the astounding three-dimensional illusion of a holographic image.

The tendency to make such an incorrect assumption is not an isolated phenomenon. Evidence through teaching fine art to undergraduate students suggests that they encounter much of the art they look at (or explore for research) online. They view multiple images of the same work, made up of gallery installation shots, close-up details and peripherally photographed views, which are scattered across the web or in printed catalogues, and often believe they have 'seen' the original. It is only when they then encounter the actual work that they realise they have not seen the work, only its documentation.

A major shift in public perception of holography was directly related to two significant developments of the technical process in the United States and USSR. Emmett Leith and Juris Upatnieks developed an off-axis geometry, which used coherent, monochromatic, laser light, to record and display holograms. This more powerful, coherent, light source, and the optical pathway they developed, allowed the recording of 'actual' three-dimensional objects (rather than Gabor's 'flat' transparencies).

The results of their research were disseminated to the scientific community (Leith and Upatnieks, 1962) but an article in the widely circulated magazine 'Scientific American' (Leith and Upatnieks, 1965) brought it to the attention of a wider audience, interested in broad scientific topics. Culturally this was a positive, future-looking, period with an enthusiastic public awareness of new technologies stimulated by the space race and rapid scientific advances. This publication informed a different demographic beyond the technical papers previously published in field-specific scientific conferences and journals. Unsurprisingly, the 'improvement' on Gabor's original invention, and its implied connection with photography (in the title of the article and through the text), captured the public view and that of popular media. Generalised information about optical holography spread quickly, with many popular articles appearing worldwide (Johnston, 2006). Subsequently, public interest and expectations about holography increased, which fuelled many inaccuracies - one of the main ones being that holograms were full-colour.

This public confusion about the realistic nature of holograms was evident during the late 1970s when two important exhibitions of large-scale pulsed laser holograms, produced by Nick Phillips (1933–2009) at the Royal Academy of Arts, London, proved extremely successful and not only attracted the attention of an 'art' sensitive public, visiting a renowned institute, but also that of the media. Some of the reports and images disseminated around this time blurred the line between what was possible with the technique and high public expectation. It was difficult to be sure whether the images from these exhibitions, reproduced in glossy colour supplements, might be full colour.

Leith and Upatnieks did propose colour holographic recording as early as 1964 (Leith and Upatnieks, 1964), but there was a need for the technical components to become available and accessible, mainly multiple colour, coherent, laser light and high-resolution recording materials.

During the same period, Yuri Denisyuk demonstrated significant developments, which also used an off-axis system of recording and display (Denisyuk, 1962). This incorporated a single beam of coherent laser light which passed through the holographic recording plate, bathed a three-dimensional object in laser light, and the subsequently reflected light from that object was combined with the laser light shining through the plate, to generate the required microscopic interference pattern. The advantage

 doi:10.1016/B978-0-12-803581-8.09611-9

of this 'single beam' technique (not to be confused with Gabor's in-line method) is that the resulting hologram could be displayed using white light. This significant development played an important part in the adoption of holography as a display (and archival process). All previous systems required expensive, and potentially dangerous, mercury arc (Gabor) or laser (Leith and Upatnieks) light. Here, inexpensive white lights, such as spotlights or direct sunlight, could be used to reconstruct the recorded three-dimensional images.

A drawback to Denisyuk's technique was that objects needed to be relatively shallow, compared to the amount of volume which was possible to record in Leith and Upatniek's technique. It appears the more accessible, English language, publications about Leith and Upatniek's activities placed them prominently in the public eye and that of the US and European media outlets. It is also worth noting that the term 'hologram' was not in common use before the mid-1960s, with terminologies such as wavefront reconstruction, diffraction microscopy, Gaboroscopy, holoscopy, wave photography and lensless photography, appearing in publications (Johnston, 2009).

The various techniques and optical processes were, theoretically, capable of recording full-colour holograms. As early as 1966, Lin and his team successfully produced the first full-colour reflection hologram, which used two different wavelengths of light, recorded and combined onto a single holographic plate (Lin *et al.*, 1966). This was, in fact, a recording of a colour photographic slide, which we might interpret as 'flat' but, like Gabor's original work, it demonstrated the feasibility of using multiple wavelengths (colours) of light which, combined, will produce the impression of full colour.

Red, Green, Blue

Most of the early holograms recorded, irrespective of the optical technique employed, used a monochromatic light source to record and display a three-dimensional object. As previously stated, the optical illusion and impact of the recorded images were so significant that the 'memory' and recall of seeing these holograms often prompted a distorted perception among viewers, thinking they had 'seen' the images in full colour.

An image recorded with red, helium-neon, laser light, for example, will display a tonal red, three-dimensional image, when displayed using the same red laser light. There will be no other colours present. Much of the early research on colour holography used the principle that it would be possible to use multiple colour lasers to record a hologram and then use the same lasers to display the results.

As white light is made up of multiple colours (wavelengths), using the light primary colours of red, green and blue will allow the recording and reconstruction of a full-colour image. This phenomenon of colour reproduction is based on the tristimulus theory of colour vision, which suggests that any colour can be reproduced (or matched) by combining the three light primary colours.

It is possible, therefore, to make a holographic recording using red laser light (helium-neon), then make a second exposure using green laser light (argon) and finally a third using blue laser light (helium-cadmium). The colour of a laser can vary, depending on the gases/materials used within the device to cause the light amplification process. This can include krypton ion (red), neodymium-doped yttrium aluminium garnet (green) and argon ion (blue).

Once processed, the developed hologram can then be illuminated with these three colours of laser light. In effect, each of the three-dimensional recordings in the hologram 'overlap' and then combine, spatially, to reconstruct the colours of the original object in their original position in space.

Three objects are presented, which overlap each other in space and combine to create a single, full-colour object/image. This is a unique property of holography, and one of the few opportunities to place visually solid, three-dimensional images of objects into the same space at the same time. Such 'overlapping' has always been possible with multiple exposure photography and use of multiple camera feeds in video, but only holography allows this in full parallax, full three dimensions.

During the 1970s, there was a considerable amount of research in the field of colour holographic recording, which took advantage of the technical improvement in light sources, recording materials and optical components (Hariharan, 1983).

Although the cost of lasers has dropped and their quality (and variety) improved, they remain complex, and in some cases fragile or dangerous (depending on their power and wavelength). This has limited their use in display. Clearly, the option of recording holograms with multiple colour laser light and then viewing them with white light is a practical solution, and a considerable amount of research has been undertaken, using reflection holography (often employing the Denisyuk process) with a significant pioneering research by Bjelkhagen (2006).

Research suggests that the optical quality and colour rendition of holographic displays can be enhanced using multiple wavelengths of laser light and more than three colour recordings (Mirlis *et al.*, 2005). The reproduction of colour has become accurate enough for it to be viable as an archival system for paintings, which can be captured using a contact recording technique, this not only accurately records the colours within the painting, but also the three-dimensional undulations of the painting surface. Each brush stroke, its thickness, and texture of the paint, is stored and can later be displayed (Bjelkhagen and Vukicevic, 2002). As an archival process, this has considerable potential – extending the visibility of paintings, which now have a restricted display life due to potential light damage. Maintaining their colour is a significant requirement of the preservation process even if, in this case, it is shown in a high-resolution facsimile.

One holographic technique, invented by Stephen Benton (1941–2003) in 1968 (Benton, 1969), impacted on the production of colour holography significantly and was extremely popular within the display, advertising and creative arts industries during the 1970–90s. This two-step process used a laser transmission hologram of a three-dimensional object, which was then 'copied', or rerecorded, onto a second holographic plate through a restricted optical 'slit', with the resulting

hologram being displayed using inexpensive white light. The optical and practical 'trade-off' here was that the display hologram produced a three-dimensional, full-parallax, image when an observer moved from left to right (or right to left) in front of the display. This allowed the observer to look 'round' objects in the hologram offering shifting points of view, as you would expect when looking at a scene through a window. You can, for example, move your point of view and look behind a three-dimensional object in the foreground so that you can see a hidden, three-dimensional object in the background, much like we do in real life. The reason why this type of hologram is of interest in connection with colour holography is that the optical production and reconstruction of this kind of hologram produces a rainbow-colour effect, which 'covers' the displayed image. These are often referred to as rainbow holograms or white-light transmission holograms.

Although parallax exists when moving left to right – when an observer moves up and down, the parallax view remains the same (so the images 'freezes') but also changes colour from red, through green, to blue. Its apparent size can also appear to shrink or expand. Although this limit in parallax viewing and shift in colour is a side effect of the production process, it presented commercial and visual advantages. The technique allowed printed mass production of holograms and prompted a surge in their use on credit cards, publications, textiles, manufacturing plastics, novelty gifts and commercial products.

Pseudo Colour and the Manipulative Artist

Benton's technique is often described as pseudo colour holography because the colours on show have no direct connection to the actual colour of the three-dimensional object which was initially recorded. This term can be applied to any hologram which produces a similar effect and covers both white-light reflection and transmission techniques. A US patent was granted for pseudo colour holography in 1991.

Pseudo colour holographic display was incorporated into a number of different systems, including the separation of two-dimensional graphic images, which could be combined (overlapped) to produce full-colour images. This had considerable impact on mass-produced commercial printing and novelty gift production (McGrew, 1982).

Although this 'inaccurate' or sometimes 'unfixed' colouration within pseudo colour holograms is viewed as a disadvantage to precise reproduction of objects, it is acknowledged by artists and designers as a flexible and creative, spatial, colour technique. Benton understood this and supported many artists in the use of his technique, offering master classes as well as accessible conference publications. A number of key artists adopted this technique (or a variety of it) and tested it to its visual and technical limits, allowing them to 'manufacture' or 'capture' light, change its colour and use it as a sculptural material in space. Artist Rudie Berkhout (1946–2008) is acknowledged as a pioneer in this field, and his colour holograms are in significant public and private collections worldwide. One aspect of the use and development of holographic colour by artists is that they are unbridled from the specific requirements of industry, results-driven, exploration. Where most optical scientists work on a particular problem which requires a quantifiable solution, often set externally by a funding agency or commercial consultation, artists set their own boundaries to develop skills and techniques relevant to their practice. Berkhout used innovative and practical solutions for recording and displaying colours in his work, which might be considered an unlikely requirement, or aberration, in an optics research programme. He was able to create and manipulate three-dimensional objects and images, which were not recorded from a corresponding (recognisable) object, and as his career progressed, this exploration became more abstract and confident.

Berkhout's work has a visual vocabulary developed over many years, and several other artists have used the technical opportunity to display 'pieces' of light and colour within space. A line, shape, mark or word can be presented, unsupported in space, using holographic colour recording techniques. Early works by artist Eduardo Kac displayed 'floating' and overlapped letters and words which made up his series of Holopoems, produced between 1983 and 1993. Unlike Berkhout, they are grounded through the use of text, which can be viewed from multiple angles, as an observer moves around the hologram, offering spatial fluidity and semantic interpolation. There are a large number of pioneering artists who produced exceptional works with colour and their contributions to the visual arts can be found in numerous exhibitions, catalogues, public and private collections.

Another pioneer in the field, artist John Kaufman, used the often unwanted side effect of reflection holography, namely the swelling of the recording emulsion, which results in a colour shift when the hologram is processed and displayed.

The majority of holograms are recorded onto a high-resolution photographic emulsion, which is then wet-processed in chemicals (similar to those used in the development of photographs). Submersion in these chemicals, and wash baths, causes the exposed photographic emulsion to swell. Once dry, the emulsion returns to its original volume/thickness (or very close to it). In white-light reflection holograms, this is an essential element in their production and processing. During display with white light, the processed emulsion absorbs unwanted wavelengths of light and then reflects the wavelength (colour) of light used in the recording. If a red laser is used to record the hologram, the thickness of the processed emulsion should cause only red light to be reflected, and the resulting holographic image would appear red. Occasionally, the processing of the hologram, or inconsistent drying of the emulsion after wet processing, can cause the emulsion to dry to a different thickness. This will change the wavelength of light absorbed and reflected during reconstruction of the hologram and subsequently change its apparent colour.

Artists and display companies have used this to their advantage, and there was a period in the 1980s when reflection holograms for commercial and archival display used emulsion which was intentionally swelled so that the resulting image would appear yellow gold, rather than the original colour of the laser used to record it.

Kaufman developed a technique using chemicals to pre-swell the emulsion on the holographic plate, record a three-dimensional object and, before processing, re-swell the emulsion to a different thickness, record another object, and often repeat

this several times until finally chemically processing the hologram. What resulted was an emulsion which dried to slightly different thicknesses and, when illuminated with white light, would absorb and reflect a variety of wavelengths in the various regions of the holographic plate. The viewer of these works would see fully three-dimensional, full-parallax, images in different colours within the recorded holographic space. This is a controlled system with long production times and a very high degree of uncertainty, as the resulting image is only visible once all the swelling, recording and processing is complete. It is not possible to return to a particular stage of the process and change it – the entire hologram would need to be remade. This limited the use of this process commercially, but enhanced its use in limited, or unique, editions by artists.

Other artists have used a less prescriptive method of emulsion swelling, which subsequently produces a more abstract spatial result. This includes swelling small parts of the emulsion across the surface of the holographic plate without predefining the result. Some artists have used their body surface as a tool for application of swelling agents onto the surface, others incorporate brushes, fine drawing devices or specially cut templates to define and contain the area of swelling. The process functions much like mark-making or the application of paint on a surface but, in this case, the area being manipulated with colour is spatial and volumetric – drawing and colouring in space.

Colour generation and 'shifting' is not limited to emulsion swelling. Several artists exploited the rainbow effect within white-light transmission holography, particularly the production and manipulation of holographic optical elements (HOEs). Already a successful process for generating optical elements with extremely specific characteristics for photonic research and development, they are also used by artists as a method of capturing, structuring and manipulating white light to explore the liminal spaces these constructions can produce. Fred Unterseher, one of the pioneers in this technique, and a significant artist/teacher of accessible holography, commented that; "The virtue of a holography class and the holography experience is to transform the way you see, to transform the way you experience; …. the value of this is to see light and to experience (it) in new ways…" (Johnston, 2006).

As the techniques and technology of holography have developed, there are now a number of commercial companies and research facilities which offer accurate full-colour analogue and digital holographic recording. This has clear commercial impact in military reconnaissance, advertising, architecture, visualisation and museum archiving. Subsequent adoption in this area should stimulate the industry and its competition with illusionistic computer graphics or virtual reality. The undeniable advantage of colour holography continues to be its ability to record and display extremely high-resolution objects in full parallax three dimensions without the need for a viewing device. Where this, perhaps, differentiates itself from developing display technologies is its perceived state of hyper-reality, evidenced by the recent significant advances in ultra-realistic holographic imaging techniques (Bjelkhagen and Brotherton-Ratcliffe, 2016) The ability to specify colours in space can allow significant applications in spatial 'signalling' or object differentiation, while artists have a physical and conceptual opportunity to explore and manipulate colour in 'real' three dimensions.

What began with Gabor in 1947 as an attempt to improve the resolution of the electron microscope has partly developed into a new technical and creative vocabulary for colouring our spatial environment.

See also: Overview: Holography

References

Benton, S.A., 1969. Hologram reconstructions with extended incoherent sources. Journal of the Optical Society of America 59, 1545–1546.

Benton, S.A., 1977. White-light transmission/reflection holographic imaging. In: Applications of Holography and Optical Data Processing: Proceedings of the International Conference, August 23–26, 1976, Jerusalem, Israel, pp. 401–409.

Bjelkhagen, H., Brotherton-Ratcliffe, D., 2016. Ultra-Realistic Imaging: Advanced Techniques in Analogue and Digital Colour Holography. Boca Raton, FL: CRC Press.

Bjelkhagen, H.I., Vukicevic, D., 2002. Color holography: A new technique for reproduction of paintings. In: Practical Holography XVI and Holographic Materials VIII: Proceedings. SPIE, 4659. doi: 10.1117/12.469251.

Bjelkhagen, H.I., 2006. Color holography: Its history, state-of-the-art, and future. In: Holography 2005: International Conference on Holography, Optical Recording, and Processing of Information: Proceedings. SPIE, 6252. doi: 10.1117/12.677173, ISBN: 9780819463111.

Denisyuk, Y.N., 1962. On the reflection of optical properties of an object in a wave field of light scattered by it. USSR Academy of Sciences (Doklady Akademii Nauk) 7 (144), 1275–1278.

Gabor, D., 1948. A new microscopic principle. Nature 161 (4098), 777–778. doi:10.1038/161777a0.

Hariharan, P., 1983. IV Colour holography. Progress in Optics. 263–324. doi:10.1016/s0079-6638(08)70279-8.

Johnston, S.F., 2006. Holographic Visions: A History of New Science. Oxford: Oxford University Press, [ISBN-13: 978-0-10-857122-3].

Johnston, S.F., 2009. The parallax view: the military origins of holography. In: Rieger, S., Schröter, J. (Eds.),Das holographische Wissen, Dortmund, Diaphane, pp. 33–57; ISBN 978-3-03734-071-4.

Leith, E.N., Upatnieks, J., 1962. Reconstructed wavefronts and communication theory. J. Opt. Soc. Am. 52 (10), 1123–1130.

Leith, E.N., Upatnieks, J., 1964. Wavefront reconstruction with diffused illumination and three-dimensional objects. Journal of the Optical Society of America 54 (11), 1295. doi:10.1364/josa.54.001295.

Leith, E.N., Upatnieks, J., 1965. Photography by laser. Scientific American 212 (6), 24–35. doi:10.1038/scientificamerican0665-24.

Lin, L.H., Pennington, K.S., Stroke, G.W., Labeyrie, A.E., 1966. Multicolor holographic image reconstruction with white-light illumination. Bell System Technical Journal 45 (4), 659–661. doi:10.1002/j.1538-7305.1966.tb01050.x.

McGrew, S., 1982. A graphical method for calculating pseudocolor hologram recording geometries. In: Proceedings of the First International Symposium on Display Holography, Lake Forest, USA (Illinois), Lake Forest College.

Mirlis, E., Turner, M.J., Bjelkhagen, H.I., 2005. Selection of optimum wavelengths for holography recording. Practical Holography XIX: Materials and Applications: Proceedings. SPIE 5742, 113–118. doi:10.1117/12.583224.

Pepper, A., 2000. Windows with memories: Creative holography in the real world. This Side Up 10, 15–18. (Summer).

High-Resolution Underwater Holographic Imaging

John Watson, University of Aberdeen, Aberdeen, United Kingdom

Introduction

Holography is well-known to most of us: many will have seen white-light display holograms in exhibitions, shops and on credit cards. The word "hologram" has passed into general usage and has come to be associated, somewhat erroneously, with any three-dimensional or pseudo-3D image regardless of how it was formed. The impact and influence of true holography, however, spreads far and wide beyond display and entertainment confines and into the whole spectrum of science, engineering, commerce and everyday life. Holography is now a standard requirement in vibration analysis, image processing, holographic-optical elements, and optical computers. The potential for high-precision imaging and accurate dimensional measurement makes holography an ideal tool for in situ imaging of plankton and other microscopic-sized aquatic organisms and particles. It is this unique ability of holography to recreate, from a flat optical sensor, a three-dimensional image that is optically indistinguishable from the original that sets it apart from other imaging techniques and makes it so useful in underwater aquatic science and engineering for imaging, identification, and measurement of aquatic organisms and particles.

A detailed knowledge of the distribution and dynamics of marine and freshwater organisms and particles, such as plankton, is crucial to our understanding of the planet's biodiversity and how these organisms affect our environment. Traditional methods of gathering in situ data on aquatic particles, such as collection using nets (which can destroy fragile organisms), electronic counting or photography, are not usually suited to observing precise spatial relationships. The overriding benefit of holography is that it permits non-invasive and non-destructive observation and analysis of organisms in their natural environment, whilst preserving their relative spatial distribution.

Initial applications of holography in underwater measurement were based on recording the holograms on photographic film or plates – so called "classical analogue holography." This method requires wet-chemical processing of the recorded hologram together with its subsequent replay in the laboratory, on dedicated (and expensive) replay systems, to create a 3D image of the original scene. Furthermore, taking classical holography out of the laboratory and into a field environment, requires the use of short-duration pulsed lasers to reduce the effects of vibration or movement in any of the optical components or in the objects being recorded. However, this does have the advantage of freezing the scene at the recording instant, thereby allowing fast moving particles to be imaged.

With the dramatic improvements in electronic sensor technology and computer performance, digital holography (DH) – sometimes known as digital holographic microscopy (DHM) – has now begun to dominate the field of aquatic holography. In contrast to classical methods, digital holography records the hologram on an electronic sensor array, such as a charge-coupled device (CCD) or complementary metal oxide semiconductor (CMOS), and the resulting interference pattern is stored in fast computer memory. Reconstruction of the digital holographic image is carried out, in a computer, by numerical simulation of the beam propagation through the hologram and the resultant image at a particular plane in space with reference to the hologram, viewed on a monitor.

Regardless of whether or not analogue or digital recording is employed, a hologram captures vastly more data than a photograph; the reconstructed scene retains the parallax of the original and is free from perspective distortion. Image resolution is high (a few micrometres under favourable conditions) over a wide depth-of-field and field-of-view. Sequential holograms can record changes within a recording volume over a defined period of time and the wide recording dynamic range allows images to be captured that would otherwise be lost in noise. Additionally, in digital holography, "holovideos" can be produced that capture the time dimension as well as spatial dimensions.

For a general background to classical analogue holography, *Optical Holography* (Hariharan, 1996) provides a good introduction; and for the basic principles of digital holography, *Digital Holography and Wavefront Sensing* (Schnars *et al.*, 2015) and *Digital Holography and Three-dimensional Television* (Poon, 2006) provide comprehensive overviews of the techniques.

Analogue Holographic Recording and Replay

Although digital holographic recording is now the dominant method of underwater holography, it makes sense to start with an outline of classical analogue holography, since this helps to explain the concepts and bring out the crucial advantages of holography for aquatic studies. Of the possible recording methods of holography (whether classical or digital) two in particular, the "in-line reference beam hologram" (ILH) and the "off-axis reference beam hologram" (OAH), find use for high-resolution in situ imaging underwater.

In an ILH, a single laser beam (**Fig. 1**) is directed through the sample volume towards the sensor (photographic or electronic). The optical interference which occurs between light diffracted by the object and the undeviated portion of the illuminating beam is recorded on the sensor. A laser is needed in recording to provide the optical coherence properties necessary to produce the optical interference pattern.

 doi:10.1016/B978-0-12-803581-8.09612-0

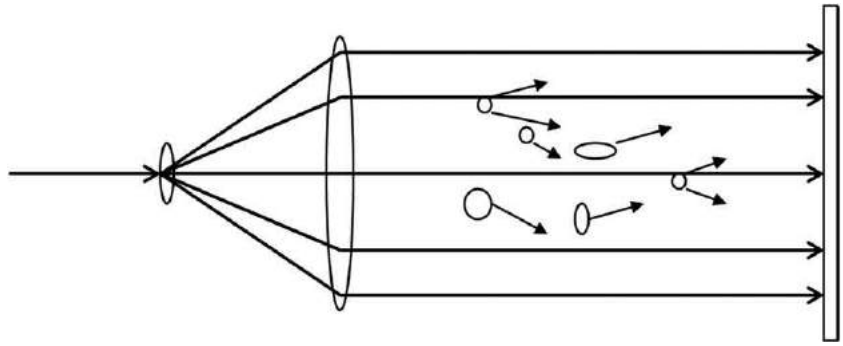


Fig. 1 Recording of an in-line hologram of suspended particles.

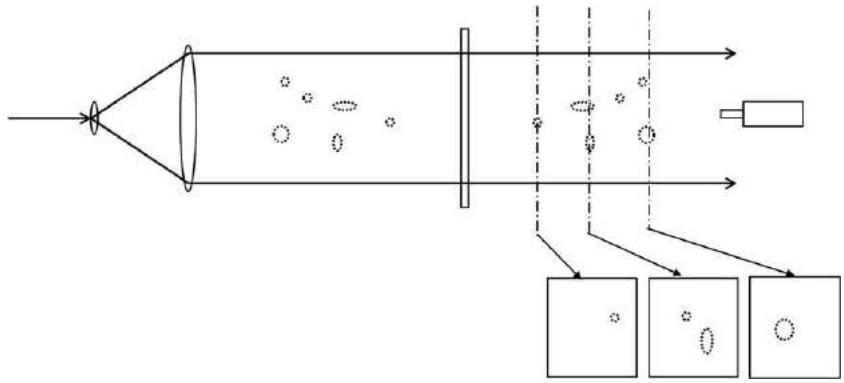


Fig. 2 Replay of an in-line hologram.

The method of reconstructing or replaying the hologram to create a 3D image is where analogue and digital holography differ. In the analogue form, the holographic film must be placed in a replica of the original reference beam (in terms of wavelength, wavefront curvature, phase and direction) and the resultant image viewed with some auxiliary optical system such as a video camera (Fig. 2). The replayed hologram simultaneously forms two images, one "virtual," the other "real," which are located on the optic axis on opposite sides of the holographic plate. These two images are co-axial, and located (for the typical case of collimated recording and replay beams) at equal distances in front of and behind the hologram plane. The virtual image appears to be located behind the hologram and seen through it (like viewing through a window – although this should not be done since the viewer will also see the straight-through laser beam). The real image is formed in space in front of the hologram, between the hologram and the viewer. Although a true 3D real image is produced, because of the dimensions of the sensor it has almost no parallax, and cannot easily be seen by the unaided eye. This image has inverted depth perspective and is seen in darkfield against the out-of-focus virtual image. Traversing a camera through the projected real image allows optical "sectioning" of the image at any plane in the image space.

In OAH, a two beam recording geometry is utilized, although both must be generated from the same laser: one beam illuminates the scene, and the other directly illuminates the holographic film at an oblique incidence angle (Fig. 3). Optical interference occurs between the diffuse light reflected from the scene and the angularly separated reference beam. On replay, the real and virtual images are angularly separated which makes their interrogation easier. OAH is primarily applied to opaque subjects of large volume. To view the "virtual image" in OAH, the processed hologram is replaced in the position in which it was originally recorded and illuminated with an exact duplicate of the original reference wave, in terms of its wavelength, curvature, and beam angle. In this mode, a virtual image is observed as if viewing the scene through a window. Other than colour, this image is almost optically indistinguishable from the original scene.

Once again for data extraction and measurement, the projected "real" image (Fig. 4) is most useful. If we illuminate an off-axis hologram from behind, with a wave which is the exact phase conjugate of the original (i.e., one which possesses the same wavelength as the original but with the opposite direction and curvature), we generate an image that appears to float in the real space in front of the observer. Often a collimated reference beam is used in recording, since reconstruction is then a simple case of turning the hologram around. No lens is needed to form this image, and it is optically identical to the original wave save that it is reversed left-to-right and back-to-front (pseudoscopic). Planar sections of the real image can be optically interrogated by directly projecting onto a video camera (often with the lens removed) which can be moved throughout the 3D volume.

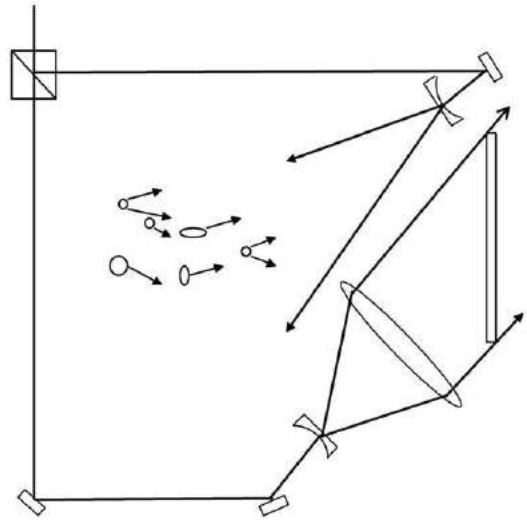


Fig. 3 Recording of an off-axis hologram.

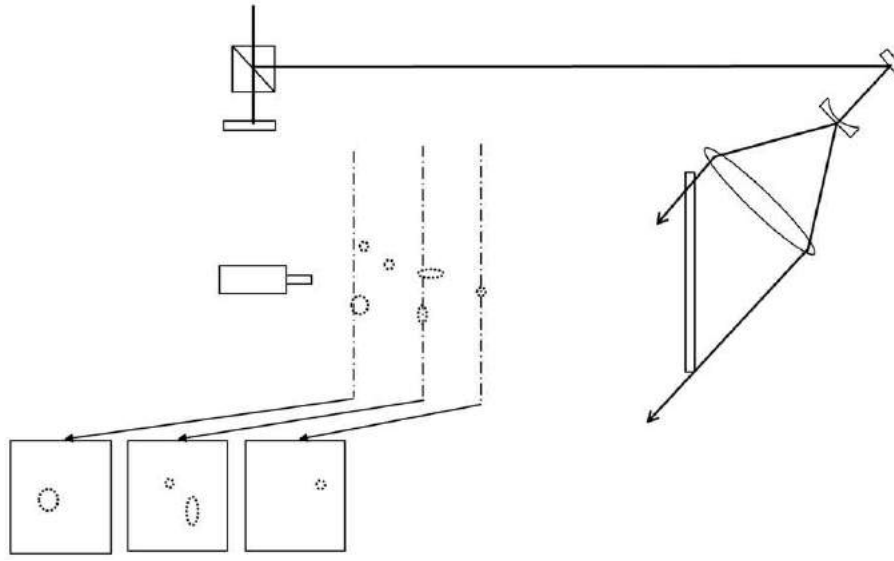


Fig. 4 Real image reconstruction from off-axis hologram.

Digital Holographic Recording and Replay

The concepts of digital holography are broadly similar to those of analogue: the main differences being the recording medium and the method of reconstruction of the holographic image. In digital holography, the interference field is stored directly into computer memory and the hologram is replayed by numerical simulation of the propagation of the optical field through space. In this way planar sections of the image can be recreated at any distance from the hologram plane, at any time, and displayed on a computer monitor, allowing size, location, identification and distribution of particles to be extracted. Computer-based algorithms are used to pre- and post-process the holograms, which are replayed and displayed on a TV monitor at any plane in the recording space. Importantly, the phase as well as the intensity of the wave field is retained on reconstruction. Apart from speed, convenience, and freedom from wet chemical processing, a distinct advantage of electronic recording is the ability to record holographic videos. This allows not only 3D spatial interrogation but allows visualisation of the movement of organisms within a particular scene and tracking the path of individuals.

For digital holographic particle sizing, the ILH mode of recording is generally the favored option by virtue of its geometric simplicity, and consequently its lower cost and potentially higher resolution. Because of the much higher sensitivity of electronic sensors compared with holographic photosensitive film, low power continuous-wave (c.w.) lasers can be used for illumination. In field applications though, pulsed lasers are often a better choice, particularly if the subject is fast-moving or subject to vibration,

or the holocamera cannot be held steady with respect to the target. To record large, opaque or dense aggregates of particles, OAH allows the use of front or side illumination of the subject in conjunction with a separate side or off-axis (angular) reference beam.

One major limitation, though, of digital recording is the lower resolution of electronic sensors compared with that of photographic film. The sensor must be able to resolve the finely spaced interference pattern recorded in the hologram; thus the grain or pixel size has a great bearing on imaging capabilities. The maximum spatial frequency of the resolvable interference pattern depends on the angle between the reference beam and the object beam. The typical grain size of 30 nm or so for holographic film is about two-orders of magnitude greater than that of a typical electronic sensor of about 3 μm pixels. Accordingly the achievable fringe frequency is much greater and reference beam angles up to 45° or so are possible with film, whereas the electronic sensor can only support beam angles up to about 10°. It is this requirement which so far makes in-line recording the dominant method of DH, since the reference beam angle is essentially zero.

Generally the computer algorithms developed for digital recording and numerical replay of a hologram are based upon the Fresnel–Kirchhoff formulation of a complex monochromatic wavefield propagating from a scene through the aperture of an optical sensor. The algorithms employed achieve computational efficiency by applying appropriate approximations and simplifications to make the Fresnel–Kirchhoff equation more suitable for digitisation and computer implementation. In most cases, the integral is implemented as a Fourier transform, thereby allowing the use of existing Fast Fourier Transform (FFT) libraries that can be easily incorporated into software. It is possible to recreate the intensity and phase distribution of the wavefield at any plane in the reconstructed volume of the hologram; thereby simulating the effect of traversing an image sensor through an optically replayed hologram (but with the added advantage that phase information can also be extracted).

Subsea Holography and Holocameras

In general, the usefulness of holography for accurate inspection and measurement is dependent on its ability to reproduce an image of the object which is low in optical aberrations and high in image resolving power. Regardless of whether or not the hologram is recorded underwater, high resolution, high contrast, and low noise are the dominant requirements. All the primary monochromatic aberrations (spherical aberration, astigmatism, coma, distortion, and field curvature) of any optical system may be present in the holographic image. For precise and accurate reconstruction of the real image, the reconstruction reference beam should be an exact copy of the original (where possible), but opposite in curvature (the phase conjugate) of that used in recording. Under these conditions, the lateral, longitudinal, and angular magnifications of the real image with all equal unity and aberrations can be reduced to a minimum.

When holography is applied subsea, the hologram is recorded in water but the image is replayed, in the laboratory (if classical) in air, or by computer (if digital). Because of the refractive index mismatch between recording and replay spaces, optical aberrations in the reconstructed image will be exaggerated, and can impair the potential for high-precision measurement. In ILH, only spherical aberration is significant since both reference and object beam angles are normal to the recording plane. However, in OAH astigmatism and coma dominate; they increase the greater the field angle of the subject in the reconstructed image. These limit resolution and introduce uncertainty in co-ordinate location. These additional aberrations are entirely due to the refractive index mismatch between object and image spaces and are unconnected with the holographic process itself. Furthermore, the water itself may be expected to influence the quality of the images produced. An increase in the overall turbidity of the water will adversely affect both ILH and OAH recording and will create background noise that will reduce image fidelity. In most underwater applications the objects (or the camera) are in motion during the exposure. The effect of this motion is to blur out the finer fringes, and thus reduce resolution and contrast.

A recurring issue in all forms of holographic recording (digital or classical) of microscopic particles, and one which imposes limits on the wider use of the technique, is extracting, processing and analyzing the vast amount of data contained in a hologram. Although analysis of classical holograms, can be performed by manual scanning of the optical reconstructions, this is tedious and time-consuming and requires high levels of concentration; automatic interrogation of the data in a dedicated reconstruction facility is essential for all practical applications of the technique. Since the main thrust of this article is on digital holography the following discussion is restricted to subsea digital methods.

Since a single holovideo may contain as much as several gigabytes of data, a first step in interpretation of digital holograms and holovideos of particles often involves localizing every particle within a frame (in x, y, z, t), and distilling particle shape and positional information from it. This facilitates application-specific processing, with subsequent image recognition, particle tracking, counting and sizing. For each video frame, the hologram is reconstructed in incremental steps and an image of each slice parallel to the sensor plane is sequentially obtained. This is equivalent to discretized simulation of the wave-field projected into real image space by an analogue hologram when reconstructed by an optical beam. When a particle coincides with a reconstruction plane it will appear in-focus, and is characterized by a maximization of image gradients at the particle edges.

Focus detection is typically the first step in hologram analysis and with suitable software, particle identification can be carried out on focused particle images and 3D plots of relative position and distribution produced. Most of these algorithms implement optimal-focus metrics which depend variously on image intensity gradient, variance, correlation, histograms and frequency-domain analysis. These metrics all rely on the premise that focused images have higher information content than blurred (out-of-focus) images due to the existence of larger gradients with higher variance across them. This leads to a greater deviation between maxima and minima in the brightness histogram and the local maximization of power in higher frequency components when the image is transformed to the frequency-domain. Due to "speckle effects" in the hologram an algorithm with good noise immunity

is required. One such algorithm relies on estimating the contour gradient around a particle and has been used to analyze a number of holographic videos recorded in the North Sea.

The first-known use of holography for recording living marine plankton was by Knox in 1966, who recorded classical in-line holograms, using a low coherence ruby laser, of a variety of living marine plankton species in a tank. Following this pioneering work, a host of subsea holographic systems, based on classical holography were developed and deployed up until the late 1990s and early 2000s (e.g., Carder, 1979; Katz 1999). Although these holocameras successfully demonstrated the power and ability of the technique, they were big, bulky, heavy and not easy to deploy from research vessels or remotely operated vehicles (ROVs) and autonomous underwater vehicles (AUVs).

With the development of high-resolution electronic sensors such as CCD or CMOS together with increased power and speed of modern PCs and high storage methods, digital holography took over from the early 2000s. Now holography was freed from the constraints of laboratory replay, wet-chemical processing and could be replayed in near real-time. Other advantages included recording with low-power diode lasers (under appropriate conditions) and the introduction of the time dimension allowing holographic videos to be captured.

Digital Subsea Holocameras

The first digital holocamera for underwater applications was developed by Owens and Zozulya (2000). It utilized a 10 mW continuous wave (c.w.) diode laser in an in-line configuration onto a CCD array over a maximum in-water path length of 25 cm.

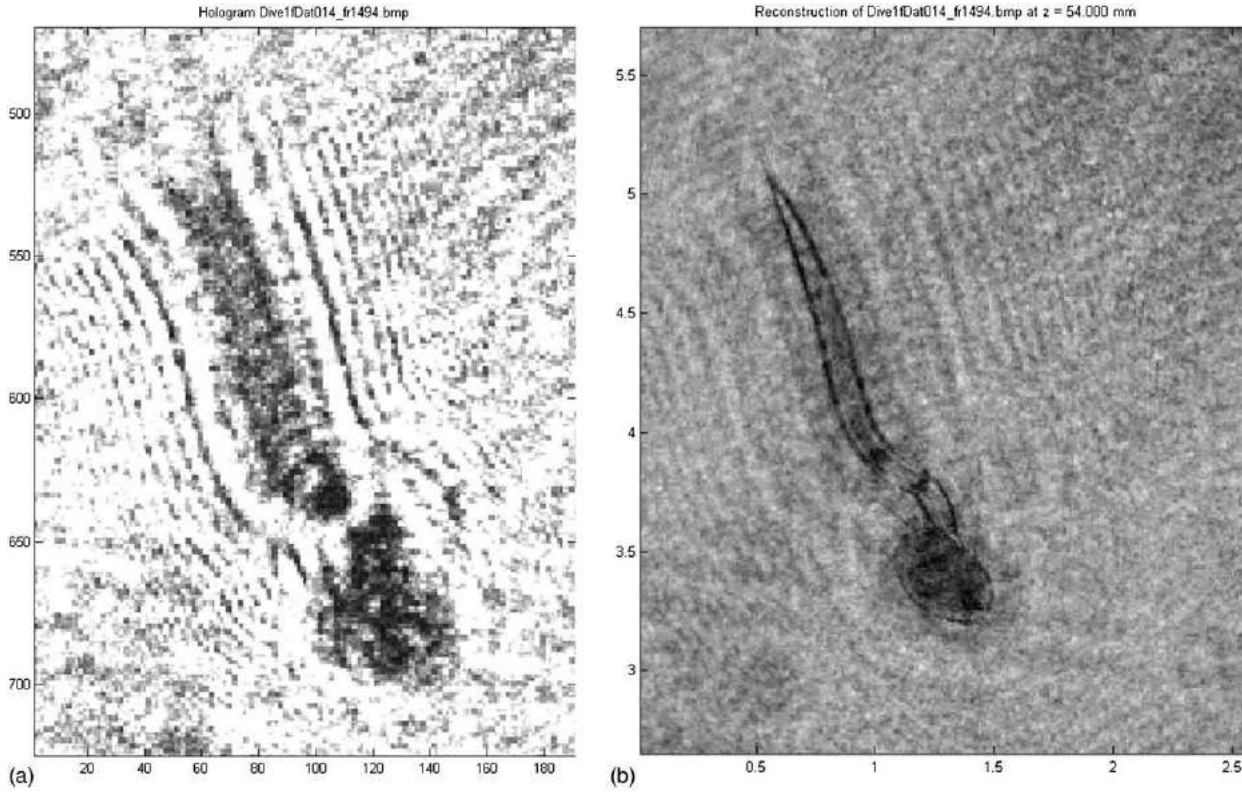


Fig. 5 Appendicularia (a) unreconstructed hologram, (b) reconstructed image (scale in mm).

Fig. 6 Internal layout of *eHolocam* showing the laser, power supply and computer on the right and the smaller CMOS housing on the left.

The use of a c.w. laser can be justified on the grounds that the enhanced sensitivity of electronic sensors over photofilm reduces exposure times to about 100 μs. However, this still limits application to slowly moving systems. This holocamera was successfully deployed to depths of about 50 m in Tampa Bay, FL.

Since then, several groups world-wide have exploited DH for underwater environmental science (e.g., Dyomin *et al.*, 2003; Jericho *et al.*, 2006; Nimmo Smith, 2008); and submersible digital holocameras are now becoming commercially available. Although these systems share common elements such as use of the in-line geometry and, usually, charge-coupled device (CCD) arrays, there are subtle differences amongst them relating to their application and deployment method. Many were developed primarily for studies of phytoplankton over sampling volumes of around 1 mm^3; their use of continuous lasers limits them to low particle velocities.

One such system is shown in **Fig. 5**; and known as "*eHoloCam*" (Sun *et al.*, 2007). It embodied a different approach from many of the others. Although still adopting in-line holography, *eHoloCam* utilized a collimated (parallel) beam from a pulsed frequency-doubled Nd-YAG. For subsea purposes, a laser with a wavelength in the blue-green region of the spectrum will match the peak transmission window of seawater. A frequency-doubled Nd-YAG laser with a wavelength of 532 nm is a good choice, although ruby lasers operating at 694 nm are also acceptable if transmission distance is not an issue. *eHoloCam* was designed with larger planktonic species in mind and featured a large recording volume of 36.5 cm^3 over a 453 mm path length to allow the sparser zooplankton species to be captured rather than the more abundant, and smaller, phytoplankton.

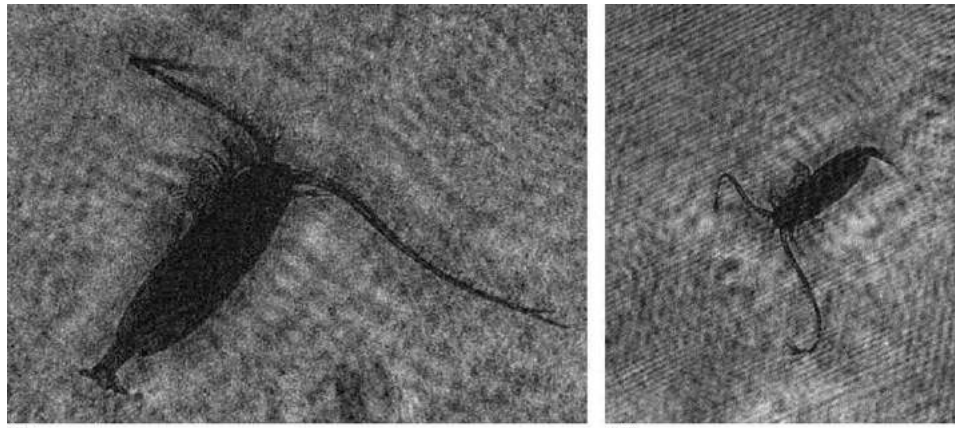


Fig. 7 Reconstruction of a calenoid copepod.

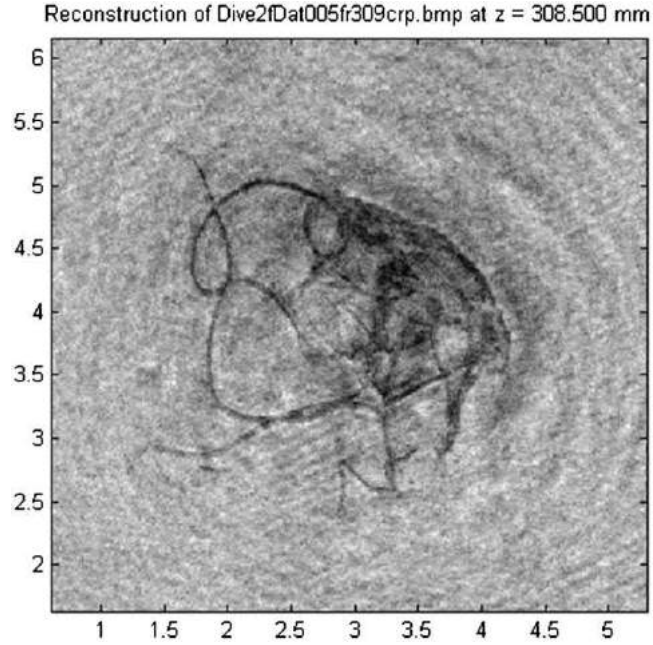


Fig. 8 Reconstruction of a jellyfish larvae (scale is mm).

eHoloCam consists of two water-tight housings: the larger housing contains the laser, a single board computer, storage hard drives and beam forming optics (Fig. 6). The second, smaller, housing contains a CMOS sensor of 2208 × 3000, 3.5 μm square pixels on a 10.50 mm × 7.73 mm area. Both housings were designed to an operational pressure of 30 MPa (3 km depth). Sapphire windows allowed beam transmission through the water from laser to sensor. Holograms were recorded at video frame rates up to 25 Hz.

eHoloCam was deployed in the North Sea from the RV Scotia (Marine Scotland Science, Aberdeen, Scotland) on four cruises over 1 year (December 2005 to December 2006). It was operated from a sampler frame Auto-Recording Instrumented Environmental Sampler(ARIES – Marine Scotland Science, Aberdeen, Scotland) and the short pulse duration (4 ns) allowed it to be towed at up to 4 knots (about 2 m s^{-1}) to depths of 450 m. Water flows through the "jaws" of eHoloCam laterally and perpendicular to the beam path. The system is self-contained in power and data storage and executes embedded control software. The use of ILH and pulsed laser allowed hologram recording at energies of less than a millijoule.

Hologram frames have been reconstructed using the angular spectrum algorithm and manually extracted from the recorded holovideos. Fig. 6 shows an Appendicularia firstly as an unreconstructed hologram and then shows the reconstructed, focused image. Fig. 7 shows a calenoid copepod and Fig. 8 shows a jellyfish larvae.

Conclusions

Since its inception in the early 1960s, holography has shown itself to be particularly useful for studies of microscopic species in the aquatic environment. Its non-destructive and non-invasive nature lend itself particularly for this environment. We have seen that holography in both its classical analogue form and its digital form provides significant advances and features not present in conventional methods of imaging. These features such as 3D, optical sectioning, freedom from parallax and perspective effects, high depth-of-field all offer important benefits to aquatic biologists in their drive to understand the complexities of our aquatic environment. In particular, digital holography also offers the ability to record holographic videos which allow the preservation of the time dimension as well as particle location. Now with the improvements in electronic sensors and computer technology, in-line digital holography is well on its way to becoming a standard tool for studies of microscopic organisms in the sea. Specifically holography aids biologists in their studies of the behavior of organic species such as plankton and other marine organisms and aquatic particles in the oceans of the world. With further development holography is set to become one of the most valuable tools available for aquatic biologists.

See also: Overview: Holography

References

Carder, K.L., 1979. Holographic microvelocimeter for use in studying ocean particle dynamics. Optical Engineering 18, 524–525.
Dyomin, V.V., Polovtsev, I.G., Makarov, A.V., *et al*., 2003. Submersible holocamera for microparticle investigation: Problems and solutions. Atmos Oceanic Optics 16, 778–785.
Hariharan, P., 1996. Optical Holography, second ed. Cambridge: Cambridge University Press.
Jericho, S.K., Garcia-Suserquia, J., Xu, W., Jericho, M.H., Kreuzer, H.J., 2006. Submersible digital in-line holographic microscope. Review of Scientific Instruments 77, 043706.
Katz, J., 1999. Submersible holocamera for detection of particle characteristics and motions in the ocean. Deep-Sea Research 46, 1455–1481.
Nimmo Smith, W.A.M., 2008. A submersible three-dimensional particle tracking velocimetry system for flow visualization in the coastal ocean. Limnology & Oceanography: Methods 6, 96–104.
Owen, R.B., Zozulya, A.A., 2000. In-line digital holographic sensor for monitoring and characterizing marine particulates. Optical Engineering 39, 2187–2197.
Poon, T.C., 2006. Digital Holography and Three-Dimensional Television. New York, NY: Springer.
Schnars, U., Falldorf, C., Watson, J., Jüptner, W., 2015. Digital Holography and Wavefront Sensing. Heidelberg: Springer.
Sun, H., Hendry, D., Player, M., Watson, J., 2007. In situ electronic holographic camera for studies of plankton. IEEE Journal of Oceanic Engineering 32, 373–382.

Further Reading

Burns, N.M., Watson, J., 2011. A study of focus metrics and their application to automated focusing of inline transmission holograms. Imaging Science Journal 59 (2), 90–99.
Burns, N.M., Watson, J., 2014. Robust particle outline extraction and its application to digital in-line holograms of marine organisms. Optical Engineering 53 (11), 8. doi:10.1117/1.OE.53.11.112212.
Gopalan, B., Katz, J., 2015. Turbulent shearing of crude oil mixed with dispersants generates long microthreads and microdroplets. Physical Review Letters 104 (054501), 1–4.
Knox, C., 1966. Holographic microscopy as a technique for recording dynamic microscopic subjects. Science 153, 989–990.
Watson, J., 2013. Subsea imaging and vision: An introduction. In: Watson, J., Zielinski, O. (Eds.), Subsea Optics and Vision. Cambridge: Elsevier, pp. 17–34. ISBN: 978-0-85709-341-7.
Watson, J., Burns, N.M., 2013. Subsea holography and submersible holocameras. In: Watson, J., Zielinski, O. (Eds.), Subsea Optics and Imaging. Cambridge: Elsevier, pp. 294–326. ISBN: 978-0-85709-341-7.

Digital Holographic Display

Daping Chu, Jia Jia, and Jhensi Chen, University of Cambridge, Cambridge, United Kingdom

Overview of a Holographic Display

History of Holography

Holography was invented by Dennis Gabor in 1948 for the correction of electron lens aberration to improve electron microscope resolution (Dennis, 1948). The word of holography is composed of 'holo-' and '-graphy', which stand for 'whole' and 'writing', respectively. The advantage of being able to reconstruct the wavefront in space was appreciated gradually, and the research works on various holographic applications started to develop.

The electron wave based holography was not immediately implemented in optical applications because of the lack of adequate coherent light source at the time. Mercury arc lamps were used commonly as the light source and the image quality of the holographic experiments was poor due to the low coherence of light. The situation changed completely after the laser was invented in 1960, and the first practical optical hologram with recorded 3D objects was realized in 1962 by Yuri Denisyuk in the Soviet Union (Denisyuk, 1962) and Leith and Upatnieks at the University of Michigan (Leith and Upatnieks, 1962).

Although people have been hoping to realize 3D images through holography for decades, as shown frequently in novels and sci-fi movies, there has been almost no actual implementation commercially so far. Instead, holography finds its applications in holographic optical elements (HOE), such as holographic gratings and holographic lenses, and in interferometry as a standard optical measurement technology, known as the non-destructive testing and holographic microscopy. It also found its way back in the aberration correction for optical components, as what Gabor intended to at the beginning. Furthermore, holography is used in anti-counterfeiting and authentication, such as the holograms laminated on credit cards and identification documents. It also has the potential to be used in optical data storage and optical computing. Nevertheless, this review will focus on the development and issues related to holographic displays, in particular digital holographic displays.

Principle of Holography

A light wave, or an electromagnetic (EM) wave more generally, can be described by using its amplitude and phase. Consider a certain area, such as a square or a circle, in a space which the light flux propagates through. The basic concept of holography is to record the complex light field (both amplitude and phase) in this area and then reconstruct it for various applications. This principle can also be extended to non-EM situations, such as electron waves used in the electron holography. For display applications, the light waves to be considered are in the visible range with a wavelength between 400 and 700 nm.

Recording a complex light field was done traditionally by using photosensitive materials, which response to light intensity rather than phase. Using a single coherent light source, the object light scatters from an object and interferes with the reference light in space. The interference pattern carries the information of both the amplitude and phase of the wavefront of the object light. Such a pattern is called a hologram, or to be precise an intensity hologram when recorded on the photosensitive material. To reconstruct the recorded wavefront of the object, a reference light is used to illuminate the intensity hologram from the same angle as it was recorded.

Fig. 1 illustrates the conventional way of recording a hologram and reconstructing a holographic image. Detailed mathematical description of these processes can be found in Benton and Bove (2008). In practice, it requires suitable equipment and finds experimental skills to successfully produce high quality holograms and reconstructed images.

Using intensity holograms produced by interference to record and reconstruct wavefront is not the only method of holography. The main reason to use an intensity hologram was historical because of the difficulty in modulating the phase of light directly in the past. As the technology progresses, it is now possible to use an array of mega pixels on a phase-only spatial light modulator (SLM) to form a phase-only hologram and modulate the phase of light. This allows us to manipulate the wavefront of light directly without the need to manipulate its intensity, hence the maximum light efficiency.

Both intensity and phase-only holograms can be generated by using appropriate SLMs. Digital holographic displays use those SLMs which are pixelated.

Digital Holography and Computer Generated Hologram

Holographic interference patterns, i.e., holograms, carry the information of light propagation, and can be sampled (recorded) discretely. This discrete information can be kept and reconstructed digitally, and this technique is called digital holography and the corresponding holograms and displays called digital holograms and digital holographic displays, respectively. Computer generated hologram (CGH) is a part of digital holography, and it specifically means those digital holograms which are generated by calculations on computers. Calculation of CGHs involves light propagation calculations and complex information approximation (further information see Section Computer Generated Holograms). Note that the wave information which a CGH carries can be

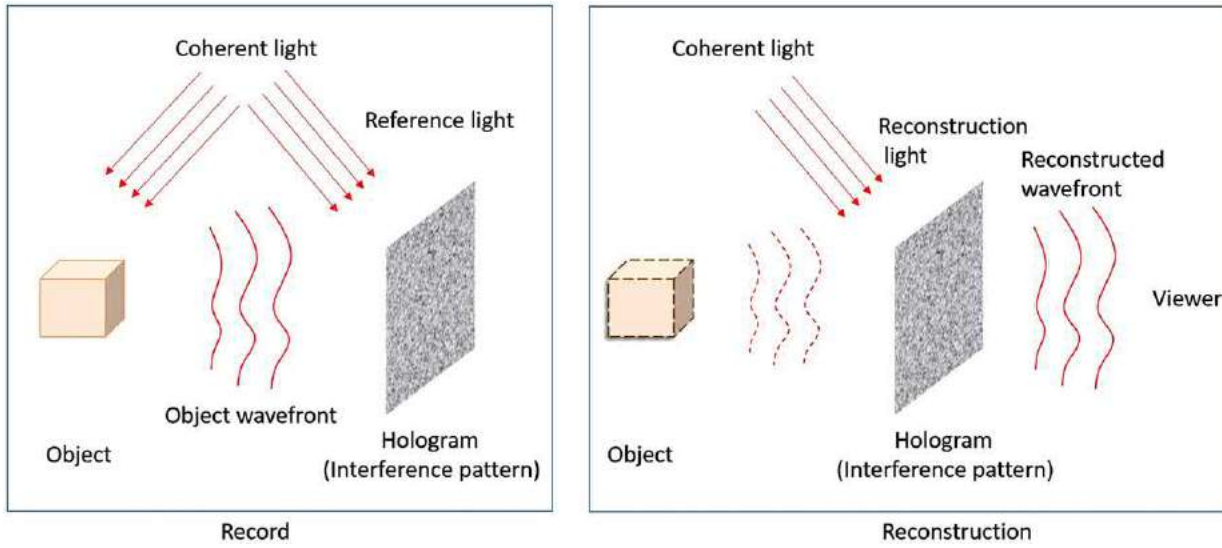


Fig. 1 Illustration of the recording and reconstruction of a 3D scene using holography.

any kind of light propagations, including image and beam propagations. The hardware to load CGH and reconstruct images is mostly of a pixelated structure except few exceptions, and this will be discussed in Section Hardware for a Holographic Display.

Hardware for a Holographic Display

Holograms can be recorded on photosensitive materials such as silver halide films from Geola (Zacharovas *et al.*, 2001), holographic photopolymer from Bayfol (Jurbergs *et al.*, 2009) and DuPont (Gambogi *et al.*, 1994), as described in Section Principle of Holography. These photosensitive materials are one-off use in general and not suitable for dynamic display applications (Berneth *et al.*, 2013; Bjelkhagen *et al.*, 2008; Cody *et al.*, 2012; Stevenson, 1997). Here, we will introduce only the materials and devices suitable for dynamic holographic displays, i.e., for holographic video displays.

Passively Dynamic Holographic Display Devices

One straightforward approach to display holographic images at video rate is to reconstruct sequential static images from sequential static holograms, just as the way which movies are shown using static image films. To implement this approach, updatable mediums, on which holograms can be recorded and removed repeatedly, were proposed and researched, such as a photorefractive polymer (PRP) films (Blanche *et al.*, 2010a,b; Kinashi *et al.*, 2012; Tay *et al.*, 2008) and optical addressed spatial light modulators (OASLMs) (Shrestha *et al.*, 2015; Slinger *et al.*, 2004; Stanley *et al.*, 2004) and azobenzene-based devices (Rasmussen *et al.*, 1999; Shishido, 2010). They can be regarded as passive dynamic holographic display devices (PDHDDs). "Passive" here means that the hologram information is provided from another device, and a PDHDD only carries it passively.

PRP films are simple in configuration and can support nanometer size grains, but they need high voltages ($>$kVolts) for its operation. OASLMs, on the other hand, can be operated at low voltages ($<$10 Volts), but with much larger reported spot sizes. Researches on both devices have improved their performance over the years, leading to better resolution, higher sensitivity, faster response time and larger area with good uniformity. A paper in 2015 shows that it is able to achieve sub-micro resolution for an OASLM (Shrestha *et al.*, 2015). In addition, updatable hologram materials have shown further potential in applications, such as data storages and futuristic optical displays for optical computing systems.

To use updatable holographic films for dynamic holographic images projection, it requires holograms to be written on the films in real time. This is not normally feasible if the diffraction patterns come from a real object directly. However, it is possible to do so if they are provided by active dynamic holographic display devices (ADHDDs), which will be introduced in the next section. Nowadays, almost all reported demonstrations of updatable holographic films are using ADHDDs to write target diffraction patterns on them.

One advantage of using a PDHDD device is that it can provide a large aperture and hologram size for the holographic image reconstruction. Existing ADHDDs all have limited apertures and hologram sizes, and this limits the minimum image spot size it can reconstruct (Hecht, 1998). The use of a PDHDD allows holograms from an ADHDD to be written and tiled up on it, to provide a better image quality in reconstruction. Nevertheless, as spatial light modulators (SLMs) develop over the time, 4k $\times$ 2k chips are now commercialized (Jasper Display Corp., 2017; Texas Instruments, 2017) and 8k $\times$ 4k chips are also available (Senoh *et al.*, 2013), and hence the aperture issue is of a less concern.

Active Dynamic Holographic Display Devices

An ADHDD is the kind of SLMs which modulate light directly by themselves. In holographic display applications, two types of ADHDDs are widely used, pixelated type SLMs and acoustic-optic modulators.

A pixelated SLM consists of a large number usually millions of pixels, each of them can be controlled individually to modulate either the amplitude or the phase of light in normally a digital manner. It is possible to control even both of the amplitude and phase simultaneously through cascade approaches (Hsieh *et al.*, 2007; Jesacher *et al.*, 2008a). The pixel pitch has been brought down over from 15–20 μm a few years ago to 1–5 μm now (Jasper Display Corp., 2017; Senoh *et al.*, 2013; Isomae *et al.*, 2016; Oppenheim and Lim, 1981). It is expected that the pixel pitch will become even smaller in future. Holographic images are reconstructed usually with a plane wave illuminating the holograms uploaded on the pixelated area of an SLM. Approaches to generate holograms from computation will be introduced in Section Computer Generated Holograms.

Liquid crystal on silicon (LCOS) devices are probably the most commonly used pixelated SLMs in a holographic display. LCOS combines the functionality of the cutting-edge CMOS integrated circuit technology with a nonlinear electro-optic material which could be switched by low voltages. The device operates in the reflective mode and its construction is similar to that of a liquid crystal cell apart from that a silicon backplane with an reflection layer is used in place of one of the traditional glass substrates. There are various types of LCOS, including grey-level amplitude-only, grey-level phase-only ones as well as binary amplitude-only and phase-only ones using ferroelectric liquid crystals (ferroelectric LCOSs). A cross section sketch of a typical phase-only LCOS device is shown in **Fig. 2** (Zhang *et al.*, 2014). In holographic displays, phase-only LCOSs are normally used because of its high light efficiency and good reconstruction image quality (Oppenheim and Lim, 1981). Other applications of phase-only LOCSs can be found in Collings *et al.* (2011). Ferroelectric LCOSs are also often used because they have very high frame rate. Temporal based methods are used in the image reconstruction by Ferroelectric LCOS SLMs, to compensate the low image quality due to the binary nature. The advantage of LCOS SLM technology includes: benefit on the development of silicon CMOS technology, integration of high performance driving circuitry on the silicon chip, high pixel fill factor, high quality process technology for excellent pixel mirror reflectivity, and scalability to smaller feature size and higher pixel number.

Digital micro-mirror devices (DMDs) are also a widely used in holographic displays. Although a DMD can only support binary amplitude modulation, its high frame rate (up to 30 kHz (Texas Instruments, 2017)) makes it attractive since it can support a large information bandwidth for a holographic video display with large image size and viewing angle.

Apart from LCOS and DMD, there are other SLMs developed previously which are less used for holographic displays nowadays, such as liquid crystal transmissive SLMs which have low fill factors. There are also some devices under development but not yet fully realized, such as diffractive nano-devices (Sun *et al.*, 2013) and flexo-electro LCOS (Chen *et al.*, 2009).

In addition to the pixelated digital devices, an acoustic-optic modulator (AOM) has also been applied in holographic display. An AOM is an analog-type SLM without pixelation structure and it is often used in laser Q-switches (Hecht, 1998). It modulates light using the acoustic-optic effect of certain materials, which refractive index changes when an acoustic wave passes through it. The acoustic wave modulated periodic index changes form an effective grating which modulates the light accordingly. Different to a pixelated SLM, which reconstructs the object light for the whole target image simultaneously, an AOM delivers one light beam at a single time. This means that with an AOM the object light is reconstructed beam by beam. In other words, an image is generated pixel by pixel in time sequence. Well-known holographic displays built on AOM(s), the Mark system, will be introduced in Section Selected Holographic Display Systems.

Bandwidth Requirements for Decent Display Size and Viewing Angle

We define the information bandwidth of a device or display system as the product of its spatial bandwidth (number of pixels to provide at a given time) and temporal bandwidth (refresh rate or frame rate).

To generate a genuine holographic 3D image from diffraction, we can calculate how much spatial information or the number of pixels is required to support the given specifications of an image. The viewing angle, $\Delta\theta$, is:

$$\Delta\theta = \sin^{-1}(W/2d) \tag{1}$$

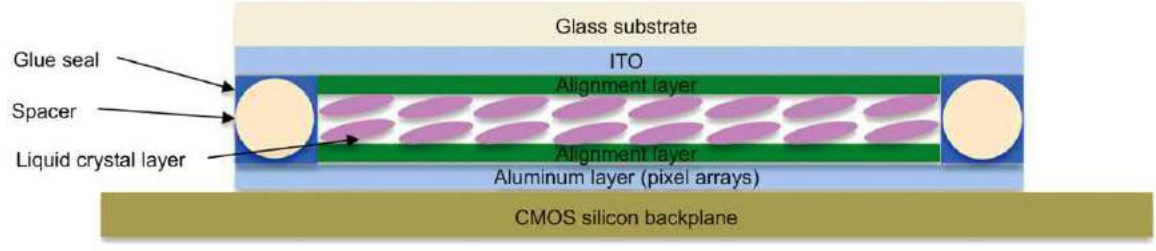


Fig. 2 Cross section of a typical phase-only LCOS device. Reproduced from Zhang, Z., You, Z., Chu, D.P., 2014. Fundamentals of phase-only liquid crystal on silicon (LCOS) devices. Light Sci. Appl. 3, e213.

and the image size, Δh, is:

$$\Delta h = d \sin \theta_{\mathrm{d}} = d\lambda/p \tag{2}$$

where W is the SLM width in one dimension, d the distance between the SLM and the reconstructed image, θ_{d} the diffraction angle for the 1st order, λ the wavelength, p the hologram pixel pitch.

In the far field $d \gg W$, $\Delta\theta \sim W/2d$ and the product of Eqs. (1) and (2) gives the number of pixels required, n, as:

$$n \simeq 2\Delta h \Delta\theta/\lambda \tag{3}$$

This result is the same as that obtained from a Fourier holographic display geometry in Stanley *et al.* (2000).

Assuming the targeted viewing angle, θ, is tiled by a number of $\Delta\theta$ and the targeted image size, H, is tiled by several Δh. For a holographic display of two dimensions with the image sizes, H_x and H_y, and the corresponding viewing angles, θ_x and θ_y, respectively, the spatial bandwidth or the overall number of pixels, N_{tot}, needed to reconstruct such an image will be from Eq. (3):

$$N_{\mathrm{tot}} = N_x N_y \simeq 4 H_x H_y \theta_x \theta_y / \lambda^2 \tag{4}$$

Multiplying this number of pixels with the required frame rate will give us the necessary information bandwidth.

Considering a holographic video display with a display size of 200 mm × 200 mm, viewing angles of 30° × 20° and frame rate of 60 Hz, it would require the system to have an information bandwidth capable of delivering at least $\sim 10^{12}$ pixels/s for just the red colour ($\lambda \sim 633$ nm). If we want to support full colour with additional green and blue, the information bandwidth required will be more than tripled.

However, existing display devices described in this section can only manage to deliver at most $\sim 10^{10}$ pixels/s, which is two orders less than that required. Besides, the available computational speed and data bus transfer rate at present are also nowhere near the levels needed for real time videos. Such hardware limitations significantly restrict the development and deployment of holographic 3D displays.

Selected Holographic Display Systems

In the last a few decades, there have been many holographic display prototypes reported. We review here a few representative systems, to illustrate how different holographic displays are developed.

QinetiQ system

QinetiQ developed a holographic 3D display which is known as the so-called Active Tiling (AT) display system (Stanley *et al.*, 2000; Slinger *et al.*, 2001), in which an electrically addressed SLM (EASLM) is used to write the hologram onto an OASLM time-sequentially. The OASLM works as the information keeper. After all the holograms are written onto an OASLM array, the images are read out at once. Because the EASLM can be driven very quickly and the OASLM array can reconstruct a hologram at a larger size, the AT system potentially can deliver several giga-pixels at once.

In 2004, QinetiQ achieved an active tiling system that combines 4 channels as shown in **Fig. 3(a)** (Stanley *et al.*, 2004), each of which contains one 1024 × 1024 × 750 Hz binary EASLM (ferroelectric LCOS) and a 5 × 5 OASLM array with the total equivalent resolution of 5120 × 5120. As a result, the system displayed a CGH of 20,480 pixels wide × 5120 pixels high at a pixel pitch of 6.6 μm and a frame rate for the whole hologram up to 30 Hz, which is equal to 3000 Mpixel per second. As a result, full parallax holographic images with the size of 140 mm wide were reconstructed as shown in **Fig. 3(b)** (Stanley *et al.*, 2004).

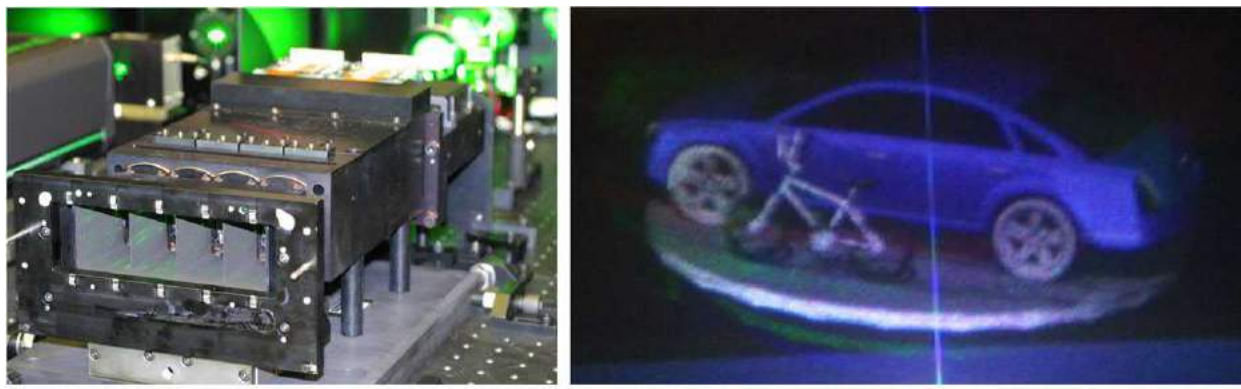


Fig. 3 (a) The QinetiQ 4 channel Active Tiling system; (b) image of an animated movie, showing a full parallax and 140 mm wide holographic image replayed by using a 100 Mpixel Active Tiling system. Reproduced from Stanley, M., Smith, M.A., Smith, A.P., *et al.*, 3D electronic holography display system using a 100-megapixel spatial light modulator. Proc. SPIE 5249, 297.

Mark systems

The Mark holographic systems are developed by Massachusetts Institute of Technology (MIT) Media Lab. There have been a few versions already since the first Mark I system was developed in early 1990s (Lucente, 1994; Smalley, 2006; Smalley *et al.*, 2007). The work in 2013 (Smalley *et al.*, 2013) showed anisotropic waveguide-based modulators (AWM) with more than 40 channels. Each channel can deliver equivalent to 100 million pixels per second (100 Mpixels/s). The total amount of information in pixels reaches the order of Gpixels/s. The diffracted output of light from the modulator was scanned into 2D with horizontal and vertical galvanometric mirrors. The reported holographic video monitor based on AWM is shown in **Fig. 4**. The holographic stereogram images were generated with a native resolution of 296 pixels × 156 pixels for one view, then the image resolution was improved to 29,600 pixels × 156 pixels by scanning, and finally to a composite resolution of 355,200 pixels × 156 pixels by stitching 12 of these images together. The size of display area is 35 mm × 20 mm. The system is still under development, and new versions of Mark IV and V are in the pipeline (Gneiting *et al.*, 2016).

National institute of information and communications technology (NICT)

NICT used 16 units of 4K × 2K-pixel SLMs to project holograms as shown in **Fig. 5**, with an information delivery rate in the order of 10^9 pixels/s (Sasaki *et al.*, 2014a,b). To tile multiple SLMs together seamlessly, a complex optical system is designed. The experimental device was able to produce full-parallax colourful holographic 3D at video rate of 20 fps with an image size of 74 mm × 42 mm and a horizontal viewing-zone angle of 5.6 deg × 2.8 deg. They also developed 8k × 4k LCOS chips and tilted three of them together to deliver 6 billion pixels/s and create a real time holographic 3D colour display (Senoh *et al.*, 2013, 2014).

Agency for science, technology and research (A*STAR)

In comparison with the NICT system, A*STAR system not only used physical tiling but also employed optical scan-tiling of high-speed SLMs to increase the total pixel count of holograms, as shown in **Fig. 6** (Lum *et al.*, 2013). 24 units of high-speed ferroelectric LCOS of 1280 × 1080 pixels in resolution were physically tiled to form a 3 × 8 SLM array. The optical output from this array was tiled in space to form a large image equivalent to that of 36 × 8 SLMs by using a horizontal galvanometric scanning mirror with 12 steps. A large computer-generated hologram for the 3 × 8 SLM array was calculated and divided to upload onto

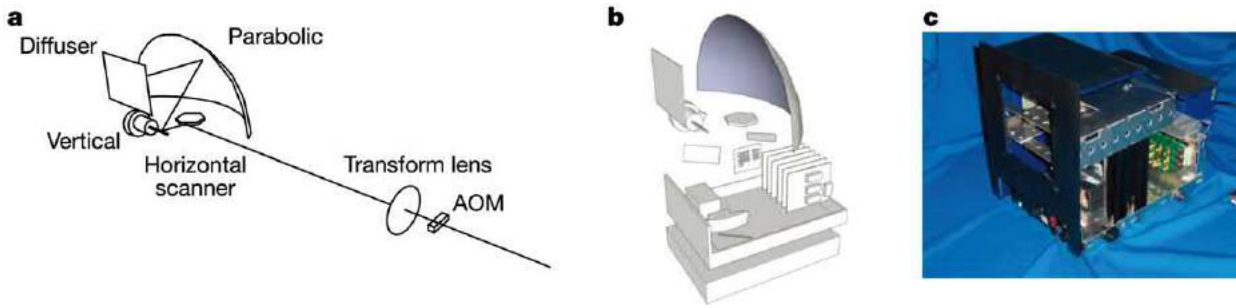


Fig. 4 A holographic video system based on AWMs, (a) optical layout, (b) folding of optical path, and (c) assembled system. Reproduced from Smalley D.E., Smithwick, Q.Y.J., Bove, V.M., Barabas, J., Jolly, S., 2013. Anisotropic leaky-mode modulator for holographic video displays. Nature 498, 313–317.

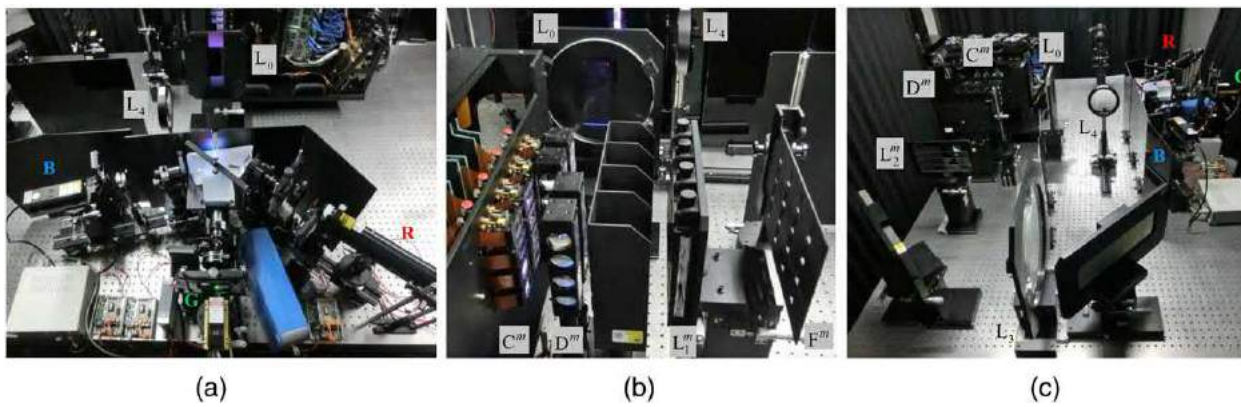


Fig. 5 The NICT (a) light source optical unit, and (b) image reading out optical unit, and (c) an overview of the whole optical system (Sasaki *et al.*, 2014).

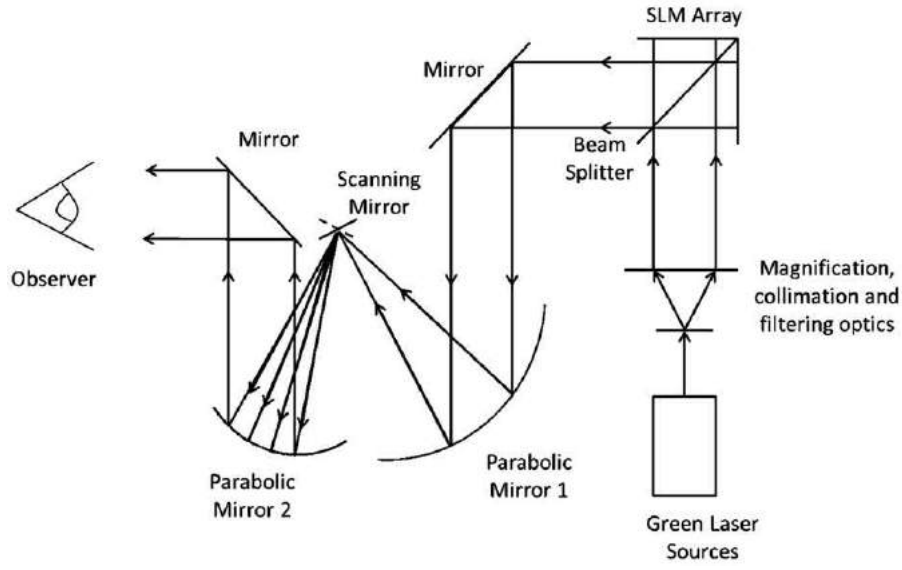


Fig. 6 The optical layout diagram of the A*STAR system. Reproduced from Lum, Z.M.A., Liang, X., Pan, Y., Zheng, R., Xu, X., 2013. Increasing pixel count of holograms for three-dimensional holographic display by optical scan-tiling. Optical Engineering 52, 15802–15802.

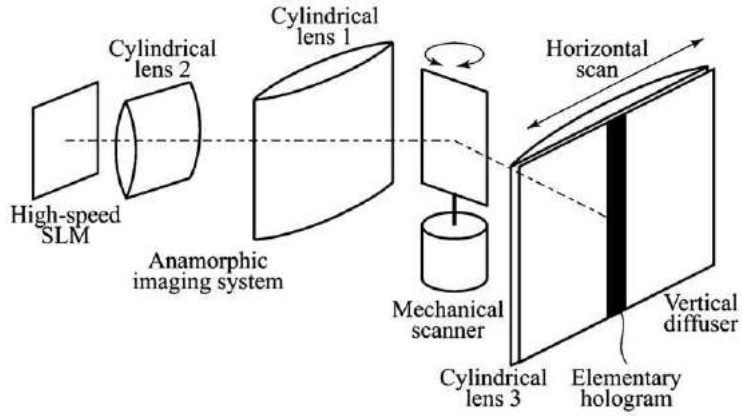


Fig. 7 The TAUT scree-scanning holographic display system. Reproduced from Takaki,Y., Okada, N., 2009. Hologram generation by horizontal scanning of a high-speed spatial light modulator. Appl. Opt. 48, 3255–3260.

individual SLMs (Pan *et al.*, 2013). A full-colour and full-parallax three-dimensional image of 10 in. in width was projected at a rate of 60 frames per second.

Tokyo university of agriculture and technology (TAUT)

TAUT developed a one-dimensional (1D) scanning approach called screen-scanning system, which consists of two orthogonally aligned cylindrical lenses (Takaki *et al.*, 2015; Takaki and Okada, 2010, 2009). The sub-holograms generated from a DMD was de-magnified in the horizontal direction and magnified along the vertical direction. The vertically expanded elementary holograms were then scanned horizontally using a galvanometric scanner onto a vertical diffuser screen to achieve a large screen size. As the pixel pitch is de-magnified horizontally, the horizontal viewing angle is enlarged. The vertical parallax is removed by the use of the vertical diffuser, making this display system one of the "horizontal parallax only" (HPO) systems. An illustration of the system is shown in **Fig. 7** (Takaki and Okada, 2009).

Besides, TAUT also developed the resolution distribution technique (Takaki and Hayashi, 2008) and the viewing-zone scanning system (Takaki and Fujii, 2014). The resolution redistribution applies multiple shifted light sources to re-shape the effective reconstructed window, which is more suitable for one direction scanning. The viewing-zone scanning system employs a magnification image system to increase the hologram in both horizontal and vertical direction. In this case, the viewing angle is reduced. Then the reduced viewing zone is scanned by the horizontal scanner to enlarge the viewing zone, as shown in **Fig. 8** (Takaki and Fujii, 2014).

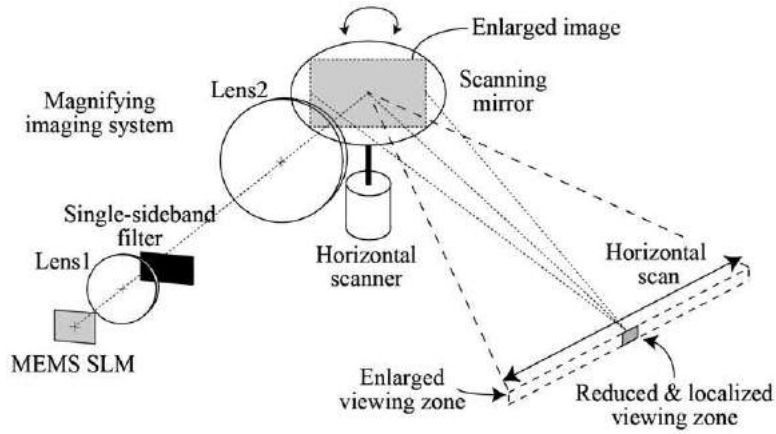


Fig. 8 Viewing zone-scanning holographic display system. Reproduced from Takaki,Y., Fujii,K., 2014. Viewing-zone scanning holographic display using a MEMS spatial light modulator. Opt. Express 22, 24713–24721.

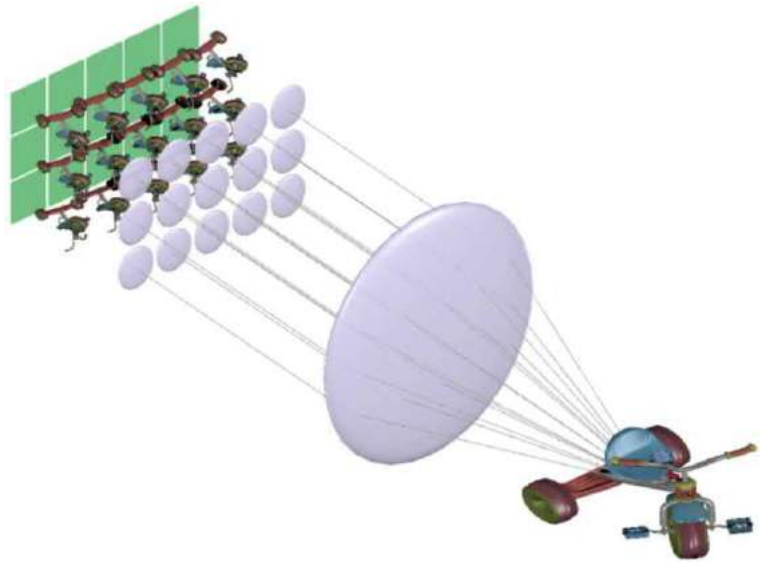


Fig. 9 Course integral holographic display system. Reproduced from Chen, J.S., Smithwick, Q.Y.J., Chu, D.P., 2016. Coarse integral holography approach for real 3D color video displays. Opt. Express 24, 6705–6718.

Course integral holographic display (CIH)

CIH display uses coarse integral optics to angularly tile low spatial bandwidth sub-holograms to form 3D images of a decent size with full parallax features and wide viewing angles in both horizontal and vertical directions, as shown in **Fig. 9** (Chen *et al.*, 2016). The result reported in 2015 (Chen *et al.*, 2016) showed full-colour holographic 3D images consisting of 141.6 Mpixels for each colour with a size of 49.6 mm × 49.6 mm, a frame rate of 23.33 Hz and viewing angles of 12° × 3.2°. Compared with integral imaging and holographic display, CIH has a better performance on producing accommodation depth cue than integral imaging, and larger viewing angles than that of a normal holographic display.

A diffraction-based scanning 3D colour video holographic display

A diffraction-based scanning 3D colour video holographic display was developed employing tiled gratings and a vertical diffuser, as shown in **Fig. 10** (Jia *et al.*, 2017). Its main idea is to use a rotational tiled grating, which is slim and lightweight, to replace a galvanometer or polygon mirror, both of which have a limited scanning speed. It can significantly increase the distributing capability of the scanner, and accommodate the information provided from a high-speed SLM, such as a DMD, without any waste. Its result reported in 2017 (Jia *et al.*, 2017) showed a full colour HPO holographic display projecting 3D images with the size of 60 mm × 30 mm and horizontal viewing angle of 37° at a frame rate of 75 Hz.

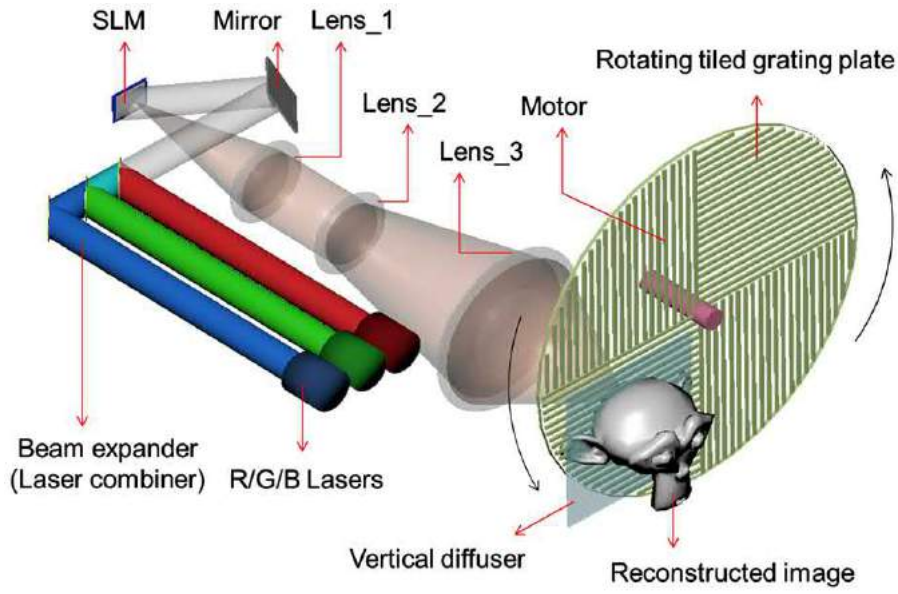


Fig. 10 A diffraction-based scanning 3D colour video holographic display system. Reproduced from Jia, J., Chen, J.S., Yao, J., Chu,D.P., 2017. A scalable diffraction-based scanning 3D colour video display as demonstrated by using tiled gratings and a vertical diffuser. Scientific Reports 7, 44656.

360 degrees holographic display

Apart from the most common front-viewing holographic display systems, there are various systems to produce images with a horizontal viewing angle of 360 degrees (full-around) (Butler *et al.*, 2011; Inoue and Takaki, 2015; Jones *et al.*, 2007; Kakue *et al.*, 2015; Miyazaki *et al.*, 2012; Sando *et al.*, 2016; Takaki and Nakamura, 2014; Takaki and Uchida, 2012; Xia *et al.*, 2013; Yoshida, 2016). This requires a huge amount of data and an information distribution system with high capacity. DMD (s), which can provide a high frame rate, are often used in 360 degrees system for prototype demonstrations. For the information distribution system, there are various approaches, such as the uses of array of projectors (Inoue and Takaki, 2015), tilted mirror (Sando *et al.*, 2016) and parabolic mirror (Butler *et al.*, 2011; Kakue *et al.*, 2015), respectively.

Holography can be integrated with the concept of 360 degrees viewing to support accommodation cue for every view (Inoue and Takaki, 2015; Lim *et al.*, 2016; Teng *et al.*, 2012). A rotating tilted mirror was used to distribute holographic images projected from a high speed SLM (Teng *et al.*, 2012). A parabolic mirror was also used to project a floating holographic image on the table (Lim *et al.*, 2016). The largest diameter of the screen was 35.2 mm among these systems. To support a large image size, a large mirror is necessary but it will increase the weight of the rotation component and slow down the scanning speed. The rotating off-axis lens, which can be made light-weight, was used to distribute holographic images to 360 degrees (Inoue and Takaki, 2015).

To summarise, a comparison of the bandwidth property of the selected holographic displays is shown in **Fig. 11**. The bandwidth improving strategies are spatial tiling, temporal tiling or hybrid tiling methods. No matter which technology is used for a holographic display, the most fundamental principle is to utilize the information bandwidth, in other words, the total amount of information which the device can provide per second to reconstruct holographic images.

Light Sources

Light sources for holographic displays are also developed to improve the quality of holographic images. Holograms, which takes the advantage of interference and diffraction to reconstruct images, require a coherent light source. However, a high level of coherence also produces speckle as a side-effect. To reduce speckle, various approaches are proposed. Some of them are to search a suitable light source (Kozacki and Chlipala, 2016; Mori *et al.*, 2014; Pan *et al.*, 2014; Pang *et al.*, 2015; Zhao *et al.*, 2012), which has an appropriate coherence (low enough to reduce the speckle but high enough to maintain the image quality). Although the image sharpness and speckle are related to both spatial and temporal coherences, studies show that the spatial coherence can be linked directly to the image sharpness and the temporal coherence to the speckle (Deng and Chu, 2017). To provide more natural colour from holograms, lasers with a wide range of wavelengths are researched (Sarakinos *et al.*, 2013) to refine the colour match and the commercial light sources for holography are also developed (Bjelkhagen and Brotherton-Ratcliffe, 2014). Currently, these sources already exist, and the challenge is to reduce its cost (by searching for a cheaper alternative or simplifying its fabrication) and make them affordable.

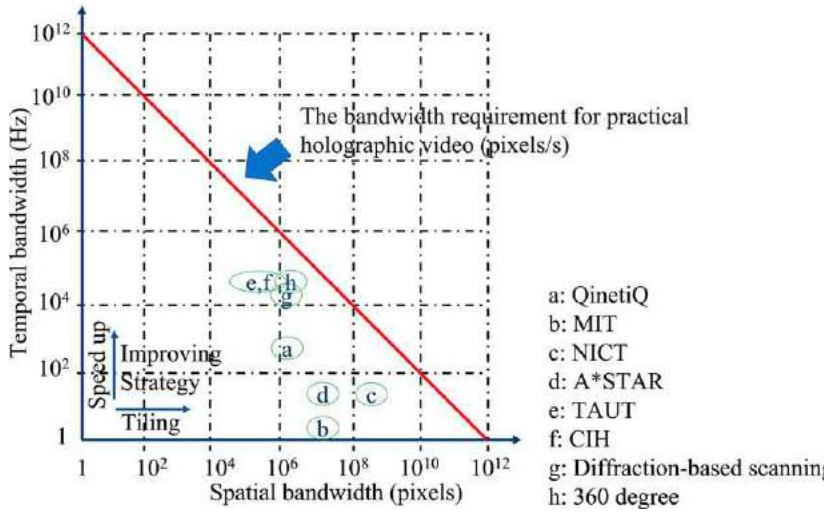


Fig. 11 The bandwidth property of selected holographic display systems.

Computer Generated Holograms

The process of calculating a digital hologram has three steps: (1) define the data of the object to be reconstructed, which can be a virtually created object or the graphic data of a real object; (2) calculate the complex field propagated from the object to the hologram plane according to the wave propagation principles; (3) transfer the calculated hologram onto a film or a SLM, using a phase-only, amplitude-only, binary or any specific format.

In general, the films or SLMs cannot control the amplitude and the phase of a complex field simultaneously, unless using some cascaded approaches mentioned earlier (Hsieh *et al.*, 2007; Jesacher *et al.*, 2008a).

To transfer hologram on films, there are techniques to encode and decode the complex field information of the targeted wave plane: detour holograms and kinoform holograms. The detour-phase hologram, which was invented by Brown and Lohmann, (1969), is a technique transforming the complex information into binary patterns printable by most plotting devices. The principle of making a detour-phase hologram is to divide the hologram into separate cells, each of which contains several pixels. The total number of pixels of the "on" state represents the amplitude and the positions of these pixels encode the phase. Note that although it is named as "phase", detour-phase technique is an amplitude modulation approach. It produces an approximation of the planned modulation and involves with errors, and it has been less used after SLMs being widely used.

The detour-phase technique was later replaced by phase contour interferogram proposed by Lee (1979, 1974). Phase contour interferogram is an encoding approach, which plots the pattern along the hologram phase contour. It can be adapted to plotters to produce a better approximation than that of detour-phase technique. It can also be applied on amplitude-only SLMs which can modulate the intensity in grey levels pixel by pixel.

Unlike pure amplitude modulations, the kinoform hologram supports a pure phase modulation and provides a better efficiency (Lesem *et al.*, 1969). The concept of kinoform holograms is based on an assumption that the phase carries the majority of information about the objects and the loss of amplitude information is acceptable (Oppenheim and Lim, 1981).

Along with the development of SLMs, amplitude-only and phase-only SLMs are commercialized and easily available. The name of detour, contour and kinoform are less used and people simply call amplitude-only or phase-only holograms. Interestingly, while DMD, a popular SLM for holographic displays, is a binary device, it does not apply detour hologram techniques. One reason is that detour-phase hologram has a limited accuracy. Besides, the necessary pixel count per second for a holographic image is high, and it is preferred to keep each pixel as one independent cell, rather than to spend multiple pixels to form an equivalent cell.

Overall, the transformation from a complex hologram to a phase-only, amplitude-only, detour or binary hologram (either phase or amplitude) is only after the complex wave propagation calculation is done. To calculate the complex amplitude and phase, there are mainly two principles. One is to adopt the wave propagation equation on each object element (point, polygon, etc.) and calculate their accumulation on the hologram plane. Another is to simply take the Fourier transformation (FT) of the target 2D image as the hologram, since the reverse FT of this pattern is the target 2D image. This corresponds to the optics fact that two planes on both sides of a lens are FT of each other. Since Fourier hologram is easier to calculate and it becomes the common implementation for 2D image hologram projection. Following we will introduce algorithms about the complex hologram calculation in 2D/2.5D/3D images.

Phase Retrieval Approach for 2D Images

To calculate holograms, we should have both amplitude and phase information of the object light. The intensity of an object is fixed according to the target while the phase of the object light at the object surface can be changed to the arbitrary number since

eyes cannot directly detect the phase of light. Meanwhile, most films and SLMs cannot modulate amplitude and phase of light simultaneously (note that there are few approaches can do this, but they still cannot be widely used). In the phase-only modulation case, the amplitude information of the calculated hologram is lost and this directly degrades the reconstruction quality of the image. Since the phase of object light is flexible, it is possible to find a phase pattern to form a phase-only hologram, which has the least loss of image quality. This action of searching is called "phase retrieval", which is an optimization problem. The fast Fourier transformation (FFT) is the common mathematics tool used in phase retrieval. However, the FFT assumes that a hologram is an infinitely periodic arrangement without considering the aperture effect, hence the pixels in the Fourier plane may be wider than expected and overlap with each other. The consequent noise is on the image plane. Therefore, a multiple iteration method is necessary to minimise this problem.

Gerchberg-Saxton (GS) algorithm is probably the most famous approach for phase retrieval (Gerchberg and Saxton, 1972). The principle of GS algorithm is to calculate the converging solution through iterative FFT between two planes by fixing the amplitude components in one plane (the 2D image plane) and optimize the phase components in the other plane (the hologram plane). Please refers to Gerchberg and Saxton, (1972) for detailed procedure steps.

Different improvements of the GS algorithm have been proposed. The Fienup algorithm (Fienup, 1978) accelerates the converging speed by feeding the errors between the actual reconstruction image and the target image as a feedback into the iteration process. The Fienup with don't-care (Fidoc) method was proposed in 1986 (Akahori, 1986) and the further improved one in 1990 (Wyrowski, 1990). It is a specific case of the Fienup algorithm by applying the don't-care area on the image plane. The don't-care areas may be placed around or within the image, which is well known as zero padding and zero circumscribing. The Fidoc algorithm performs better than GS and Fienup algorithms in terms of image quality but at the cost of increased computational load and number of pixels because of the additional noise disposal area.

In commercialization, a handheld holographic video projector with 60 Hz frame rate and real-time CGHs was developed in CAPE of Cambridge University and demonstrated by Alps Electric in Japan in 2006, as shown in **Fig. 12** (SPIE, 2017).

Holographic Stereogram for 2.5D Images

Between 2D and 3D images, we should also define the 2.5D image. By 2.5D we mean that it provides some but not all depth cues for 3D visual perception. For example, a display using multiple 2D images can provide motion parallax for 3D perception but there is a lack of accommodation cue. Stereogram, which provides 3D perception by binocular cue, is another example of a 2.5D image.

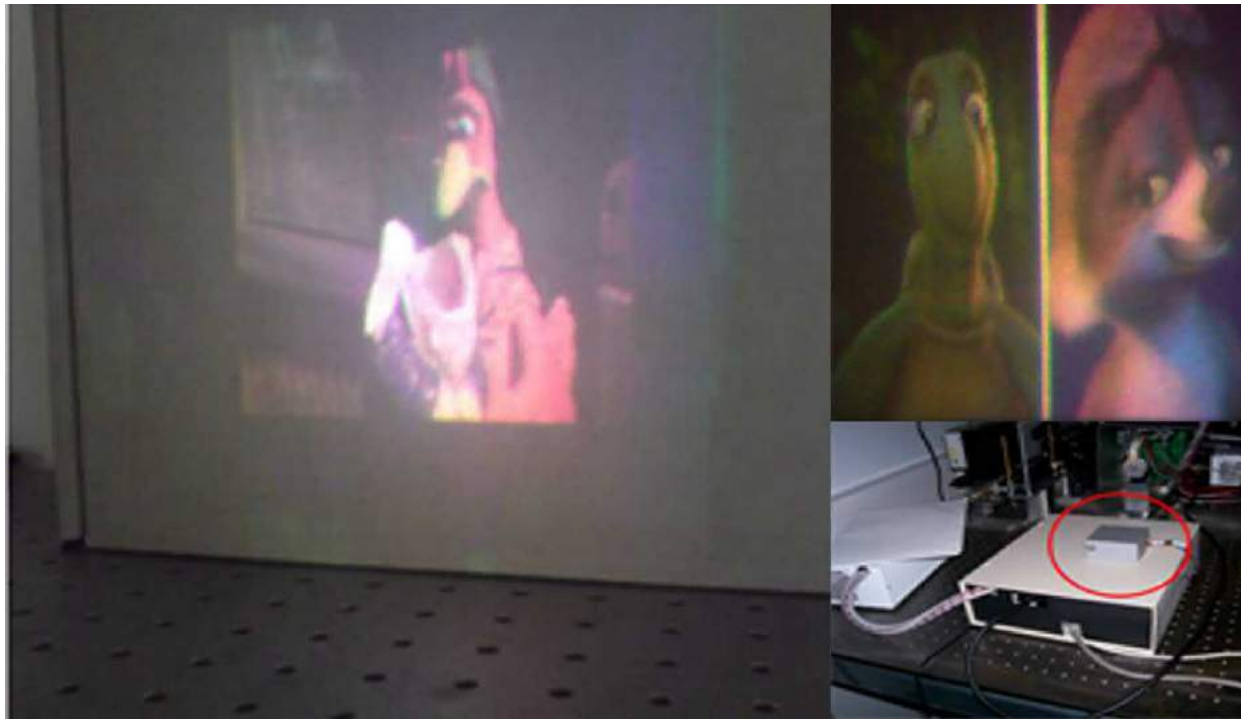


Fig. 12 A handheld holographic video projector with real-time CGHs developed in CAPE of Cambridge University and made by Alps Electric in Japan. Reproduced from SPIE, 2017. Available at: http://spie.org/newsroom/1015-projection-display-using-computer-generated-phase-screens?highlight=x2408&ArticleID=x20068 (accessed 21.05.17).

The concept of multiple 2D images is also used in holographic display, and it is known as holographic stereogram. In early days, it records 2D pictures of the object taken from different angles onto different parts of the hologram by moving a slit step by step (Benton and Bove, 2008). Later on in 1980s, computer calculation of holographic stereogram was also realized. The principle of implementation is to divide the hologram into multiple segments as the basic elements, get the 2D projection with different angles of the 3D object on the hologram plane to get the ray information, transform the ray information (including intensity and angle) to the diffraction pattern on each segment, and then form the final hologram. The way to transform a ray into a diffraction pattern is to conduct Fourier transformation (FT) on a 2D image representing the ray direction and intensity in the image domain. Since the diffraction calculation can be done by FFT, the calculation load of computer holographic stereograms is much less than that of the conventional CGH techniques for 3D images (which will be introduced in Section 3D Images (Point/Polygon/Layer Based Methods and Improvements).

Another advantage of holographic stereogram is that it can produce occlusion effect directly because of its use of 2D projection of the 3D object, and it can be integrated with 3D graphics without much additional computation (Barabas *et al.*, 2011; Sando *et al.*, 2013), while occlusion culling methods for conventional 3D image hologram calculation involves with heavy computation (Ichikawa *et al.*, 2013a,b; Matsushima and Kondoh, 2004; Matsushima and Nakahara, 2009; Underkoffler, 1997; Zhang *et al.*, 2011).

The concept of using individual rays is an approximation to the 3D wave propagation. To improve this approximation, phase-added stereogram was proposed (Gao *et al.*, 2015; Kang *et al.*, 2016; Yamaguchi *et al.*, 1993). It calculates the directional information of the light from the object and uses a phase factor to yield a better coherent wavefront. As a result, the depth range of the image is wider while the amount of calculation stays almost the same. In commercialization, Geola and Zebra Imaging print customized 3D holographic stereogram holograms.

3D Images (Point/ Polygon/Layer Based Methods and Improvements)

To calculate the complex wave plane of one object, the information of location and intensity should be considered. The object data can come from a virtual object or the graphics data reconstructed from a real object by a 3D camera. To simplify the calculation procedure, a 3D object is broken down to smaller basic units and the wave propagation of each unit is calculated and added up to form the final hologram.

According to the type of the basic unit in use, 3D image CGH methods can be classified into three categories. One is the point-source method, which samples 3D objects by cloud points (Eldeib and Yabe, 1996; Stein *et al.*, 1992). Another is the polygon-based method (Matsushima, 2010, 2008; Matsushima and Shimobaba, 2009), which discretizes 3D objects into a number of polygons, usually triangles. The third one is image-based approach (Barabas *et al.*, 2011; Chen and Chu, 2015; Jia *et al.*, 2014) which hybrids graphic rendered 2D images with other approaches to generate 3D effects.

The point-based method is also known as the coherent ray tracing (CRT) method. It treats every point on the object as a self-illuminating point source and calculates the emitted spherical wave to all the pixels on a hologram plane using a light propagation function (Matsushima and Takai, 2000; Nishitsuji *et al.*, 2015; Weng *et al.*, 2011). To support a high quality 3D image, the object needs to be sampled with a large number of points and this leads to a high computational load.

To improve the calculation speed, Lucente proposed the look-up-table (LUT) method in 1993 (Lucente, 1993) to store all the possible diffraction patterns for the points offline. The point diffraction patterns are directly read out from the table during the online calculation. However, the necessary size of the LUT is huge, such as hundreds of gigabytes, and it is overwhelming even for a modern computer. Several improved LUT methods (Gao *et al.*, 2015; Dong *et al.*, 2014; Jia *et al.*, 2013; Jiao *et al.*, 2017; Kim and Kim, 2008; Kim *et al.*, 2013, 2012, 2008, 2015; Kwon *et al.*, 2016; Pan *et al.*, 2009) were proposed to reduce the memory size down to the order of kilobytes.

Another method to improve the calculation speed of point-based methods is the wave recording plane (WRP) technique developed by Phan *et al.* (2014) and Shimobaba *et al.* (2009, 2010). A WRP is a virtual plane that is located at a near distance to the objects. Because the short distance and the limited diffraction angle, each object point will only diffract light within a small area onto the WRP. Then the information on the WRP propagates to the hologram plane by Fresnel diffraction. With the WRP technique, a hologram of 2048×2048 pixels in size, representing an image scene composed of 4×10^6 points, can be generated in less than 25 ms (Tsang *et al.*, 2011).

On the other hand, in the polygon-based method, the 3D object is represented by thousands of polygons rather than millions of points. It essentially reduces the number of basic units while the computation for each basic unit becomes more complicated. The computation of a polygon unit involves a rotational transformation in the frequency domain to get the pattern corresponding to the normal angle, and the angular spectrum approach which can calculate the complex information along the propagation axis (Matsushima, 2008). In 2008, a fully analytical method was developed (Ahrenberg *et al.*, 2008). It proposes a new structure to conduct polygon based methods (FFT is no longer used). The method was implemented on GPU for parallel computing in order to achieve a faster speed and better image quality (Liu *et al.*, 2010). When polygon-approach was developed, the overall calculation speed was expected to be faster than that of other approaches. However, point-based and image-based methods were also improved, and now polygon-based approach remains the slowest approach.

The image-based approach was firstly used in holographic stereogram to provide multiple 2D images for different views. This approach, later on, is applied to the calculation of real 3D images. A hybrid of point-based approach and image-based approach was proposed (Barabas *et al.*, 2011; Jia *et al.*, 2014), so that the occlusion cue is provided by graphics while the accommodation cue

is provided by point-based method. The layer-based method was also proposed (Chen and Chu, 2015; Chen *et al.*, 2014). It uses 2D rendered depth map to divide a 3D image into several layers. The hologram of each 2D image is calculated, by adding a lens of a specific focal length to the FT of that layer. Finally, holograms of individual layers are added together to form the hologram of the 3D image. It takes advantage of the factor that eyes have a limited resolution of accommodation cue, so few layers are enough to project a good accommodation cue. Research shows that only 28 depth layers are required to produce a full depth of field from 25 cm to infinite distance for human eyes (Akeley, 2004; Rolland *et al.*, 1999).

Error Reduction in CGH

Existing SLMs, such as LCOS, can only display either the amplitude or the phase component of light. The loss of either amplitude or phase information brought up the phase retrieval issue, as discussed in Section Phase Retrieval Approach for 2D Images. Different phase retrieval approaches, such as GS and Fienup methods, optimize the reconstruction image quality by using an iterative calculation. In contrast, the loss of either amplitude or phase information may lead to heavy errors in the holograms calculated by using wavefront propagation functions discussed in 3D Images (Point/Polygon/Layer Based Methods and Improvements). The random phase was first added to the object points prior to the generation of the digital hologram to minimise this error, but the random phase makes the speckle noise worse (Pan *et al.*, 2014; Makowski *et al.*, 2016; Shimobaba *et al.*, 2016,a,b; Takaki and Taira, 2016; Takaki and Yokouchi, 2011).

One solution to this problem is to cascade two SLMs and reconstruct the complex amplitude hologram (Hsieh *et al.*, 2007; Leportier *et al.*, 2016). In some implementations, an optical element (can be a prism or grating) is used to merge the wave propagations from different regions of the same device in order to achieve the complex modulation (Jesacher *et al.*, 2008a,b; Liu *et al.*, 2011; Song *et al.*, 2012). Another effective approach is to place a moving diffuser at the imaging plane (Kuratomi *et al.*, 2010; Pan and Shih, 2014). It can average the noise, but requires an extra mechanical device.

To deal with this problem through calculation, the error reduction methods are proposed which can be grouped into temporal-based and spatial-based methods. For the temporal-based method, multiple sub-frames, each of which represents the same object scene added with different random phase patterns, are displayed rapidly in sequence to the observers to average the speckle noise (Buckley, 2011; Buckley *et al.*, 2006). It is also known as the random averaging method. The superposition of N holograms with uncorrelated phase patterns can reduce the speckle noise to $1/\sqrt{N}$ at the cost of computational load and the reduction of effective frame rate.

The spatial-based error reduction method is also known as the error diffusion method (Tsang and Poon, 2013; Tsang *et al.*, 2014). It converts a digital complex hologram into a phase-only hologram without using a random phase pattern. The process is accomplished with an error diffusion mechanism. It scans each hologram pixel sequentially, updates it with the values it was past to, then pass the error to the unvisited neighborhood, and then carry on. Compared with the temporal-based method, this method does not consume the frame rate of the SLM, and the reconstructed images are very similar to those obtained with the original complex holograms. The downside of spatial-based error distribution is its additional calculations. It is noted that the error diffusion methods can be applied directly on an existing complex amplitude hologram, without knowing the original data of the object image. This allows it to be applied on holograms calculated by any approach as a post-processing approach.

Holographic Display VR/AR

Virtual reality (VR) has been a technology dream for decades. It is now gradually commercialized after the launches of Oculus Rift, HTC Vive and Sony PlayStation VR between 2015 and 2016. At the same time augmented reality (AR) is also attracting public attention. One of the representative examples is Microsoft Hololens, released in 2016. It successfully applies a holographic waveguide combiner to be part of the AR headset while the image contents stay as 2D. On the other hand, holography can be used as true 3D image projectors and it attracts efforts on building a prototype of holographic HMD. This section introduces these two holographic applications in VR/AR.

Holographic Waveguide/Lightguide Combiner in VR/AR HMD

In early 1990s, the concept of a holographic waveguide display was patented in which a waveguide hologram was used to couple a collimated image into a glass-based waveguide, and transmit the images by Total Internal Reflection (TIR). The spectral and angular Bragg selectivity is the key functionality in the holographic combiners both in reflective or transmission holograms. Compared with a transmission hologram, a reflective hologram can achieve a narrower spectral bandwidth and a wider angular bandwidth at the same time. A larger angular bandwidth can produce a larger field of view, and narrower spectral bandwidth can reduce the chromatic aberration.

Since the waveguide can be made very thin and light, there are companies such as Sony Ltd, Konica/Minolta and Nokia/Vuzix which have applied this concept to the HMDs. Their works mainly focused on improving the light efficiency and reducing the chromatic aberration in waveguide due to the wavelength dependency of holograms.

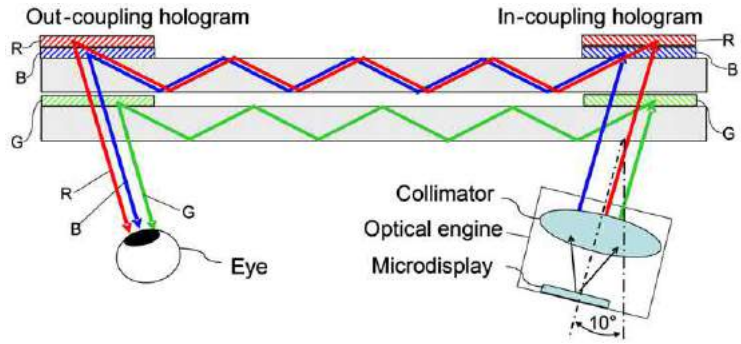


Fig. 13 Sony holographic waveguide combiner. Reproduced from Mukawa, H., Akutsu, K., Matsumura, S., 2008.8.4: distinguished paper: A full color eyewear display using holographic planar waveguides. In: SID Symposium Digest of Technical Papers 39, pp. 89–92.

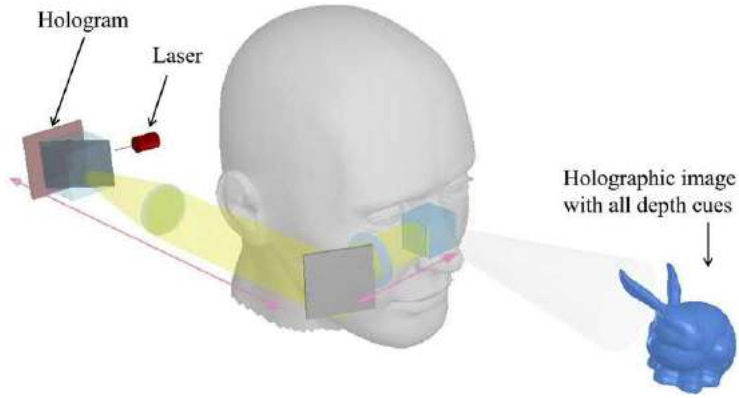


Fig. 14 A holographic 3D head-mounted display. Reproduced from Chen,J.-S., Chu, D.P., 2015. Improved layer-based method for rapid hologram generation and real-time interactive holographic display applications. Opt. Express 23, 18143–18155.

In Sony and Konica/Minolta's holographic waveguide combiners, the traditional reflective holograms are used with high diffraction efficiency because of the Bragg selectivity. The chromatic aberration in waveguide can be compensated by using three superimposed holograms, such as in Sony's approach shown in **Fig. 13** (Mukawa *et al.*, 2008). Each hologram is exposed to a specific wavelength. Otherwise, Konica/Minolta and other researchers exposed one hologram three times to different wavelengths with the corresponding illuminating angle (Kress and Starner, 2013; Shi *et al.*, 2012). One drawback of using holographic waveguides is its low diffraction efficiency due to the material limitation, and its use in HMDs is replaced by free space (Zhou *et al.*, 2017), free-form optics (Hu and Hua, 2014; Hua and Javidi, 2014) or prisms (Love *et al.*, 2009).

Holographic 3D Display in VR/AR HMD

Conventional stereoscopic HMDs may cause eye fatigue because of the conflict of accommodation and convergence depth cues. Applications of holographic 3D displays in HMD is one solution to provide accommodation cue and remove the fatigue.

A number of prototypes have been built (Chen and Chu, 2015; Li *et al.*, 2016; Tanjung *et al.*, 2010; Yoneyama *et al.*, 2014, 2013). The concept of a holographic 3D HMD is shown in **Fig. 14** (Chen and Chu, 2015), in which the holographic 3D images generated by the SLM are projected into human pupil directly. The necessary bandwidth (both optical and computational) for a holographic HMD is much less than that of a holographic display because of the small size of a pupil, therefore holographic HMDs are much more feasible in terms of bandwidth. Nevertheless, there are still issues remain to be solved, including speckle noise, small field of view (FOV), chromatic aberration and small eye-box. This field is still relatively new, and its development will take time before any solid impact is produced.

Key Challenges and Conclusions

Although huge efforts and funding have been put into the development of digital holographic displays for several decades, the process of its commercialization is still very slow. The key challenge is that the user experience, including the image size, viewing angle and image quality, is not as good as that of traditional 2D displays. In comparison with 2D displays, a high quality 3D

display requires a pixel size smaller than 1 μm and information bandwidth larger than 10^{12} pixels/s for a decent size display screen. This is difficult to achieve by the existing technology. Such a huge amount of data also leads to the heavy computational load for rendering the 3D data and generate CGHs. Moreover, it is difficult to realize the real-time display because of the limited data transmission bandwidth of the data bus. For example, a DMD can run at a high frame rate but pre-uploading of the images is required, which can take a couple of minutes or even more.

To move holographic 3D display technologies forward into a mass market is not easy. We need to develop suitable technologies and products to overcome all the obstacles. Fortunately, future SLMs have the potential to be made with sub-micron pitches (Shrestha *et al.*, 2015; Isomae *et al.*, 2016) and to be driven at a fast speed (Forth Dimensions Displays, 2017), in order to produce the required amount of data. High speed computers and fast CGH calculation algorithms make it possible to generate holograms in real time (Ichihashi *et al.*, 2012), and customized data board for transmitting holograms may solve the data bus issue.

Finally, to reconstruct only the wavefront at the eye position using eye-tracking technology can significantly reduce the amount of information to be processed. Similarly, applying digital holographic 3D technology for HMDs requires much less amount of information if compared with traditional holographic 3D displays.

See also: Module: Digital Holography

References

Ahrenberg, L., Benzie, P., Magnor, M., Watson, J., 2008. Computer generated holograms from three dimensional meshes using an analytic light transport model. Appl. Opt. 47, 1567–1574.

Akahori, H., 1986. Spectrum leveling by an iterative algorithm with a dummy area for synthesizing the kinoform. Appl. Opt. 25, 802–811.

Akeley, K., 2004. Achieving Near-Correct Focus Cues Using Multiple Image Planes. PhD Thesis. Stanford University.

Barabas, J., Jolly, S., Smalley, D.E., Bove, V.M., 2011. Diffraction specific coherent panoramagrams of real scenes. Proc. SPIE 7957, 795702.

Benton, S.A., Bove, J.V.M., 2008. Holographic Imaging. Wiley-Interscience.

Berneth, H., Bruder, F.-K., Fäcke, T., *et al.*, 2013. Holographic recordings high beam ratios Improved Bayfol®HX photopolymer. Proc. SPIE 8776, 877603.

Bjelkhagen, H.I., Brotherton-Ratcliffe, D., 2014. Ultrarealistic imaging the future of display holography. Optical Engineering 53, 112310.

Bjelkhagen, H.I., Crosby, P.G., Green, D.P.M., Mirlis, E., Phillips, N.J., 2008. Fabrication ultra-fine-grain silver halide recording material color holography. Proc. SPIE 6912, 691209.

Blanche, P.-A., Bablumian, A., Voorakaranam, R., 2010a. Future photorefractive based holographic 3D display. Proc. SPIE 7619, 76190L.

Blanche, P.-A., Bablumian, A., Voorakaranam, R., *et al.*, 2010b. Holographic three-dimensional telepresence using large-area photorefractive polymer. Nature 468, 80–83.

Brown, B.R., Lohmann, A.W., 1969. Computer-generated binary holograms. IBM J. Res. Dev. 13, 160–168.

Buckley, E., 2011. Real-time error diffusion for signal-to-noise ratio improvement in a holographic projection system. Journal of Display Technology 7, 70–76.

Buckley, E., Cable, A., Lawrence, N., Wilkinson, T., 2006. Viewing angle enhancement for two- and three-dimensional holographic displays with random superresolution phase masks. Appl. Opt. 45, 7334–7341.

Butler, A., Hilliges, O., Izadi, S., *et al.*, 2011. Vermeer direct interaction with a 360° viewable 3D display. In: Proceedings of the 24th Annual ACM Symposium User Interface Software Technology, UIST'11 (ACM, 2011), pp. 569–576.

Chen, J., Morris, S.M., Wilkinson, T.D., Freeman, J.P., Coles, H.J., 2009. High speed liquid crystal over silicon display based on the flexoelectro-optic effect. Opt. Express 17, 7130–7137.

Chen, J.-S., Chu, D., Smithwick, Q., 2014. Rapid hologram generation utilizing layer-based approach and graphic rendering for realistic three-dimensional image reconstruction by angular tiling. Journal of Electronic Imaging 23, 023016.

Chen, J.-S., Chu, D.P., 2015. Improved layer-based method for rapid hologram generation and real-time interactive holographic display applications. Opt. Express 23, 18143–18155.

Chen, J.S., Smithwick, Q.Y.J., Chu, D.P., 2016. Coarse integral holography approach for real 3D color video displays. Opt. Express 24, 6705–6718.

Cody, D., Naydenova, I., Mihaylova, E., 2012. New non-toxic holographic photopolymer material. Journal of Optics 14, 015601.

Collings, N., Davey, T., Christmas, J., Chu, D.P., Crossland, B., 2011. The applications and technology of phase-only liquid crystal on silicon devices. Journal of Display Technology 7, 112–119.

Deng, Y., Chu, D., 2017. Coherence properties of different light sources and their effect on the image sharpness and speckle of holographic display. Sci. Rep. 7 (1), 5893.

Denisyuk, Y.N., 1962. On the reflection of optical properties of an object in a wave field of light scattered by it. Doklady Akademii Nauk SSSR 144, 1275–1278.

Dennis, G., 1948. A new microscopic principle. Nature 161, 777–778.

Dong, X.-B., Kim, S.-C., Kim, E.-S., 2014. MPEG-based novel look-up table for rapid generation of video holograms of fast-moving three-dimensional objects. Opt. Express 22, 8047–8067.

Eldeib, H., Yabe, T., 1996. Computer-generated holograms: A fast ray-tracing approach. Journal of King Saud University - Computer and Information Sciences 8, 139–151.

Fienup, J.R., 1978. Reconstruction of an object from the modulus of its Fourier transform. Opt. Lett. 3, 27–29.

Forth Dimensions Displays, 2017. Available at: http://www.forthdd.com/ (accessed 21.05.17).

Gambogi Jr, W.J., Weber, A.M., Trout, T.J., 1994. Advances and applications of DuPont holographic photopolymers. Proc. SPIE 2043, 2.

Gao, C., Liu, J., Li, X., *et al.*, 2015. Accurate compressed look up table method for CGH in 3D holographic display. Opt. Express 23, 33194–33204.

Gerchberg, R.W., Saxton, W.O., 1972. A practical algorithm for the determination of the phase from image and diffraction plane pictures. Optik ((Jena)) 35, 237.

Gneiting, S., Smalley, D.E., Qaderi, K., *et al.*, 2016. Optimizations for robust, high-efficiency, waveguide-based holographic video. In: 2016 IEEE 14th International Conference Industrial Informatics (INDIN), pp. 576–581.

Hecht, E., 1998. Optics. Addison-Wesley.

Hsieh, M.-L., Chen, M.-L., Cheng, C.-J., 2007. Improvement of the complex modulated characteristic of cascaded liquid crystal spatial light modulators by using a novel amplitude compensated technique. Optical Engineering 46.070501–070501–3.

Hu, X., Hua, H., 2014. High-resolution optical see-through multi-focal-plane head-mounted display using freeform optics. Opt. Express 22, 13896–13903.

Hua, H., Javidi, B., 2014. A 3D integral imaging optical see through head-mounted display. Opt. Express 22, 13484–13491.

Ichihashi, Y., Oi, R., Senoh, T., Yamamoto, K., Kurita, T., 2012. Real-time capture and reconstruction system with multiple GPUs for a 3D live scene by a generation from 4K IP images to 8K holograms. Opt. Express 20, 21645–21655.

Ichikawa, T., Yamaguchi, K., Sakamoto, Y., 2013a. Realistic expression for full-parallax computer-generated holograms with the ray-tracing method. Appl. Opt. 52, A201–A209.
Ichikawa, T., Yoneyama, T., Sakamoto, Y., 2013b. CGH calculation with the ray tracing method for the Fourier transform optical system. Opt. Express 21, 32019–32031.
Inoue, T., Takaki, Y., 2015. Table screen 360-degree holographic display using circular viewing-zone scanning. Opt. Express 23, 6533–6542.
Isomae, Y., Shibata, Y., Ishinabe, T., Fujikake, H., 2016. Optical phase modulation properties of 1 μm-pitch LCOS with dielectric walls for wide-viewing-angle holographic displays. In: SID Symposium Digest of Technical Papers 47, 1670–1673.
Jasper Display Corp., 2017. Available at: http://www.jasperdisplay.com/success-stories/ (accessed 21.05.17).
Jesacher, A., Maurer, C., Schwaighofer, A., Bernet, S., Ritsch-Marte, M., 2008a. Near perfect hologram reconstruction with a spatial light modulator. Opt. Express 16, 2597–2603.
Jesacher, A., Maurer, C., Schwaighofer, A., Bernet, S., Ritsch-Marte M., 2008b. Full phase and amplitude control of holographic optical tweezers with high efficiency. Opt. Express 16, 4479–4486.
Jia, J., Chen, J.S., Yao, J., Chu, D.P., 2017. A scalable diffraction-based scanning 3D colour video display as demonstrated by using tiled gratings and a vertical diffuser. Scientific Reports 7, 44656.
Jia, J., Liu, J., Jin, G., Wang, Y., 2014. Fast and effective occlusion culling for 3D holographic displays by inverse orthographic projection with low angular sampling. Appl. Opt. 53, 6287–6293.
Jia, J., Wang, Y., Liu, J., *et al*., 2013. Reducing the memory usage for effective computer-generated hologram calculation using compressed look-up table in full-color holographic display. Appl. Opt. 52, 1404–1412.
Jiao, S., Zhuang, Z., Zou, W., 2017. Fast computer generated hologram calculation with a mini look-up table incorporated with radial symmetric interpolation. Opt. Express 25, 112–123.
Jones, A., McDowall, I., Yamada, H., Bolas, M., Debevec, P., 2007. Rendering for an interactive 360° light field display. ACM Trans. Graph 26.
Jurbergs, D., Bruder, F.-K., Deuber, F., *et al*., 2009. New recording materials for the holographic industry. Proc. SPIE 7233, 72330K.
Kakue, T., Nishitsuji, T., Kawashima, T., *et al*., 2015. Aerial projection of three-dimensional motion pictures by electro-holography and parabolic mirrors. Scientific Reports 5, 11750.
Kang, H., Stoykova, E., Yoshikawa, H., 2016. Fast phase-added stereogram algorithm for generation of photorealistic 3D content. Appl. Opt. 55, A135–A143.
Kim, S.-C., Dong, X.-B., Kim, E.-S., 2015. Accelerated one-step generation of full-color holographic videos using a color-tunable novel-look-up-table method for holographic three-dimensional television broadcasting. Scientific Reports 5, 14056.
Kim, S.-C., Dong, X.-B., Kwon, M.-W., Kim, E.-S., 2013. Fast generation of video holograms of three-dimensional moving objects using a motion compensation-based novel look-up table. Opt. Express 21, 11568–11584.
Kim, S.-C., Kim, E.-S., 2008. Effective generation of digital holograms of three-dimensional objects using a novel look-up table method. Appl. Opt. 47, D55–D62.
Kim, S.-C., Kim, J.-M., Kim, E.-S., 2012. Effective memory reduction of the novel look-up table with one-dimensional sub-principle fringe patterns in computer-generated holograms. Opt. Express 20, 12021–12034.
Kim, S.-C., Yoon, J.-H., Kim, E.-S., 2008. Fast generation of three-dimensional video holograms by combined use of data compression and lookup table techniques. Appl. Opt. 47, 5986–5995.
Kinashi, K., Wang,Y., Tsujimura,S., Sakai,W., Tsutsumi, N., 2012. Dynamic holographic images using photorefractive composites.In: Biomedical Optics 3-D Imaging. Optical Society of America, p. JM3A.58.
Kozacki, T., Chlipala, M., 2016. Color holographic display with white light LED source and single phase only SLM. Opt. Express 24, 2189–2199.
Kress, B., Starner, T., 2013. A review of head-mounted displays technologies and applications for consumer electronics. Proc. SPIE 8720, 87200A.
Kuratomi, Y., Sekiya, K., Satoh, H., *et al*., 2010. Speckle reduction mechanism in laser rear projection displays using a small moving diffuser. J. Opt. Soc. Am. A 27, 1812–1817.
Kwon, M.-W., Kim, S.-C., Kim, E.-S., 2016. Three-directional motion-compensation mask-based novel look-up table on graphics processing units for video-rate generation of digital holographic videos of three-dimensional scenes. Appl. Opt. 55, A22–A31.
Lee, W.-H., 1974. Binary synthetic holograms. Appl. Opt. 13, 1677–1682.
Lee, W.-H., 1979. Binary computer-generated holograms. Appl. Opt. 18, 3661–3669.
Leith, E.N., Upatnieks, J., 1962. Reconstructed wavefronts and communication theory. J. Opt. Soc. Am. 52, 1123–1130.
Leportier, T., Park, M.C., Kim, T., 2016. Numerical alignment of spatial light modulators for complex modulation in holographic displays. Journal of Display Technology 12, 1000–1007.
Lesem, L.B., Hirsch, P.M., Jordan, J.A., 1969. The kinoform: A new wavefront reconstruction device. IBM Journal of Research and Development 13, 150–155.
Li, G., Lee, D., Jeong, Y., Cho, J., Lee, B., 2016. Holographic display for see-through augmented reality using mirror-lens holographic optical element. Opt. Lett. 41, 2486–2489.
Lim, Y., Hong, K., Kim, H., *et al*., 2016. 360-degree tabletop electronic holographic display. Opt. Express 24, 24999–25009.
Liu, J.-P., Hsieh, W.-Y., Poon, T.-C., Tsang, P., 2011. Complex Fresnel hologram display using a single SLM. Appl. Opt. 50, H128–H135.
Liu, Y.-Z., Dong, J.-W., Pu, Y.-Y., *et al*., 2010. High-speed full analytical holographic computations for true-life scenes. Opt. Express 18, 3345–3351.
Love, G.D., Hoffman, D.M., Hands, P.J.W., *et al*., 2009. High-speed switchable lens enables the development of a volumetric stereoscopic display. Opt. Express 17, 15716–15725.
Lucente, M., 1994. Diffraction-specific fringe computation for electro-holography. PhD, Massachusetts Institute of Technology.
Lucente, M.E., 1993. Interactive computation of holograms using a look-up table. Journal of Electronic Imaging 2, 28–34.
Lum, Z.M.A., Liang, X., Pan, Y., Zheng, R., Xu, X., 2013. Increasing pixel count of holograms for three-dimensional holographic display by optical scan-tiling. Optical Engineering 52, 15802.
Makowski, M., Shimobaba, T., Ito, T., 2016. Increased depth of focus in random-phase-free holographic projection. Chin. Opt. Lett. 14, 120901.
Matsushima, K., 2008. Formulation of the rotational transformation of wave fields and their application to digital holography. Appl. Opt. 47, D110–D116.
Matsushima, K., 2010. Wave-field rendering in computational holography: The polygon-based method for full-parallax high-definition CGHs. In: 2010 IEEE/ACIS Proceedings of the 9th International Conference Computer Information Science, pp. 846–851.
Matsushima, K., Kondoh, A., 2004. A wave-optical algorithm for hidden-surface removal in digitally synthetic full-parallax holograms for three-dimensional objects. Proc. SPIE 5290, 90.
Matsushima, K., Nakahara, S., 2009. Extremely high-definition full-parallax computer-generated hologram created by the polygon-based method. Appl. Opt. 48, H54–H63.
Matsushima, K., Shimobaba, T., 2009. Band-limited angular spectrum method for numerical simulation of free-space propagation in far and near fields. Opt. Express 17, 19662–19673.
Matsushima, K., Takai, M., 2000. Recurrence formulas for fast creation of synthetic three-dimensional holograms. Appl. Opt. 39, 6587–6594.
Miyazaki, D., Akasaka, N., Okoda, K., Maeda, Y., Mukai, T., 2012. Floating three-dimensional display viewable from 360 degrees. Proc. SPIE 8288, 82881H.
Mori, Y., Fukuoka, T., Nomura, T., 2014. Speckle reduction in holographic projection by random pixel separation with time multiplexing. Appl. Opt. 53, 8182–8188.
Mukawa, H., Akutsu, K., Matsumura, I., *et al*., 2008. 8.4: Distinguished paper: A full color eyewear display using holographic planar waveguides. In: SID Symposium Digest of Technical Papers 39, pp. 89–92.
Nishitsuji, T., Shimobaba, T., Kakue, T., Arai, D., Ito, T., 2015. Simple and fast cosine approximation method for computer-generated hologram calculation. Opt. Express 23, 32465–32470.

Oppenheim, A.V., Lim, J.S., 1981. The importance of phase in signals. Proceedings of the IEEE 69, 529–541.
Pan, J.-W., Shih, C.-H., 2014. Speckle reduction and maintaining contrast in a LASER pico projector using a vibrating symmetric diffuser. Opt. Express 22, 6464–6477.
Pan, Y., Xu, X., Liang, X., *et al.*, 2013. Large-pixel-count hologram data processing for holographic 3D display. Proc. SPIE 8644, 86440F.
Pan, Y., Wang, Y., Liu, J., Li, X., Jia, J., 2014. Improved full analytical polygon-based method using Fourier analysis of the three-dimensional affine transformation. Appl. Opt. 53, 1354–1362.
Pan, Y., Xu, X., Solanki, S., *et al.*, 2009. Fast CGH computation using S-LUT on GPU. Opt. Express 17, 18543–18555.
Pang, X.-N., Chen, D.-C., Ding, Y.-C., *et al.*, 2015. Image quality improvement of polygon computer generated holography. Opt. Express 23, 19066–19073.
Phan, A.-H., Piao, M., Gil, S.-K., Kim, N., 2014. Generation speed and reconstructed image quality enhancement of a long-depth object using double wavefront recording planes and a GPU. Appl. Opt. 53, 4817–4824.
Rasmussen, P.H., Ramanujam, P.S., Hvilsted, S., Berg, R.H., 1999. A remarkably efficient azobenzene peptide for holographic information storage. Journal of the American Chemical Society 121, 4738–4743.
Rolland, J.P., Krueger, M.W., Goon, A.A., 1999. Dynamic focusing in head-mounted displays. Proc. SPIE 3639, 463–470.
Sando, Y., Barada, D., Yatagai, T., 2013. Hidden surface removal of computer-generated holograms for arbitrary diffraction directions. Appl. Opt. 52, 4871–4876.
Sando, Y., Barada, D., Yatagai, T., 2016. Optical rotation compensation for a holographic 3D display with a 360 degree horizontal viewing zone. Appl. Opt. 55, 8589–8595.
Sarakinos, A., Zervos, N., Lembessis, A., 2013. Holofos: An optimized LED illumination system for color reflection holograms display. Proc. SPIE 8644, 86440I.
Sasaki, H., Yamamoto, K., Ichihashi, Y., Senoh, T., 2014a. Image size scalable full-parallax coloured three-dimensional video by electronic holography. Scientific Reports 4, 4000.
Sasaki, H., Yamamoto, K., Wakunami, K., *et al.*, 2014b. Large size three-dimensional video by electronic holography using multiple spatial light modulators. Scientific Reports 4, 6177.
Senoh, T., Ichihashi, Y., Oi, R., Sasaki, H., Yamamoto, K., 2013. Study of a holographic TV system based on multi-view images and depth maps. Proc. SPIE 8644, 86440A.
Senoh, T., Wakunami, K., Ichihashi, Y., *et al.*, 2014. Multiview image and depth map coding for holographic TV system. Optical Engineering 53, 112302.
Shi, R., Liu, J., Zhao, H., *et al.*, 2012. Chromatic dispersion correction in planar waveguide using one-layer volume holograms based on three-step exposure. Appl. Opt. 51, 4703–4708.
Shimobaba, T., Kakue, T., Ito, T., 2016. Random phase-free computer holography and its applications. Proc. SPIE 9867, 98670M.
Shimobaba, T., Makowski, M., Nagahama, Y., *et al.*, 2016. Color computer-generated hologram generation using the random phase-free method and color space conversion. Appl. Opt. 55, 4159–4165.
Shimobaba, T., Masuda, N., Ito, T., 2009. Simple and fast calculation algorithm for computer-generated hologram with wavefront recording plane. Opt. Lett. 34, 3133–3135.
Shimobaba, T., Nakayama, H., Masuda, N., Ito, T., 2010. Rapid calculation algorithm of Fresnel computer-generated-hologram using look-up table and wavefront-recording plane methods for three-dimensional display. Opt. Express 18, 19504–19509.
Shishido, A., 2010. Rewritable holograms based on azobenzene-containing liquid-crystalline polymers. Polym. J. 42, 525–533.
Shrestha, P.K., Chun, Y.T., Chu, D.P., 2015. A high-resolution optically addressed spatial light modulator based on ZnO nanoparticles. Light Sci. Appl. 4, e259.
Slinger, C.W., Bannister, R.W., Cameron, C.D., *et al.*, 2001. Progress and prospects for practical electroholographic display systems. Proc. SPIE 4296, 18.
Slinger, C.W., Cameron, C.D., Coomber, S.D., *et al.*, 2004. Recent developments computer-generated holography: Toward a practical electroholography system interactive 3D visualization. Proc. SPIE 27, 5290.
Smalley, D.E., 2006. Integrated optics for holographic video. M.E.: Massachusetts Institute of Technology.
Smalley, D.E., Smithwick, Q.Y.J., Bove, V.M., 2007. Holographic video display based on guided-wave acousto-optic devices. Proc. SPIE 6488, 64880L.
Smalley, D.E., Smithwick, Q.Y.J., Bove, V.M., Barabas, J., Jolly, S., 2013. Anisotropic leaky-mode modulator for holographic video displays. Nature 498, 313–317.
Song, H., Sung, G., Choi, S., *et al.*, 2012. Optimal synthesis of double-phase computer generated holograms using a phase-only spatial light modulator with grating filter. Opt. Express 20, 29844–29853.
SPIE, 2017. Available at: http://spie.org/newsroom/1015-projection-display-using-computer-generated-phase-screens?Highlight=x2408&ArticleID=x20068 (accessed 21.05.17).
Stanley, M., Conway, P.B., Coomber, S.D., *et al.*, 2000. Novel electro-optic modulator system for the production of dynamic images from giga-pixel computer-generated holograms. Proc. SPIE 3956, 13.
Stanley, M., Smith, M.A., Smith, A.P., *et al.*, 2004. 3D electronic holography display system using a 100-megapixel spatial light modulator. Proc. SPIE 5249, 297.
Stein, A.D., Wang, Z., Leigh Jr., J.S., 1992. Computer-generated holograms: A simplified ray-tracing approach. Computers in Physics 6, 389–392.
Stevenson, S.H., 1997. DuPont multicolor holographic recording films. Proc. SPIE 3011, 231.
Sun, J., Timurdogan, E., Yaacobi, A., Hosseini, E.S., Watts, M.R., 2013. Large-scale nanophotonic phased array. Nature 493, 195–199.
Takaki, Y., Fujii, K., 2014. Viewing-zone scanning holographic display using a MEMS spatial light modulator. Opt. Express 22, 24713–24721.
Takaki, Y., Hayashi, Y., 2008. Increased horizontal viewing zone angle of a hologram by resolution redistribution of a spatial light modulator. Appl. Opt. 47, D6–D11.
Takaki, Y., Matsumoto, Y., Nakajima, T., 2015. Color image generation for screen-scanning holographic display. Opt. Express 23, 26986–26998.
Takaki, Y., Nakamura, J., 2014. Generation of 360-degree color three-dimensional images using a small array of high-speed projectors to provide multiple vertical viewpoints. Opt. Express 22, 8779–8789.
Takaki, Y., Okada, N., 2009. Hologram generation by horizontal scanning of a high-speed spatial light modulator. Appl. Opt. 48, 3255–3260.
Takaki, Y., Okada, N., 2010. Reduction of image blurring of horizontally scanning holographic display. Opt. Express 18, 11327–11334.
Takaki, Y., Taira, K., 2016. Speckle regularization and miniaturization of computer-generated holographic stereograms. Opt. Express 24, 6328–6340.
Takaki, Y., Uchida, S., 2012. Table screen 360-degree three-dimensional display using a small array of high-speed projectors. Opt. Express 20, 8848–8861.
Takaki, Y., Yokouchi, M., 2011. Speckle-free and grayscale hologram reconstruction using time-multiplexing technique. Opt. Express 19, 7567–7579.
Tanjung, R.B.A., Xu, X., Liang, X., *et al.*, 2010. Digital holographic three-dimensional display of 50-Mpixel holograms using a two-axis scanning mirror device. Optical Engineering 49, 025801. –025801–9.
Tay, S., Blanche, P.-A., Voorakaranam, R., *et al.*, 2008. An updatable holographic three-dimensional display. Nature 451, 694–698.
Teng, D., Liu, L., Wang, Z., Sun, B., Wang, B., 2012. All-around holographic three-dimensional light field display. Optics Communications 285, 4235–4240.
Texas Instruments, 2017. Available at: http://www.ti.com/lsds/ti/dlp/advanced-light-control/ultraviolet/ultraviolet-products.page#p2439=32,552 (accessed 21.05.17).
Texas Instruments, 2017. Available at: http://www.ti.com/lsds/ti/dlp-technology/markets/dlp-products-for-4k-ultra-hd.page (accessed21.05.17).
Tsang, P., Cheung, W.-K., Poon, T.-C., Zhou, C., 2011. Holographic video at 40 frames per second for 4-million object points. Opt. Express 19, 15205–15211.
Tsang, P.W.M., Jiao, A.S.M., Poon, T.-C., 2014. Fast conversion of digital Fresnel hologram to phase-only hologram based on localized error diffusion and redistribution. Opt. Express 22, 5060–5066.
Tsang, P.W.M., Poon, T.-C., 2013. Novel method for converting digital Fresnel hologram to phase-only hologram based on bidirectional error diffusion. Opt. Express 21, 23680–23686.
Underkoffler, J.S., 1997. Occlusion processing and smooth surface shading for fully computed synthetic holography. Proc. SPIE 19, 3011.
Weng, J., Shimobaba, T., Oikawa, M., Masuda, N., Ito, T., 2011. Fast recurrence algorithm for computer-generated-hologram. In: Digital Holography Three-Dimensional Imaging. Optical Society of America, p. DTuC21.
Wyrowski, F., 1990. Diffractive optical elements iterative calculation of quantized, blazed phase structures. J. Opt. Soc. Am. A 7, 961–969.
Xia, X., Liu, X., Li, H., *et al.*, 2013. A 360-degree floating 3D display based on light field regeneration. Opt. Express 21, 11237–11247.
Yamaguchi, M., Hoshino, H., Honda, T., Ohyama, N., 1993. Phase-added stereogram: Calculation of hologram using computer graphics technique. Proc. SPIE 1914, 25.

Yoneyama, T., Ichikawa, T., Sakamoto, Y., 2014. Semi-portable full-color electro-holographic display with small size. Proc. SPIE 9006, 900617.
Yoneyama, T., Yang, C., Sakamoto, Y., Okuyama, F., 2013. Eyepiece-type full-color electro-holographic binocular display with see-through vision. In: Digital Holography Three-Dimensional Imaging.Optical Society of America, p. DW2A.11.
Yoshida, S., 2016. fVisiOn 360-degree viewable glasses-free tabletop 3D display composed of conical screen and modular projector arrays. Opt. Express 24, 13194–13203.
Zacharovas, S.J., Rodin, A.M., Ratcliffe, D.B., Vergnes, F.R., 2001. Holographic materials available Geola. Proc. SPIE 4296, 206.
Zhang, H., Collings, N., Chen, J., *et al*., 2011. Full parallax three-dimensional display with occlusion effect using computer generated hologram. Optical Engineering 50, 074003–074005. [074003–].
Zhang, Z., You, Z., Chu, D.P., 2014. Fundamentals of phase-only liquid crystal on silicon (LCOS) devices. Light Sci. Appl. 3, e213.
Zhao, Y., Cao, L., Zhang, H., He, Q., 2012. Holographic display with LED illumination based on phase-only spatial light modulator. Proc. SPIE 8559, 85590B.
Zhou, L., Chen, C.P., Wu, Y., *et al*., 2017. See-through near-eye displays enabling vision correction. Opt. Express 25, 2130–2142.

Holography: Computer Generated Holograms

WJ Dallas, University of Arizona, Tucson, AZ, United States
AW Lohmann[†]

From the Classical Hologram to the Computer Generated Hologram: CGH

Holography is a method to form images. The method consists of two steps: recording and reconstruction. In the recording step, some interference fringes are recorded on a photographic plate (**Fig. 1(A)**). The two interfering waves come from the object and from a reference light source. After being developed by the usual photo-chemical treatment the photographic plate is called a hologram. In the reconstruction step, the hologram acts as a diffracting object that is illuminated by a replica of the reference light. Due to diffraction, the light behind the hologram is split into three parts (**Fig. 1(B)**). One part proceeds to the observer who perceives a virtual object where the genuine object had been during the recording step.

This was a brief description of "classical – holography". In "computer holography" the first step, recording, is synthetic. In other words, the propagation of the light from the object to the photographic plate is simulated digitally. The simulation includes the addition of the object wave to the reference wave and the subsequent modulus square process, which describes the transition from the two complex amplitudes to the hologram irradiance:

$$|u_O(x) + u_R(x)|^2 = I_H(x) \tag{1}$$

The amplitude transmittance of the hologram is

$$T_H(x) = c_0 + c_1 I_H(x) = c_1 u_O u_R^* + \ldots \tag{2}$$

The coefficients c_0 and c_1 have to do with the photochemical development. In the reconstruction process (**Fig. 1(B)**) the transmittance T_H is illuminated by the reference beam. Hence, we get

$$u_R(x)T_H(x) = c_1 u_O(x)|u_R(x)|^2 + \ldots \tag{3}$$

If the reference intensity $|u_R(x)|^2$ is constant, this term represents a reconstruction of the object wave $u_O(x)$. The omitted parts in Eq. (2) and in Eq. (3) describe those beams that are ignored by the observer (**Fig. 1(B)**). The hologram transmittance $T_H(x)$ is an analog signal:

$$0 \le T_H \le 1 \tag{4}$$

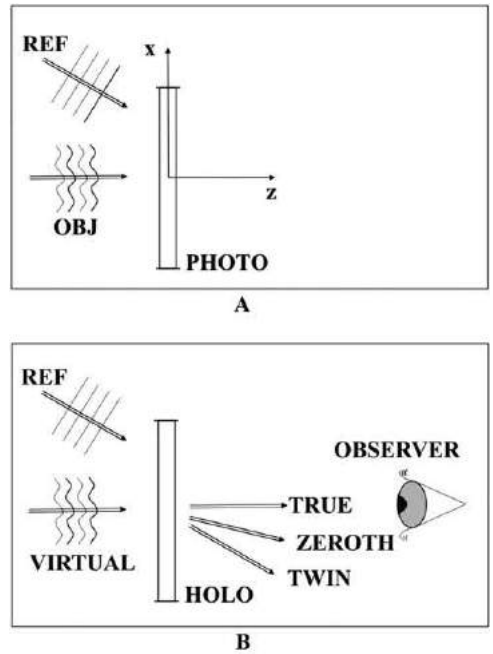


Fig. 1 Holographic image formation in two steps: (A) Recording the hologram. (B) Reconstructing the object wave.

[†]Deceased.

 doi:10.1016/B978-0-12-803581-8.09617-X

50 years ago, when computer holograms were invented (Brown and Lohmann, 1966; Lohmann and Paris, 1967), the available plotters could produce only binary transmittance patterns. Hence, it was necessary to replace the analog transmission $T_H(x)$ by a binary transmission $B_H(x)$. That is possible indeed, without loss of image quality. In addition, binary computer holograms have better light efficiency and a better signal-to-noise ratio. They are also more robust when being copied or printed.

Another advantage of a computer generated hologram (CGH) is that the object does not have to exist in reality. That is important for some applications. The genuine object may be difficult to manufacture or, if it is three-dimensional, difficult to illuminate. The CGH may be used to implement an algorithm for optical image processing. In that case, the term "object" becomes somewhat fictitious. We will discuss more about applications, in Section Some CGH Applications, towards the end of this entry.

From a Diffraction Grating to a Fourier CGH

Now we will explain, in some detail, the "Fourier type" CGH. Fourier holograms are more popular than two other types: "image-plane" CGH's and "Fresnel-type" CGH's. The Fresnel-type will be treated in a section on "software" because of some interest in 3-D hologram's (Section About some CGH algorithms). Certain hardware issues will appear in Section On Some Hardware Aspects.

The explanation of Fourier CGH's begins with grating diffraction (Fig. 2). The only uncommon feature in Fig. 2 is the off-axis location of the source. It is arranged such that the plus-first diffraction order will hit the center of the output plane.

We will now convert a simple Ronchi grating (Fig. 3(A)) into a Fourier CGH. The Ronchi grating transmittance can be expressed as the Fourier series:

$$G(x) = \sum_{(m)} C_m \exp\left(\frac{2\pi i\, m\, x}{D}\right);\; C_0 = \frac{1}{2};\; C_m = \frac{\sin(\pi\, m/2)}{\pi\, m} \tag{5}$$

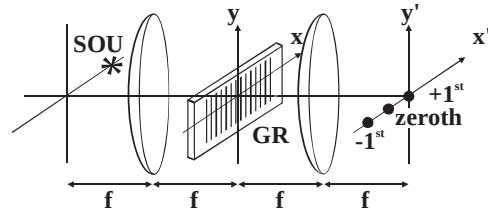


Fig. 2 Grating diffraction, but with an off-axis source.

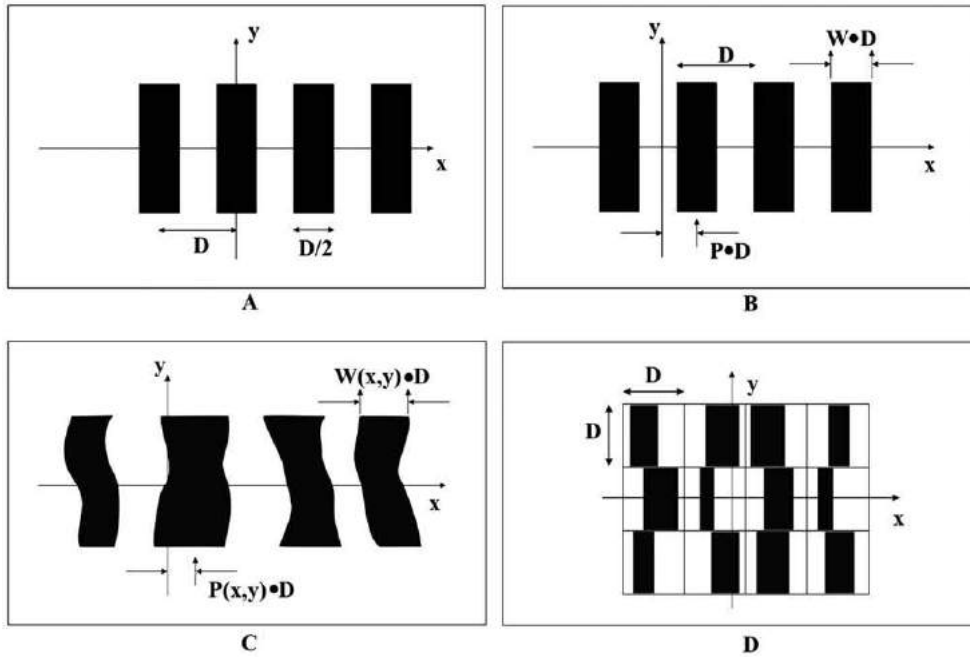


Fig. 3 Modifying a grating into a hologram in 3 steps: (A) Rhonchi grating. (B) Modified width and location of the grating bars. (C) Distorted grating bars. (D) As in Fig. 3(C), but now discrete modifications.

Now we shift the grating bars by an amount PD. And, we modify the width from $D/2$ to WD, as shown in **Fig. 3(B)**. The Fourier coefficients are now

$$C_m = \frac{\exp(2\pi imP)\,\sin(\pi mW)}{\pi m} \tag{6}$$

The C_{+1} coefficient is responsible for the complex amplitude at the center of the readout plane:

$$C_{+1} = \frac{\exp(2\pi iP)\,\sin(\pi W)}{\pi} \tag{7}$$

We are able now to control the beam magnitude, $A = \left(\frac{1}{\pi}\right)\sin(\pi W)$, by the relative bar width W and the phase $\phi = 2\pi P$ by the relative bar shift P. This particular kind of phase shift is known as "detour phase" (Hauk and Lohmann, 1958).

So far, the light that ends up at the center of the exit plane, leaves the grating as a plane wave with a wave vector parallel to the optical axis. Now we distort the grating on purpose (**Fig. 3(C)**):

$$W \to W(x,y);\ P \to P(x,y) \tag{8}$$

These distortions should be mild enough that the plus-first order light behind the distorted grating can be described by the complex amplitude:

$$\left(\frac{1}{\pi}\right)\sin[\pi W(x,y)]\,\exp[2\pi iP(x,y)] = u_H(x,y) \tag{9}$$

This is a generalization of Eq. (7). The phase $2\pi P(x,y)$ covers the range from $\frac{-\pi}{2}$ to $\frac{\pi}{2}$ if the shift is bounded by $|P| \le \frac{1}{2}$. That is enough to cover the range of complex amplitudes within the circle of $|u_H| \le \frac{1}{\pi}$. We will return shortly to the condition "mild-enough distortions" in the context of **Fig. 4**.

Most plotters move their pens only in x-direction and y-direction and printers place dots on a grid. Therefore it is convenient to replace the continuous (x,y) variations of W and P by piece-wise constant variations as in **Fig. 3(D)**. That restriction turns out to be acceptable, as demonstrated by the first computer holograms. The theoretical proof involves the sampling theorem (Lohmann and Paris, 1967).

Recall that the light propagation through a 2-f system (**Fig. 2**, between grating and output plane) can be described by a Fourier transformation:

$$\iint u_H(x,y)\,\exp[-2\pi i(xx' + yy')/\lambda f]\,dxdy = \tilde{u}_H(x'/\lambda f, y'/\lambda f) \tag{10}$$

Suppose now that we wish to see a particular image: $v(x',y')$. Hence, we request that

$$v(x',y') = \tilde{u}_H(x'/\lambda f, y'/\lambda f)RECT(x'/\lambda f) \tag{11}$$

The RECT function indicates that only the PLUS first diffraction order is of interest. As a consequence we have to select the grating distortions $W(x,y)$ and $P(x,y)$ such that the hologram $u_H(x,y)$ is the Fourier transform of $v(x',y')$:

$$u_H(x,y) = \iint v(x',y')\,\exp[2\pi i(xx' + yy')/\lambda f]\,dx'dy' \tag{12}$$

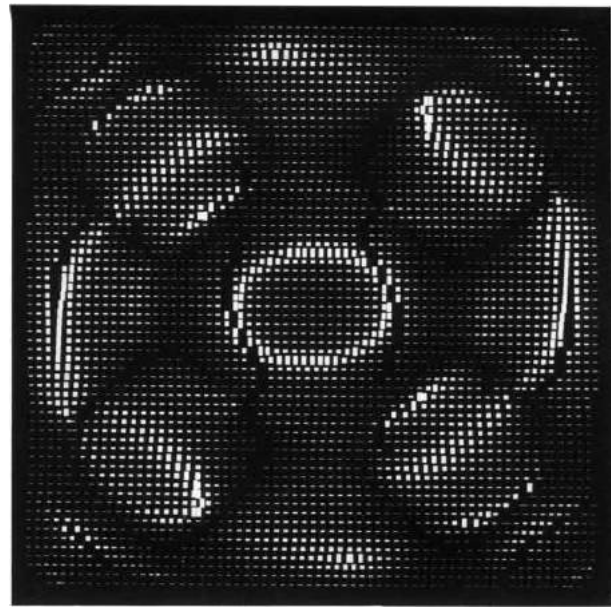


Fig. 4 CGH with 64×64 cells.

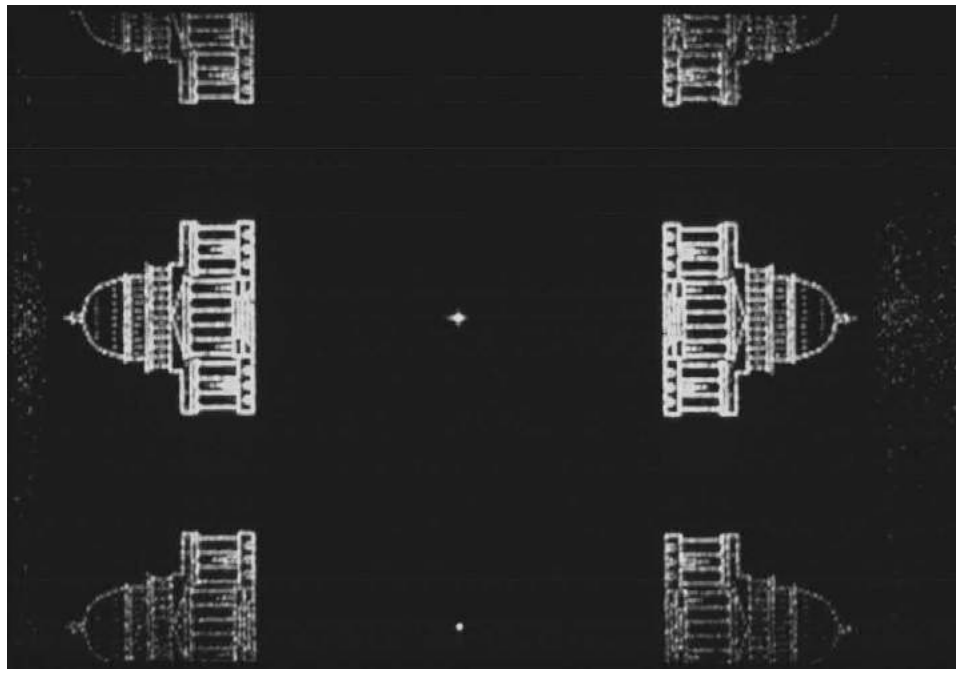


Fig. 5 A reconstruction from a 1024 × 1024 CGH. The true image and the symmetrical twin image appear. The zeroth-order is blocked. Reproduced from Brown, B.R., Lohmann, A.W., 1969. IBM J. R&D 13, 160.

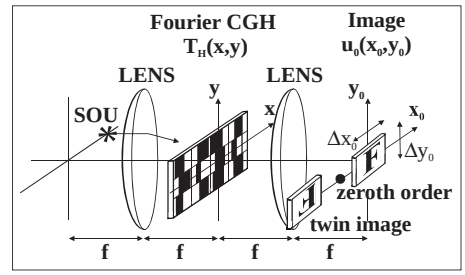


Fig. 6 Reconstruction setup for a Fourier CGH.

The condition "mild enough distortions" has been treated in the literature. There is not enough space here to present that part of the CGH theory. But we show an actual CGH with 64 × 64 cells (**Fig. 4**). This CGH is no longer a grating. But, it shows enough resemblance to our heuristic derivation. Hence, the CGH theory appears to be trustworthy. A CGH looks less regular if the image v (x',y') contains a random phase that simulates a diffuser. The diffuser spreads light across the CGH and so levels out unacceptably large fluctuations in the amplitude.

With 512 × 512 or 1024 × 1024 cells in a CGH one can obtain an image of the quality seen in **Fig. 5** (Brown and Lohmann, 1969). The reconstruction setup is shown in **Fig. 6**, here with the letter F as the image.

About Some CGH Algorithms

The computational effort for getting the desired hologram amplitude of a Fourier hologram u_H is reasonable due to the powerful FFT algorithm. To use it the data must be discrete. In other words the CGH consists of cells, centered at $x_m = mD$, $\gamma_n = nD$ as shown in **Fig. 3(D)**. The questions now are: "What is the proper cell size D?" and "How many cells are needed?" The answers to those questions depend on the parameters of the image. If the zero order (see **Fig. 6**) should be at the distance Δx_0, then the quasi-grating period D should obey:

$$\lambda f / D \geq \Delta x_0 \tag{13}$$

And if the size of an image pixel should be as fine has δx_0 the size Δx_H of the hologram should obey:

$$\lambda f / \Delta x_H \leq \delta x_0 \tag{14}$$

This condition is plausible because the finite size Δx_H of the hologram acts like a resolution-limiting aperture. The combination of Eqs. (13) and (14) yields

$$\Delta x_0/\delta x_0 \leq \Delta x_H/D \tag{15}$$

This means that the number of image pixels $\Delta x_0/\delta x_0$ is bounded by the number of CGH cells $\Delta x_H/D$. The generalization to two dimensions is straightforward. Again, Fourier CGH's benefit from the fact that the light propagation from the virtual object to the hologram plane can be described by a Fourier transformation. Hence, the FFT can be used.

And now we move to the synthesis of a Fresnel hologram. A Fresnel hologram is recorded at a finite distance z from the object. Hence, we have to simulate the wave propagation in free space from the object (at $z=0$) to the hologram at a distance z. This free space propagation can be described (in the paraxial approximation) as a convolution of object and quasi-spherical wave:

$$u_0(x) \rightarrow \int_{-\infty}^{\infty} u_0(x') \exp\left[\frac{-i\pi(x-x')^2}{\lambda z}\right] dx' = u(x,z) \tag{16}$$

In the Fourier domain the corresponding operation is:

$$\tilde{u}_0(\mu) \rightarrow \tilde{u}_0(\mu) \exp\left[-i\pi\lambda z\mu^2\right] = \tilde{u}(\mu,z) \tag{17}$$

The indirect execution of Eq. (16), based on Eq. (17), is easy: a Fourier transformation of the object converts $u_0(x)$ into $\tilde{u}_0(\mu)$, which is then multiplied by the quadratic phase factor, yielding $\tilde{u}(\mu,z)$. An inverse Fourier transformation produces $u(x,z)$. The two Fourier transformations consist typically of $4N \log N$ multiply/add operations. N is the number of pixels: image size Δx_0, divided by the pixel size δx_0. In the case of a rectangular object $u_0(x,y)$, the number of pixels is $\Delta x_0\ \Delta y_0/\delta x_0\ \delta y_0$.

And now to the algorithmic effort for computing the convolution. The $u_0(x)$ has N pixels. And, the lateral size of the exponential is essentially $M = z/\delta z$, where δz means "depth of focus". The M describes the lateral size of the diffraction cone, measured in pixelsize units δx_0. The convolution involves MN multiply/adds. A comparison of these two algorithms is dictated by the ratio

$$M/4\log N \tag{18}$$

The direct convolution is preferable if the distance z is fairly small. The integration limits of Eq. (16) depend on the oscillating phase $\pi x^2/\lambda z$. Effectively, nothing is contributed to this integral, when this phase varies by more than 2π over the length, δx_0, of an object pixel. Note that the pixel size of $u(x,z)$ does not change with the distance z, because the bandwidth $\Delta\mu(z) = \Delta\mu_0$ is z-independent. This is so because the power spectrum is z-invariant:

$$|\tilde{u}(\mu,z)|^2 = |\tilde{u}_0(\mu)|^2 \tag{19}$$

It is easy to proceed from here to the CGH synthesis of a 3-D object. One starts with an object detail $u_1(x)$ at z_1, for example, a house as in **Fig. 7**. Then, one propagates to the next object detail at z_2. This second detail may act as a multiplication (transmission) and as addition (source elements). This process is repeated as the wave moves towards the plane of the hologram. There, a reference wave has to be added just as in the Fourier case.

A very powerful algorithm for the CGH synthesis isiterative Fourier transform algorithm (IFTA) (Lesem *et al.*, 1969; Gerchberg and Saxton, 1972; Fienup, 1981). Suppose, we want to design a Fourier CGH, which looks like a person and whose reconstructed image shows the signature of that person (**Fig. 8(A)** and **(B)**). In other words, the two intensities

$$|u_H(x,y)|^2 \text{ and } |\tilde{u}_H(\mu,(v))|^2 \tag{20}$$

are determined. But, the phase distributions of u_H and $\tilde{u}_H$ are still at our disposal. The IFTA algorithm will yield those phases in many cases at the expense of ten, hundred or more Fourier transformations.

The IFTA algorithm is sometimes called "Fourier Ping-Ping" because it bounces back and forth between the (x,y) and the (μ,v) domain. The amplitudes $|u_H|$ and $|u_0|$ are enforced at every opportunity. But the associated phases are free to vary and to converge, eventually. If so, u_H and u_0 are completely known. The IFTA will not always converge, for example not if

$$|u_H(x,y)| = \delta(x,y) \text{ and } |\tilde{u}_H(\mu,(v))| = \delta(\mu,(v)) \tag{21}$$

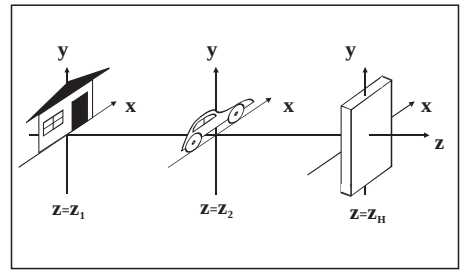


Fig. 7 Synthesis of a hologram for a three-dimensional object.

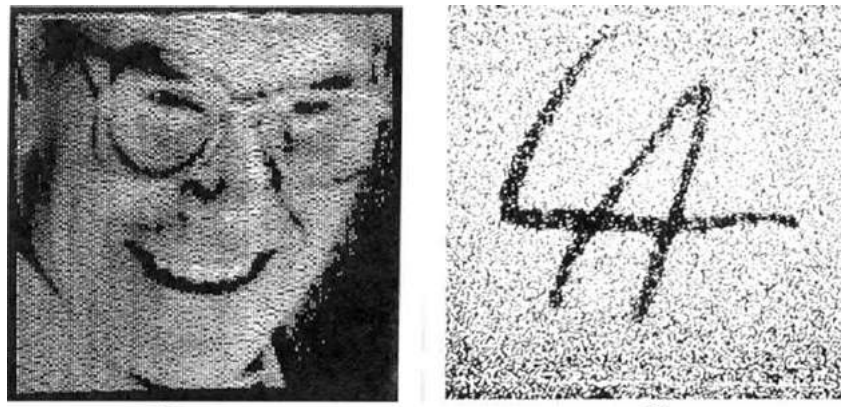


Fig. 8 IFTA-CGH: (A) The Fourier CGH. (B) The optical reconstruction thereof. Courtesy of Bartelt, H.O.

But the chances are fairly good if both amplitudes are very often close to one and seldom near zero.

Finally, a few words about image holograms. An image plane CGH is trivial from an algorithmic point of view. The simulated propagation from object plane to image plane is usually an identity, apart from the finite bandwidth of the image forming system. Nevertheless, image plane holograms can be quite useful, for example for interferometric testing of a non-spherical mirror of an astronomical telescope (Hansler, 1968; McGovern and Wyant, 1971; Ichioka and Lohmann, 1972). The CGH is laid out to provide a wavefront that is suitable for an interferometric "null test". The CGH serves as a "synthetic prototype".

A fourth kind of CGH, to be named near-field CGH may emerge, perhaps. It operates in the near-field. And, it employs evanescent waves as they occur if object details are comparable in size with, or even smaller than, the wavelength. Conceivable topics are "super resolution" and "sub-lambda lithography". Fundamentals and history of classical evanescent holography are reported in Bryngdahl (1973).

On Some Hardware Aspects

And now a few comments about the hardware aspects of the CGH's. Remembering the CGH cells structure in **Fig. 3(D)** we saw the amplitude encoded as relative width W (Eq. (6)). Instead of the width one may use the relative height $H \leq 1$ as the amplitude parameter (Lohmann and Sinzinger, 1995). More efficient, but more difficult to manufacture are saw tooth shaped hologram cells (Brown and Lohmann, 1969). The saw tooth prism may be approximated by a phase stair, which is possible with today's lithographic technology. The corresponding theory is called phase quantization (Goodman and Silvestri, 1970; Dallas, 1971).

Some CGH Applications

Optical Filtering

Some applications had been mentioned already before, for example, 3-D Display (see **Fig. 7**). The data volume can be very high. Hence, shortcuts such as eliminating the vertical perspective, are called for. The largest CGH's for visual light are used for TESTING astronomical telescopes by means of interferometry. The IFTA algorithm has been used, among other things, for designing a CGH which acts as an optimal diffuser. Speckles and light efficiency are the critical issues in diffuser design. Another project that also benefits from IFTA is beam shaping. That is a broad topic with aims such as homogenizing the output of laser, beam structuring for welding (Dresel *et al.*, 1996), cutting, and scanning.

A CGH can also be used as a spatial filter for image processing. For example, an input image may be Hilbert-transformed in the setup shown in **Fig. 9(A)** (Hauk and Lohmann, 1958). The filter (**Fig. 9(B)**) is a grating, of which one sideband is shifted by one-half of the grating period. Such a geometric shift generates the π-phase shift that is needed for the Hilbert transformation. The same setup can be used also for implementing Zernike's phase contrast (Lohmann and Paris, 1968). Now the grating is distorted such that the zero-frequency region at the optical axis is reduced in amplitude by (typically) a factor of 0.2. In addition, the phase is shifted by $\pi/2$ due to a fringe shift of a quarter period (**Fig. 10(A)**). In the image plane one can see an ordinary image on-axis, and two phase-contrast images in the plus-minus first diffraction orders. One of them has positive phase contrast, the other one has negative phase contrast (**Fig. 10(B)**). Another simple spatial filter (**Fig. 11(A)**) produces a derivative of the input $u(x,y)$: $\frac{\partial u(x,y)}{\partial x}$ (**Fig. 11(B)**). Eqs. (22)–(24) explain this differential filtering experiment.

$$u(x,y) \rightarrow \partial u(x,y)/\partial x = \mathrm{v}(x,y) \tag{22}$$

$$\tilde{u}(\mu,(v)) \rightarrow 2\pi i\mu \cdot \tilde{u}(\mu,(v)) = \tilde{\mathrm{v}}(\mu,(v)) \tag{23}$$

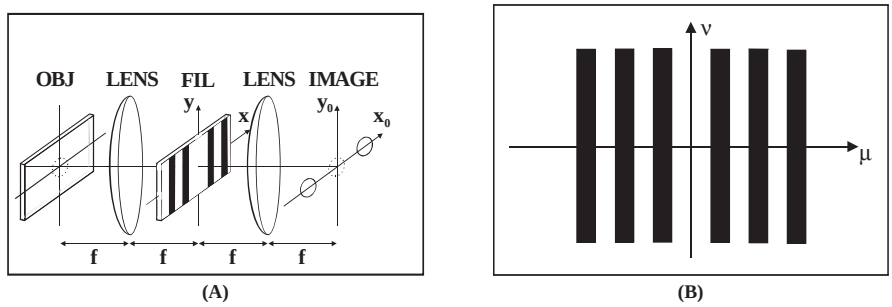


Fig. 9 (A) Hilbert setup. (B) Hilbert filter.

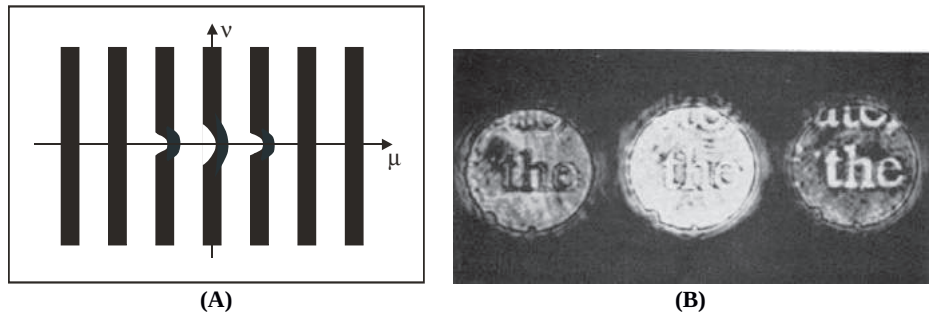


Fig. 10 (A) Phase contrast filter. (B) Phase contrast output.

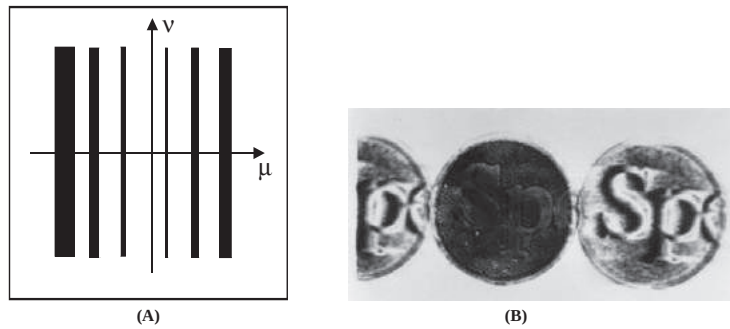


Fig. 11 (A) Differentiation filter. (B) Differentiation output.

$$\tilde{p}(\mu,(\nu)) = 2\pi|\mu| \cdot \begin{cases} +i(\mu > 0) \\ -i(\mu > 0) \end{cases} \tag{24}$$

The phase portion of the filter (**Fig. 11(A)**) is the same as for the Hilbert filter. But, the filter amplitude now enhances the higher frequencies. The object used for getting the results of **Figs. 10(B)** and **11(B)** was a phase-only object: a bleached photograph.

A very famous spatial filtering experiment is "matched filtering" (Vander Lugt, 1964; Kozma and Kelly, 1965). It can be used for pattern recognition. The original matched filters were classical Fourier holograms of the target pattern. Computer hologram's can do the same job (Brown and Lohmann, 1966), even with spatially incoherent objects (Lohmann and Werlich, 1971). Work on spectrally incoherent spatial filtering was initiated by Katyl (1972).

Fourier holograms are attractive also for data storage due to their robustness against local defects. A local error in the Fourier domain spreads out all over the image domain. The computed hologram has an advantage over classical holograms because of the freedom to incorporate any error detection and correction codes for reducing the bit error rate even further.

One of the most recent advances is in matched filtering with totally incoherent light (Andrés *et al.*, 1999; Pe'er *et al.*, 1999). This step allows moving spatial filtering out of the well-guarded laboratory into hostile outdoor environments.

Optical Vortices

The mathematical key to understanding CGH generation of optical vortices (Rozas *et al.*, 1997) is the circular harmonic. It is a function of two variables where θ is the angular member of the polar coordinate pair and n is the order of the harmonic. The

circular harmonic is

$$\exp(in\theta) \tag{25}$$

An optical wave at a plane with this complex amplitude $\exp(in\theta)$ is called an optical vortex with a topological charge of n. In order to design a CGH that converts a plane wave to an optical vortex the circular harmonoic decomposition is used. Extending Eq. (2) to polar coordinates, the transmittance of the CGH is

$$T_H(r,\theta) = c_0 + c_1 \exp(in\theta) u_R^* + \ldots \tag{26}$$

For a unit-amplitude normally-incident wave, the exiting optical vortex is numberically equal. The corresponding CGH mask is shown in **Fig. 12**.

These waves can be used, as propeller beams (Zhang *et al.*, 2010) that apply torque to components (cranks) of micro-mechanical devices. They can be used as optical traps (Ng *et al.*, 2010), for example, to trap Bose-Einstein condensates. They can also be employed in solar coronagraphs (Foo *et al.*, 2005).

Polarization CGH

Polarized light can be decomposed into two orthogonal components. For example, if the light is propogating in the z-direction, light in a plane can be decomposed into linearly polarized components: x-polarized and y-polarized. An arbitrarily polarized field can be composed by superimosing a field of x-polarized and a field of y-polarized light. This process is know as polarization synthesis (Lohmann, 1667; Noble *et al.*, 2010). An idealized synthesis process is illustrated in (**Fig. 13**) where two Fourier CGHs are inserted into a Mach-Zehnder interferometer. It combines the waves exiting both holograms into the aritrarily-polarized reconstructed image.

Determining the strength of each component is know as polarization analysis. The generation steps for calculating the encoded waves of the CGHs are illustrated in the flow diagram of **Fig. 14**.

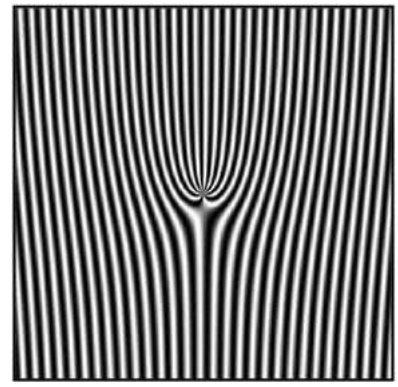


Fig. 12 Ninth-order vortex CGH.

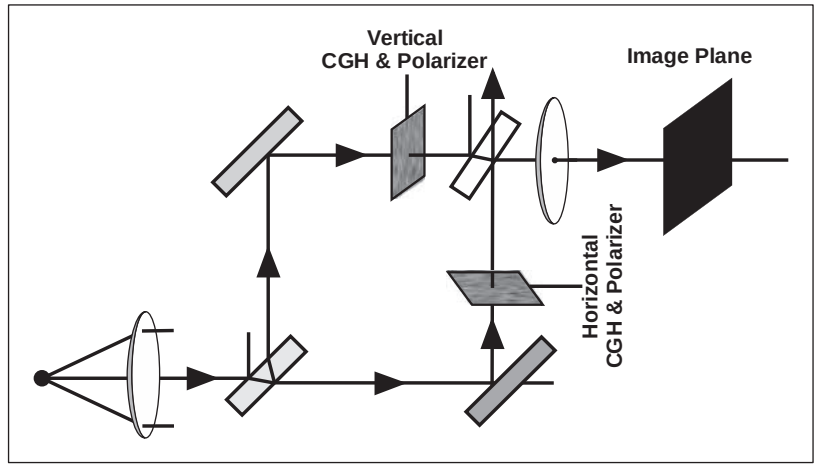


Fig. 13 Polarized reconstruction.

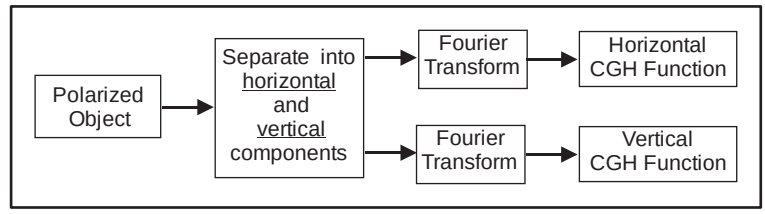


Fig. 14 Flow diagram for straight-forward designing of a polarization CGH.

Terminology

Several terms mean more or less the same as CGH: synthetic hologram, digital hologram, holographic optical element (HOE), diffractive optical element (DOE) or digital optical element.

CGH's have been made also for other waves, such as acoustical waves, ocean waves, and microwaves. Digital antenna steering can be understood as a case of computer holography.

See also: Module: Digital Holography

References

Andrés, P., *et al.*, 1999. Opt. Lett. 245, 1331.
Brown, B.R., Lohmann, A.W., 1966. Appl. Opt. 5, 967.
Brown, B.R., Lohmann, A.W., 1969. IBM J. R&D 13, 160.
Bryngdahl, O., 1973. Prog. Opt. 11, 168.
Dallas, W.J., 1971. Appl. Opt. 10, 673.
Dresel, T., Beyerlein, M., Schwider, J., 1996. Appl. Opt. 35, 4615.
Fienup, J.R., 1981. Proc. SPIE 373, 147.
Foo, G., Palacios, D.M., Swartzlander, G.A., 2005. Opt. Lett. 30 (24), 3308.
Gerchberg, R.W., Saxton, W.O., 1972. Optik 35, 237.
Goodman, J.W., Silvestri, A.M., 1970. IBM J. R&D 14, 478.
Hansler, R.L., 1968. Appl. Opt. 7, 1863.
Hauk, D., Lohmann, A.W., 1958. Optik 15, 275.
Ichioka, Y., Lohmann, A.W., 1972. Appl. Opt. 11, 2597.
Katyl, R.H., 1972. Appl. Opt. 11, 1255.
Kozma, A., Kelly, D.L., 1965. Appl. Opt. 4, 387.
Lesem, L.B., Hirsch, P.M., Jordan, J.A., 1969. IBM J. R&D 13, 150.
Lohmann, A.W., 1667. Reconstruction of vectorial wavefronts. Appl. Opt. 4, 1965.
Lohmann, A.W., Paris, D.P., 1967. Appl. Opt. 6, 1746.
Lohmann, A.W., Paris, D.P., 1968. Appl. Opt 7, 651.
Lohmann, A.W., Sinzinger, S., 1995. Appl. Opt. 34, 3172.
Lohmann, A.W., Werlich, H.W., 1971. Appl. Opt. 10, 670.
McGovern, A.J., Wyant, J.C., 1971. Appl. Opt. 10, 619.
Ng, J., Lin, Z., Chan, C.T., 2010. PhysRevLett. 104, 103601.
Noble, H., Ford, E., Dallas, W., *et al.*, 2010. Opt. Lett. 35 (20), 3423.
Pe'er, A., Wang, D., Lohmann, A.W., Friesem, A.A., 1999. Opt. Lett. 24, 1469. (2000; vol. 25, p. 776).
Rozas, D., Law, C.T., Swartzlander Jr., G.A., 1997. J. Opt. Soc. Am. B14, 3054.
Vander Lugt, A.B., 1964. IEEE IT 10, 139.
Zhang, P., Huang, S., Hu, Y., Hernandez, D., Chen, Z., 2010. Opt. Lett. 35, 3129.

Further Reading

Bryngdahl, O., Wyrowski, F., 1990. Prog. Opt. 28, 1–86.
Dallas, W.J., 1973. Appl. Opt. 12, 1179.
Dallas, W.J., 1980. Top. Appl. Phys., 41. Heidelberg: Springer, pp. 291–366.
Dallas, W.J., 1999–2003, University of Arizona, Tucson, AZ. Available at: http://www.radiology.arizona.edu/dallas/CGH.htm.
Fienup, J.R., 1982. Appl. Opt. 21, 2758.
Lee, W.H., 1978. Prog. Opt. 16, 119–232.
Sinzinger, S., Jahns, J., 1999. Micro Optics. New York: Wiley.
Wyrowski, F., Bryngdahl, O., 1991. Rep. Prog. Phys. 1481–1571.
Yaroslavskii, L.P., Merzlyakov, N.S., 1980. Methods in Digital Holography. New York: Plenum Press.

Module: Digital Holography

Wolfgang Osten, University of Stuttgart, Stuttgart, Germany

Introduction

One of the most fascinating properties of light is the fact that it encounters us both as a beam, as a wave, and as a particle. In the context of this article, however, we will only deal with the wave character of light and its sinusoidal nature. The accompanying ability of waves to interfere makes it possible to store spatial information on a two-dimensional target and to reconstruct the complete light field again. This is known as holography (Gabor, 1949). Even more, the superposition of waves and the precise knowledge of the wavelength guide us to a series of interesting methods for the high-resolution measurement of various physical quantities, such as three-dimensional distances, displacements, strains and refractive index distributions. The primary physical quantity of the wave field, which represents its three-dimensional properties, is the phase. Unfortunately, due to the high frequency of light, all media for the storage of light, whether photo-chemical (films, photoplates) or photo-electronic (CCD and CMOS detectors) can only register intensities. But the ability to interfere allows us to apply a trick: the interferometric principle. Simply by means of interferometry, phase changes are converted into measurable intensity changes. This is done in an interferometer where a primary wave front is divided into at least two wave trains, often called as object and reference wave, respectively, which are superimposed again after traveling along common or various paths. Finally, interferometry and holography enabled numerous outstanding applications such as the 3D-imaging of natural objects and scenes (Leith and Upatnieks, 1961; Denisyuk, 1962), the fabrication of diffractive optical elements (Lee, 1978), and the high-resolution inspection of static and dynamic objects having both precisely machined as well as rough surfaces (Steel, 1983; Powell and Stetson, 1965).

As already mentioned, the basic principle of holography consists in the transformation of phase changes into recordable intensity changes. Because of the high spatial frequency of these intensity fluctuations the registration of a hologram requires a light sensitive medium with adequate spatial resolution. Therefore, special photographic emulsions have dominated holographic technologies for a long period. However, the recording of holograms on electronic sensors and their numerical reconstruction is almost as old as holography itself. The first ideas and implementations were already conceived in the 1960s and 1970s (Goodman and Lawrence, 1967; Huang, 1971; Kronrod *et al.*, 1972; Demetrakopoulos and Mitra, 1974). But the most crucial step towards a practicable technology could only be made as digital cameras with reasonable space-bandwidth product and powerful processor technology became widely accessible. At the beginning of the 1990s these technical prerequisites were available and it was only a matter of time before they would be applied to the discrete recording, digital storage and numerical reconstruction of optical wave fronts, which is called digital holography. The article of Schnars and Jüptner from 1994 (Schnars and Jüptner, 1994) can be considered as a kind of initialization for a continuously growing number of implementations and applications. Meanwhile, digital holography has grown into a separate, distinct category in modern optics and has obvious implications in many fields such as microscopy (Yu *et al.*, 2014), 3D imaging (Frauel *et al.*, 2006), metrology (Kreis, 2005), display technology (Gang, 2013), material processing (Hasegawa *et al.*, 2006), data storage (Coufal *et al.*, 2000), and information processing (Yaroslavsky, 2004). A further remarkable extension of the application range of digital holography could be achieved by the application of sophisticated spatial light modulator (SLM) technology (Sutkowski and Kujawinska, 2000; Kohler *et al.*, 2006; Haist and Osten, 2015a,b), complementing the digital recording process with a matching digital optical reconstruction process. Since the late 1990s such devices have been offered by different companies (HOLOEYE Photonic AG; Lazarev *et al.*, 2012; HAMAMTSU Photonics K.K.) with the result that a bunch of new applications such as holographic micro-manipulation (Zwick *et al.*, 2010; Reicherter *et al.*, 1999; DaneshPanah *et al.*, 2010), aberration compensation (Ferraro *et al.*, 2003; Liesener *et al.*, 2004), computational microscopy (Maurer *et al.*, 2010; Haist *et al.*, 2014; Marquet and Depeursinge, 2014; Kim, 2011), holographic displays (Onural *et al.*, 2011; Lee and Kim, 2012), comparative digital holography (Osten *et al.*, 2002; Baumbach *et al.*, 2006), and remote holographic laboratories (Osten *et al.*, 2013) could be implemented. Further recent application fields are described in (Osten *et al.*, 2014).

Meanwhile, modern digital holography is more than 25 years old. But it is still an equally young and dynamic discipline. Some emerging fields where digital holography in combination with modern photonic technologies shows their great potential for new applications in technical and living sciences are:

- the implementation of sophisticated illumination, imaging and reconstruction principles for the enhancement of the resolution of holographically stored images and for the reduction of the demands on the space-bandwidth product of holographic sensors,
- the combination of digital-holographic microscopy with SLM-based holographic 3D-micromanipulation for the in vivo investigation of living cells and tissues,
- the evaluation of the spatial coherence function measured by self-referenced interferometry with the objective to reduce the degree of coherence of the light needed for holographic storage and to derive spatial and spectral information about the objects under test,
- the application of modern light sources such as supercontinuum lasers and frequency combs for the high-precision depth mapping and 3D-reconstruction of objects and scenes,

- the iterative evaluation of the propagating wavefront for avoiding sensitive interferometric detection principles,
- the evaluation of light fields transmitted or reflected from strongly scattering or hidden objects, and
- the enabling of holographic procedures and setups for remote access and comparative inspection technologies.

Direct Phase Reconstruction by Digital Holography

In digital speckle pattern interferometry DSPI (Butters and Leendertz, 1971) the object is focused onto the target of an electronic sensor. Thus an image plane hologram is formed as result of the interference with an inline reference wave. In contrast with DSPI, a digital hologram can be recorded without imaging the object onto a target. The target records the superposition of the reference and the object wave in the near-field region – a so-called Fresnel hologram (Schnars and Jüptner, 1994). The basic optical setup in digital holography for recording holograms is the same as in conventional holography, **Fig. 1(a)**. A laser beam is divided into 2 coherent beams. One beam illuminates the object and generates the object beam $u(x,y)$. The other enters the target directly and forms the reference wave $r(x,y)$. On this basis very compact solutions are possible. **Fig. 1(b)** shows a holographic camera where the interferometer, containing a beam splitter and some wave front shaping components, is mounted in front of the camera target. The objected, drafted as a ball, is illuminated from four directions. This scheme has some advantages for the measurement of the shape and displacement of the object under test (see Section "The Fresnel approximation").

For the description of the principle of digital Fresnel holography we use a simplified version of the optical setup, **Fig. 2(a)**. The object is modeled by a plane rough surface that is located in the (x,y)-plane and illuminated by laser light. The scattered wave field forms the object wave $u(x,y)$. The target of an electronic sensor (e.g. a CCD or a CMOS) used for recording the hologram is located in the (ξ,η)-plane at the distance d from the object. Following the basic principles of holography (Gabor, 1949; Hariharan, 1984) the hologram $h(\xi,\eta)$ originates from the interference of the object wave $u(\xi,\eta)$ and the reference wave $r(\xi,\eta)$ in the (ξ,η)-plane:

$$h(\xi,\eta) = |u(\xi,\eta) + r(\xi,\eta)|^2 = r \cdot r^* + r \cdot u^* + u \cdot r^* + u \cdot u^* \tag{1}$$

The transformation of the intensity distribution $h(\xi,\eta)$ into a discrete and digital gray value distribution $T(m,n)$ that is stored in the image memory of the computer is considered by a characteristic function t of the sensor. This function is in general only approximately linear:

$$T(m,n) = t[h(\xi,\eta)] \tag{2}$$

Because the sensor has a limited spatial resolution the spatial frequencies of the interference fringes in the hologram plane – the so-called micro-interferences – have to be considered. The fringe spacing g in, for example, x-direction and the corresponding spatial frequency f_x, respectively, is determined by the angle β between the object and the reference wave, **Fig. 2(b)**:

$$g = \frac{1}{f_x} = \frac{\lambda}{2\sin(\beta/2)} \tag{3}$$

with λ as the wavelength. If we assume that the discrete sensor has a pixel pitch (distance between two adjacent pixels) $\Delta\xi$ the sampling theorem requires at least 2 pixels per fringe for a correct reconstruction of the periodic function:

$$2\Delta\xi < \frac{1}{f_x} \tag{4}$$

Consequently, we obtain for small angles β:

$$\beta < \frac{\lambda}{2\Delta\xi} \tag{5}$$

Modern high-resolution CCD or CMOS chips have a pitch $\Delta\xi$ of about 4μm. In that case a maximum angle between the reference and the object wave of only 4° is acceptable. The practical consequence of the restricted angle resolution in digital holography is a limitation of the effective object size that can be stored holographically by an electronic sensor. However, this is only a technical handicap. Larger objects can be placed in a sufficient distance from the hologram, or reduced optically by imaging with a negative lens.

The reconstruction is done by illumination the hologram with a so-called reconstruction wave $c(\xi,\eta)$:

$$u'(x',y') = t[h(\xi,\eta)] \cdot c(\xi,\eta) \tag{6}$$

The transmission function of the hologram $t[h(\xi,\eta)]$ (see Eq. (2)) diffracts the wave $c(\xi,\eta)$ in such a way, that images of the object wave are reconstructed. In general, four terms are reconstructed if the wave $u'(x',y')$ propagates in space. In case of a linear characteristic function $t(h) = \alpha h + t_0$ we can write:

$$u' = T \cdot c = t(h) \cdot c \tag{7}$$

and

$$u' = \alpha\left[cu^2 + cr^2 + cur^* + cru^*\right] + ct_0 \tag{8}$$

with two relevant image terms $[cur^*]$ and $[cru^*]$, containing the object wave u and its conjugated version u^*, respectively. The appearance of the image terms depends on the concrete shape of the reconstruction wave. In general, the original reference wave

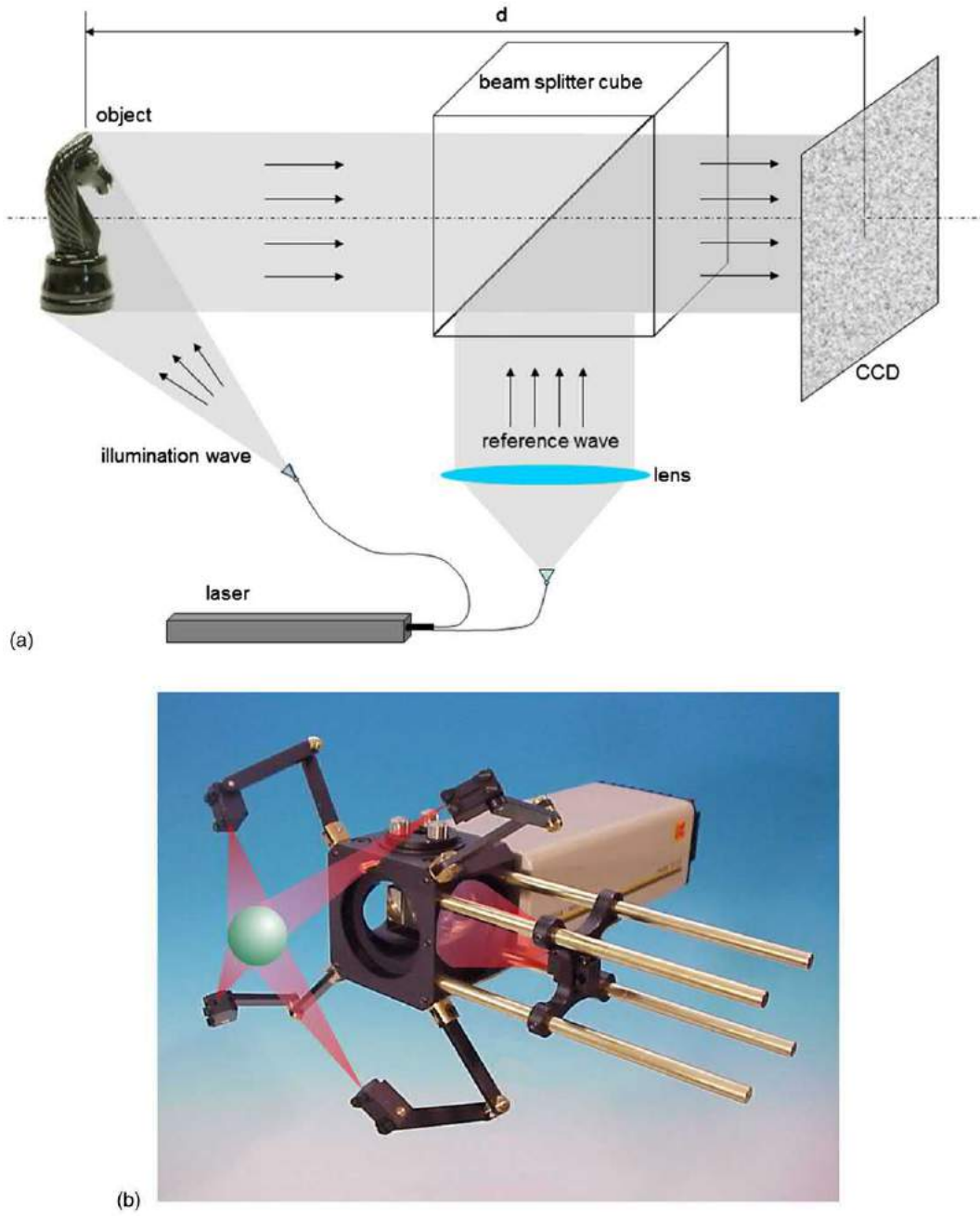


Fig. 1 Setup for recording digital holograms. (a) Schematic setup for recording a digital hologram onto a CCD target, (b) Implementation of a holographic camera by mounting a compact holographic interferometer in front of a CCD-target (four illumination directions are used). Osten, W., Seebacher, S., Jüptner, W. 2001. The application of digital holography for the inspection of micro-components. Proc. SPIE 4400, 1–15.

$c=r$ or its conjugated version $c=r^*$ is applied. In case of the conjugated reference wave a direct or real image will be reconstructed due to a converging image wave that can be imaged on a screen at the place of the original object.

However, in digital holography the reconstruction of the object wave in the image plane $u'(x',y')$ is done by numerical reproduction of the physical propagation process as shown in Fig. 3(a). The reconstruction wave, with a well-known shape equal to the reference wave $r(\xi,\eta)$, propagates through the hologram $h(\xi,\eta)$. Following the Huygens' principle each point $P(\xi,\eta)$ on the hologram acts as the origin of a spherical elementary wave. The intensity of these elementary waves is modulated by the transparency $h(\xi,\eta)$. In a given distance $d'=d$ from the hologram, a sharp real image of the object can be reconstructed as the superposition of all elementary waves. For the reconstruction of a virtual image, $d'=-d$ is used.

Consequently, the calculation of the wave field $u'(x',y')$ in the image plane starts with the pointwise multiplication of the stored and transformed intensity values $t[h(\xi,\eta)]$ with a numerical model of the reference wave $r(\xi,\eta)$. For a normally incident and

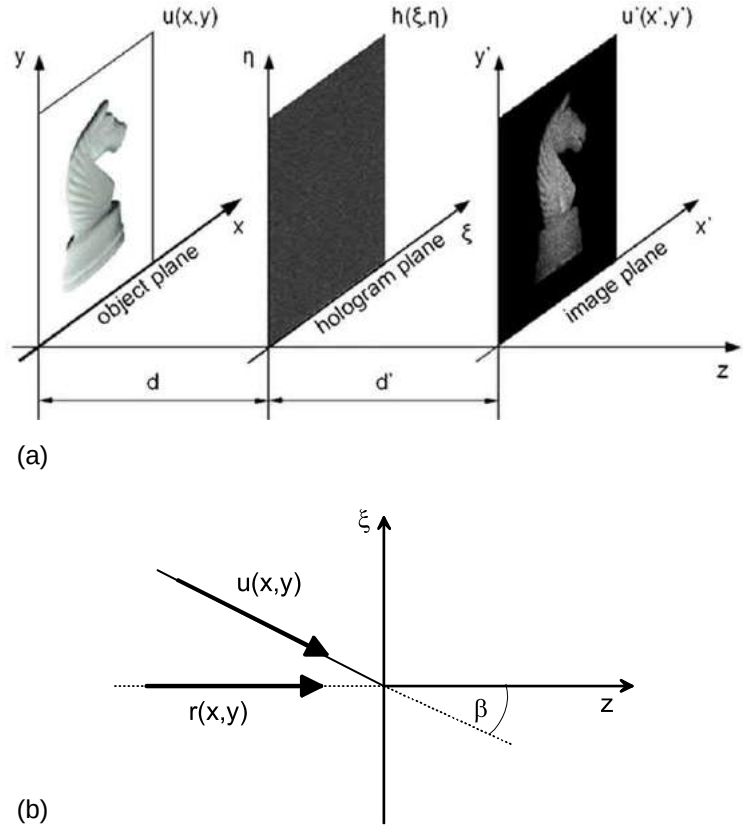


Fig. 2 Geometry for recording and reconstruction of digital holograms. (a) Schematic setup for Fresnel holography, (b) Interference between the object and reference wave in the hologram plane.

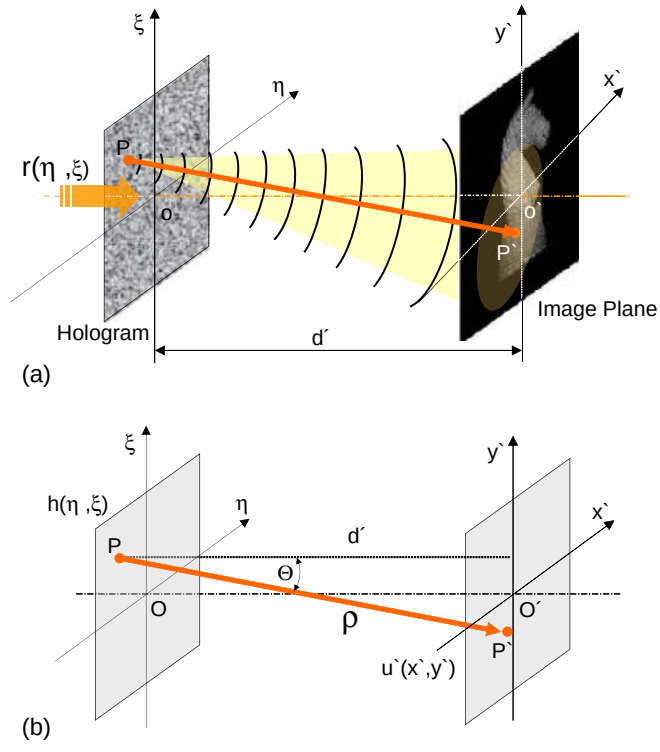


Fig. 3 Reconstruction of digital holograms. (a) Principle of wave front reconstruction, (b) light propagation by diffraction (Huygens-Fresnel principle).

monochromatic wave with unit amplitude, the reference wave can be modeled by $r(\xi,\eta)$. After the multiplication, the resulting field is propagated in free space with reference to the laws of e.g. nearfield diffraction. In the distance d′ the diffracted field $u'(x',y')$ can be found by solving the well-known Rayleigh–Sommerfeld diffraction formula, which is also known as the Huygens–Fresnel principle (Goodman, 1996):

$$u'(x',y') = \frac{1}{\mathrm{i}\lambda} \iint_{-\infty\ldots\infty} t[h(\xi,\eta)] \cdot r(\xi,\eta) \frac{\exp(\mathrm{i}k\rho)}{\rho} \cos\theta \mathrm{d}\xi\mathrm{d}\eta \tag{9}$$

with

$$\rho(\xi - x', \eta - y') = \sqrt{d'^2 + (\xi - x')^2 + (\eta - y')^2} \tag{10}$$

as the distance between a point $O'(x',y',z=d')$ in the image plane and a point $P(\xi,\eta,z=0)$ in the hologram plane and

$$k = \frac{2\pi}{\lambda} \tag{11}$$

as the wave number. The obliquity factor $\cos\theta$ represents the cosine of the angle between the outward normal and the vector joining P to O'. This term is given exactly by

$$\cos\theta = \frac{d'}{\rho} \tag{12}$$

and therefore Eq. (9) can be rewritten

$$u'(x',y') = \frac{\mathrm{d}'}{\mathrm{i}\lambda} \iint_{-\infty\ldots\infty} t[h(\xi,\eta)] \cdot r(\xi,\eta) \frac{\exp(\mathrm{i}k\rho)}{\rho^2} \mathrm{d}\xi\mathrm{d}\eta \tag{13}$$

The numerical reconstruction provides the complex amplitude of the wave front. Consequently, the phase distribution $\phi(x',y')$ and the intensity $I(x',y')$ can be calculated directly from the reconstructed complex function $u'(x',y')$:

$$\phi(x',y') = \arctan \frac{\mathrm{Im}|u'(x',y')|}{\mathrm{Re}|u'(\mathrm{x}',y')|} \quad [-\pi,\pi] \tag{14}$$

$$I(x',y') = u'(x',y') \cdot u'^{*}(x',y') \tag{15}$$

The direct approach to the phase yields several advantages for imaging and metrology applications that are discussed later.

Reconstruction Principles in Digital Holography

The theoretical background for the reconstruction of digital holograms is given by the scalar diffraction theory. Based on this, different numerical reconstruction principles are applied: the Fresnel approximation (Schnars, 1994), the convolution approach (Demetrakopoulos and Mittra, 1974), the angular spectrum method (Goodman, 1996), the lens-less Fourier approach (Takeda *et al.*, 1996; Wagner *et al.*, 1999), the phase shifting approach (Yamaguchi and Zhang, 1979), and the phase-retrieval approach (Zhang *et al.*, 2003). In this section the main techniques are briefly described.

The Fresnel Approximation

If the distance d between the object and hologram plane, and equivalently $d'=d$ between the hologram and image plane, is large compared to $(\xi–x')$ and $(\eta–y')$, so that the Fresnel approximation is fulfilled, then the distance ρ^2 in the denominator of Eq. (13) can be replaced by d'^2 and the parameter ρ in the numerator can be approximated by a binomial expansion for the square root in Eq. (10) where only the first two terms are considered (Goodman, 1996):

$$\rho \approx d' \left[1 + \frac{(\xi - x')^2}{2d'^2} + \frac{(\eta - x')^2}{2d'^2}\right] \tag{16}$$

The resulting expression for the field at (x',y') becomes

$$u'(x',y') = \frac{\exp(\mathrm{i}kd')}{\mathrm{i}\lambda \mathrm{d}'} \iint_{-\infty\ldots\infty} t[\mathrm{h}(\xi,\eta)] \cdot r(\xi,\eta) \exp\left[\frac{\mathrm{i}k}{2\mathrm{d}'}\left\{(\xi - x')^2 + (\eta - y')^2\right\}\right] \mathrm{d}\xi\mathrm{d}\eta \tag{17}$$

Eq. (17) is a convolution integral that can be expressed as

$$u'(x',y) = \iint_{-\infty\ldots\infty} t[h(\xi,\eta)] \cdot r(\xi,\eta) H(\xi - x', \eta - y') \mathrm{d}\xi\mathrm{d}\eta \tag{18}$$

with the convolution kernel

$$H(x',y') = \frac{\exp(\mathrm{i}kd')}{\mathrm{i}\lambda \mathrm{d}} \exp\left[\frac{\mathrm{i}k}{2\mathrm{d}}\left(x'^2 + y'^2\right)\right] \tag{19}$$

Another notation of Eq. (17) is found if the term $\exp\left[\frac{ik}{2d'}(x'^2+y'^2)\right]$ is taken out of the integral:

$$u'(x',y') = \frac{\exp(ikd')}{i\lambda d'} e^{i\frac{k}{2d'}(x'^2+y'^2)} \iint_{-\infty\ldots\infty} \left\{t[h(\xi,\eta)]\cdot r(\xi,\eta)\cdot e^{i\frac{k}{2d'}(\xi^2+\eta^2)}\right\} \exp\left[-i\frac{2\pi}{\lambda d'}(\xi x'+\eta y')\right] d\xi d\eta \tag{20}$$

or

$$u'(x',y') = \frac{\exp(ikd')}{i\lambda d'} e^{i\frac{k}{2d'}(x'^2+y'^2)} \mathbf{FT}_{\lambda d'}\left\{t[h(\xi,\eta)]\cdot r(\xi,\eta)\cdot e^{i\frac{k}{2d'}(\xi^2+\eta^2)}\right\} \tag{21}$$

where $\mathbf{FT}_{\lambda d'}$ indicates the 2D-Fourier transform that has been modified by a factor $1/(\lambda d')$. Eq. (21) makes clear that the diffracted wave front u' results from the Fourier transform **FT** of the digitally stored hologram $t[h(\xi,\eta)]$ multiplied by the reference wave, $r(\xi,\eta)$, and a quadratic phase factor, the so-called the chirp function $\exp\left\{\frac{i\pi}{\lambda d'}(\xi^2+\eta^2)\right\}$. This Fourier transform is scaled by the constant factor $1/(id'\lambda)$ and a phase factor that is independent of the processed hologram.

The discrete sampling in digital holography with the discrete pixel coordinates (m,n) requires the transformation of the infinite continuous integral in Eq. (20) into a finite discrete sum. This results in the finite Fresnel transform:

$$u'(m,n) = \sum_{j=0}^{M-1}\sum_{l=0}^{N-1} t[h(j\cdot\Delta\xi, l\cdot\Delta\eta)]\cdot r(j\cdot\Delta\xi, l\cdot\Delta\eta)\exp\left[\frac{i\pi}{d'\lambda}\left(j^2\Delta\xi^2+l^2\Delta\eta^2\right)\right] \times \exp\left\{-i2\pi\left(\frac{j\cdot m}{M}+\frac{l\cdot n}{N}\right)\right\} \tag{22}$$

where constants and pure phase factors preceding the sums have been omitted. The main parameters are the pixel number MxN and the pixel pitches $\Delta\xi$ and $\Delta\eta$ in the two directions which are defined by the used sensor chip.

The discrete phase distribution ϕ(m,n) of the wave front and the discrete intensity distribution $I(m,n)$ on the rectangular grid of MxN sample points can be calculated from the reconstructed complex function $u'(m,n)$ by using Eqs. (14) and (15), respectively. The wrapped phase distribution is computed directly without the need of additional temporal or spatial phase modulation like phase shifting or spatial heterodyning (Creath, 1993; Kujawinska, 1993). **Fig. 4** shows the result of the numerical reconstruction of the intensity and the phase of a tin soldier using a setup of the kind shown in **Fig. 1(a)**.

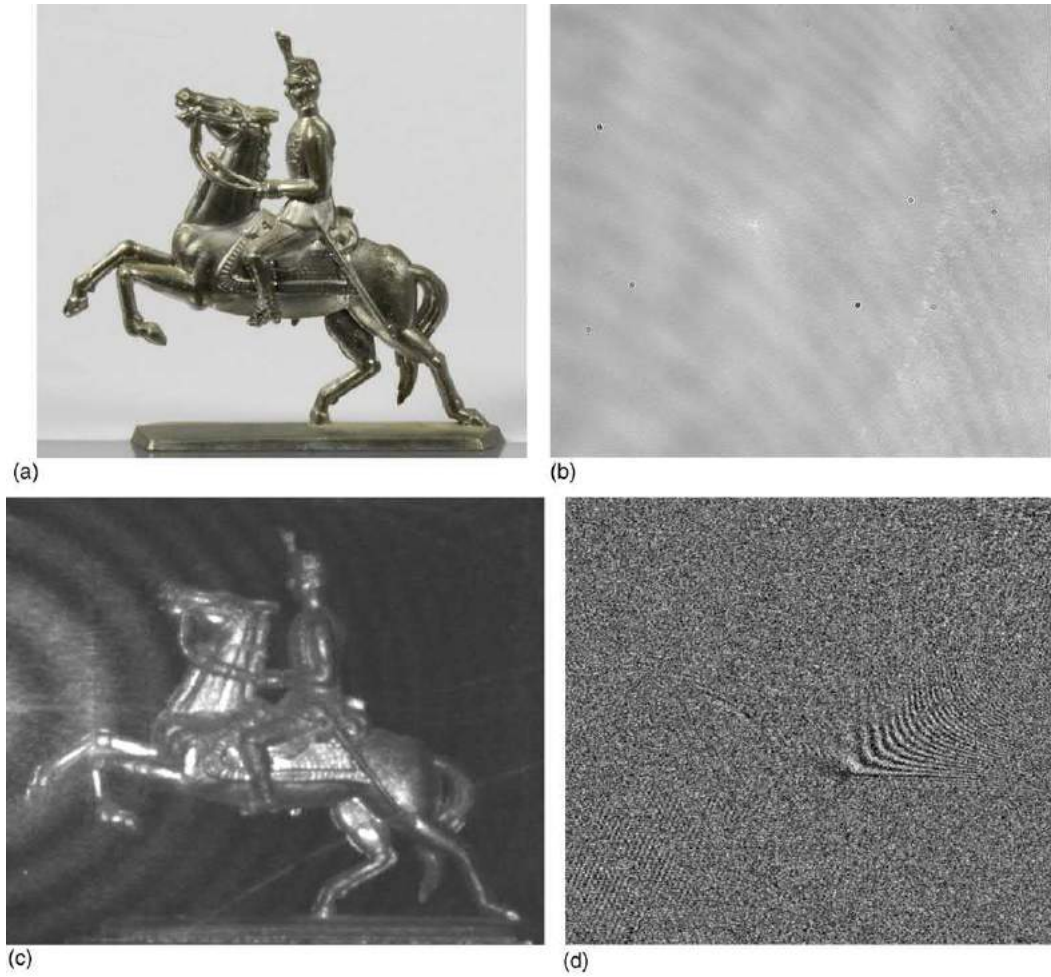


Fig. 4 (a) Tin soldier, (b) digital hologram, (c) Reconstructed Intensity $I(m, n)$, (d) Reconstructed phase distribution ϕ(m,n).

For interferometric applications with respect to displacement and shape measurement (Kreis, 2005; Osten and Ferraro, 2007) the phase difference δ(m,n) of the two reconstructed wave fields u_1'(m,n) and u_2'(m,n) corresponding to the two states of the object needs to be calculated. The algorithm can be reduced to the following equation:

$$\delta(m,n) = \phi_1(m,n) - \phi_2(m,n) \tag{23}$$

Fig. 5 shows the principle of displacement measurement by digital holography. Besides the investigation of surface displacements in the region of the wavelength, digital holography can be applied advantageously for the measurement of the shape of complex objects. In **Fig. 6** a turbine blade was investigated with the 2-wavelength contouring method (Osten *et al.*, 1998). Both, the digitally reconstructed mod 2π-phase and its demodulated version after phase unwrapping are presented (Wagner *et al.*, 2000). Phase unwrapping is necessary in digital holography since the fringe-counting problem still remains. However, there are some efficient approaches to overcome the difficulties of that process (Wagner *et al.*, 2000; Osten and Kujawinska, 2000).

Following Eq. (22), the pixel sizes in the reconstructed image along the two directions are

$$\Delta m = \frac{d'\lambda}{M\Delta\xi} \text{ and } \Delta n = \frac{d'\lambda}{N\Delta\eta} \tag{24}$$

In addition to the real image, a blurred virtual image and a bright dc-term, the zero-order diffracted field, are reconstructed. The dc-term can effectively be eliminated by preprocessing the stored hologram (Kreis and Jüptner, 1997; Seebacher *et al.*, 1997), and the different terms can be separated by using the off-axis instead of the in-line scheme. However, the spatial separation between the object and the reference field requires a sufficient space-bandwidth product of the used CCD-chip, as discussed in the section on direct phase reconstruction above.

Numerical Reconstruction by the Convolution Approach

In Section "The Fresnel approximation" we have already mentioned that the connection between the image term $u'(x',y')$ and the product $t[h(\xi,\eta)] \cdot r(\xi,\eta)$ can be described by a linear space-invariant system. The diffraction formula (17) is a superposition integral with the impulse response

$$H(x'-\xi, y'-\eta) = \frac{\exp(ikd')}{i\lambda d'} \exp\left[\frac{ik}{2d'}\left\{(\xi-x')^2 + (\eta-y')^2\right\}\right] \tag{25}$$

A linear space-invariant system is characterized by a transfer function G that can be calculated as the Fourier-transform of the impulse response:

$$G(f_x, f_y) = \mathbf{FT}\{H(\xi,\eta,x',y')\} \tag{26}$$

with the spatial frequencies f_x and f_y. Consequently, the convolution theorem can be applied, which states that the Fourier transform of the convolution $t[h(\xi,\eta)] \cdot r(\xi,\eta)$ with H is the product of the individual Fourier transforms $\mathbf{FT}\{t[h(\xi,\eta)] \cdot r(\xi,\eta)\}$ and $\mathbf{FT}\{H\}$. Thus $u'(x',y')$ can be calculated by the inverse Fourier transform of the product of the Fourier-transformed convolution partners:

$$u'(x',y') = \mathbf{FT}^{-1}\{\mathbf{FT}[t(h) \cdot r] \cdot \mathbf{FT}[H]\} \tag{27}$$

$$u'(x',y') = [t(h) \cdot r] \otimes H \tag{28}$$

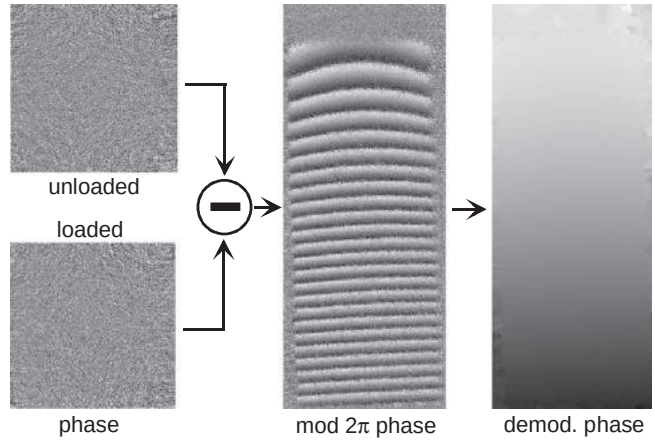


Fig. 5 The principle of digital holographic interferometry demonstrated on example of a loaded small beam. The interference phase is the result of the subtraction of the two phases which correspond to the digital holograms of the both object states.

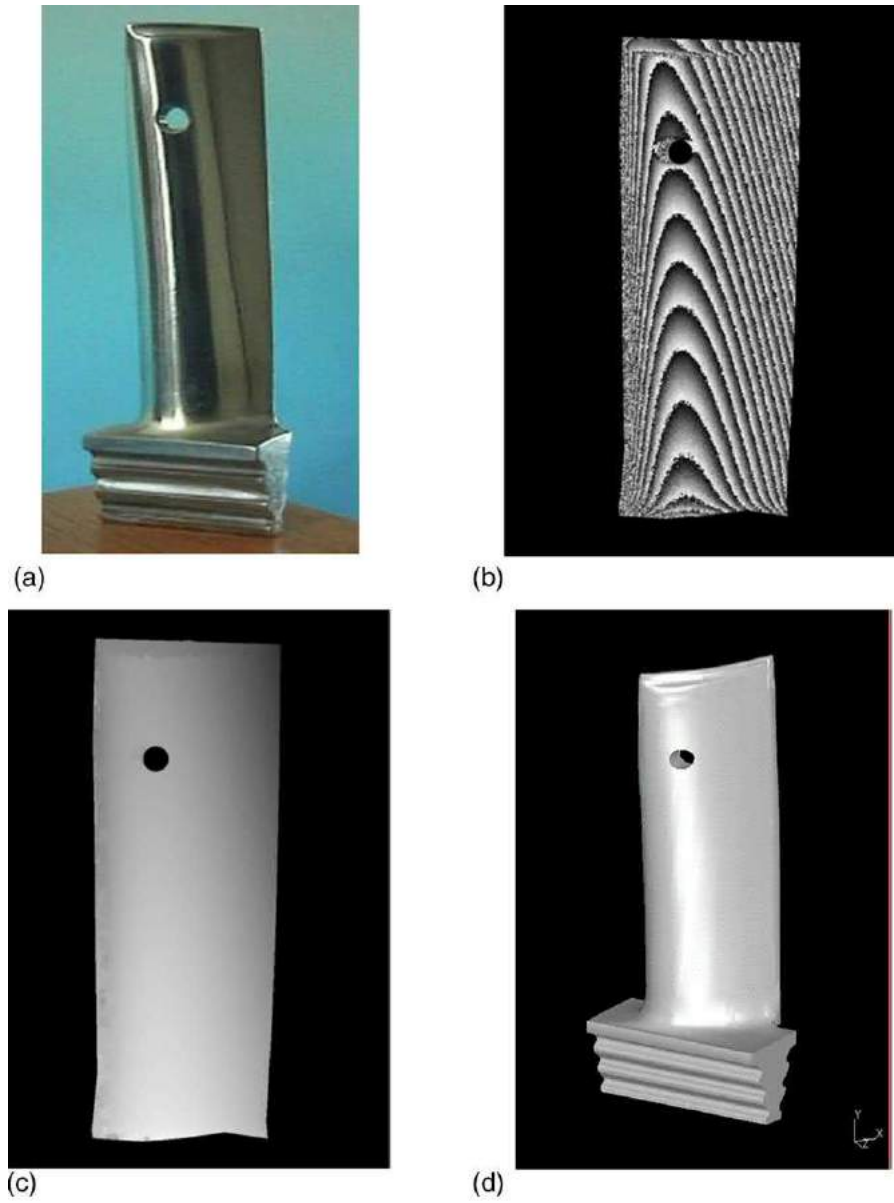


Fig. 6 Phase reconstruction by digital holographic interferometry on example of 2-wavelength contouring of a turbine blade. (a) image of the turbine blade, (b) mod-2π phase distribution, (c) unwrapped phase distribution, (d) CAD reconstruction of the turbine blade.

with $\otimes$ as the convolution symbol. The computing effort is comparatively high: two complex multiplications and three Fourier transforms. The essential difference compared to the Fresnel approximation is the different pixel size in the reconstructed field (Kreis, 2005). The pixel size is constant independent from the reconstruction distance d′, wavelength λ and pixel number m:

$$\Delta x' = \Delta\xi \text{ and } \Delta y' = \Delta\eta \tag{29}$$

Numerical Reconstruction by the Lensless Fourier Approach

We have already mentioned that the limited spatial resolution of the sensor restricts the angular resolution of the digital hologram. In fact, the sampling theorem requires that the angle between the object beam and the reference beam at any point of the electronic sensor be limited in such a way that the microinterference spacing is larger than double the pixel size. In general, the angle between the reference beam and the object beam varies over the sensor's surface, and so does the maximum spatial frequency. Thus for most holographic setups the full spatial bandwidth of the sensor cannot be used. However, it is very important to use the entire spatial bandwidth of the sensor because the lateral resolution of the reconstructed image depends on a complete evaluation of all the information one can get from the sensor.

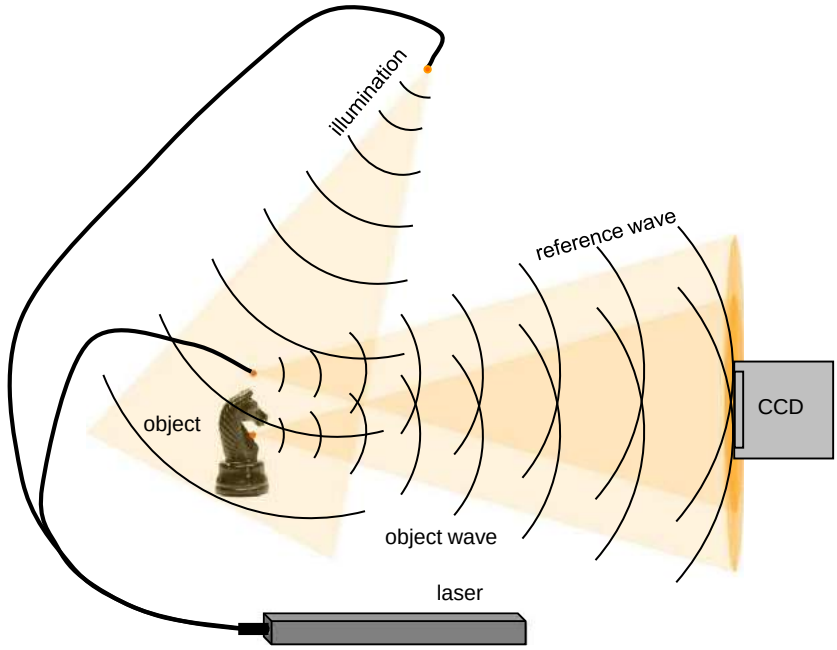


Fig. 7 Scheme of a setup for digital lensless Fourier holography of diffusely reflecting objects.

Even if the speed of digital signal processing is increasing rapidly, algorithms should be as simple and as fast to compute as possible. For the Fresnel approach and the convolution approach several fast Fourier transforms and complex multiplications are necessary. Therefore a more effective approach such as the subsequently described algorithm seems promising.

The lensless Fourier approach (Takeda *et al.*, 1996; Wagner *et al.*, 1999) is the fastest and most suitable algorithm for small objects. The corresponding setup is shown in **Fig. 7**. It allows us to choose the lateral resolution in a range from a few microns to hundreds of microns without any additional optics. Each point (ξ,η) on the hologram is again considered as a source point of a spherical elementary wave front (Huygen's principle). The intensity of these elementary waves is modulated by $t[h(\xi,\eta)]$ – the amplitude transmission of the hologram. The reconstruction algorithm for lensless Fourier holography is based on the Fresnel reconstruction. Here, again $u(x,y)$ is the object wave in the object plane, $h(\xi,\eta)$ the hologram, $r(\xi,\eta)$ the reference wave in the hologram plane and $u'(x',y')$ the reconstructed wave field.

For the specific setup of lensless Fourier holography a spherical reference wave is used with an origin in the same distance from the sensor as the object itself. In the case that $d' = -d$, $x = x'$, and $y = y'$ both virtual images are reconstructed and the reconstruction algorithm is then

$$u'(x',y') = \frac{\exp(\mathrm{i}kd)}{\mathrm{i}\lambda d} \mathrm{e}^{-\mathrm{i}\frac{k}{2d}(x^2+y^2)} \mathbf{FT}_{\lambda d'} \left\{ t[h(\xi,\eta)] \cdot r(\xi,\eta) \cdot \mathrm{e}^{-\mathrm{i}\frac{k}{2d}(\xi^2+\eta^2)} \right\} \tag{30}$$

where $\mathbf{FT}_{\lambda \mathrm{d}}$ is the two-dimensional Fourier transformation which has been scaled by a factor $1/(\lambda d)$. In this recording configuration, the effect of the spherical phase factor associated with the Fresnel diffraction pattern of the object is eliminated by the use of a spherical reference wave $r(\xi,\eta)$ with the same average curvature:

$$r(x,y) = \mathrm{const} \cdot \exp\left(\mathrm{i}\frac{\pi}{\lambda d}\left(\xi^2+\eta^2\right)\right) \tag{32}$$

This results in a more simple reconstruction algorithm which can be described by

$$u'(x',y') = \mathrm{const} \cdot \exp\left(-\mathrm{i}\frac{\pi}{\lambda d}\left(x^2+y^2\right)\right) \mathbf{FT}_{\lambda d}\{h(\xi,\eta)\} \tag{33}$$

Besides the faster reconstruction algorithm (only one Fourier transform has to be computed), the Fourier algorithm uses the full space-bandwidth product (SBP) of the sensor chip because it adapts the curvature of the reference wave to the curvature of the object wave.

Influences of Discretization

For the estimation of the lateral resolution in digital holography three different effects related to discretization have to be considered (Wagner *et al.*, 1999): averaging, sampling and the limited sensor size. We assume a quadratic sensor with *N*x*M* pixels of size $\Delta\xi$x$\Delta\xi$. Each pixel has a light sensitive region with a side length $\gamma\Delta\xi = \Delta\xi_{\mathrm{eff}}$. The quantity γ^2 is the so-called fill factor $(0 \le \gamma \le 1)$ that indicates the active area of the pixel. The sensor averages the incident light which has to be considered because of

possible intensity fluctuations over this area. The continuous expression for the intensity $I(\xi,\eta)=t[h(\xi,\eta)]$ registered by the sensor has to be integrated over the light sensitive area. This can be expressed mathematically by the convolution of the incident intensity $I(\xi,\eta)$ with a rectangle function (Goodman, 1996)

$$I^1(\xi,\eta) \propto I \otimes \mathrm{rect}_{\Delta\xi_{\mathrm{eff}},\Delta\xi_{\mathrm{eff}}} \tag{34}$$

The discrete sampling of the light field is modeled by the multiplication of the continuous assumed hologram with the two-dimensional comb-function:

$$I^2(\xi,\eta) \propto I^1(\xi,\eta) \cdot \mathrm{comb}_{\Delta\xi,\Delta\xi} \tag{35}$$

Finally, the limited sensor size requires that the comb-function has to be truncated at the borders of the sensor. This is considered by the multiplication with a two-dimensional rect-function of the size $Nx\Delta\xi$:

$$I^3(\xi,\eta) \propto I^2 \cdot \mathrm{rect}_{N\Delta\xi\cdot N\Delta\xi} \tag{36}$$

The consequences are amplitude distortion, aliasing and speckling that have to be considered in the reconstruction procedure (Wagner *et al.*, 1999; Seebacher, 2001).

Advantages of Digital Holography

Besides the electronic processing and the direct access to the phase some further advantages recommend digital holography for several applications such as metrology and microscopy (Osten, 2003).

- The availability of the independently reconstructed phase distributions of each individual state of the object and interferometer, respectively, offers the possibility to record a series of digital holograms with increased load amplitude. In the evaluation process the convenient states can be compared interferometrically. Furthermore, a series of digital holograms with increasing load can be applied to unwrap the mod2π-phase temporally (Wagner *et al.*, 2000). In this method the total object deformation is subdivided in many measurement steps in which the phase differences are smaller than 2π. By adding up those intermediate results, the total phase change can be obtained without any further unwrapping. This is an important feature, since it is essential to have an unwrapped phase to be able to calculate the real deformation data from the phase map. **Fig. 8** shows an example of such a measurement. The left image shows the wrapped deformation phase for a clamped coin which was loaded with heat. The right image shows the temporal unwrapped phase which has been obtained by dividing the total deformation into 85 sub-measurements. Thus the displacement of the object can be observed almost in real time during the loading process.
- The independent recording and reconstruction of all states gives also a new degree of freedom for optical shape measurement. In case of multi-wavelength contouring each hologram can be stored and reconstructed independently with its corresponding wavelength. This results in a significant decrease of aberrations and makes it possible to use larger wavelength differences for the generation of shorter synthetic wavelengths (Wagner *et al.*, 2000).
- Because all states of the inspected object can be stored and evaluated independently, only seven digital holograms are necessary to measure the 3D-displacement field and – shape of an object under test: one holograms for each illumination direction before and after the loading, respectively, and one hologram with a different wavelength (or a different source-point of illumination) which can interfere with one of the other holograms to do two-wavelength-contouring (or two-source-point-contouring) for shape measurement. If four illumination directions are used nine holograms are necessary (Seebacher *et al.*, 2001).
- The direct approach to the phase enables various new microscopic principles for the investigation of phase objects. They are categorized with the term Quantitative Phase Contrast Microscopy (Cuche *et al.*, 1999; Kim, 2011).

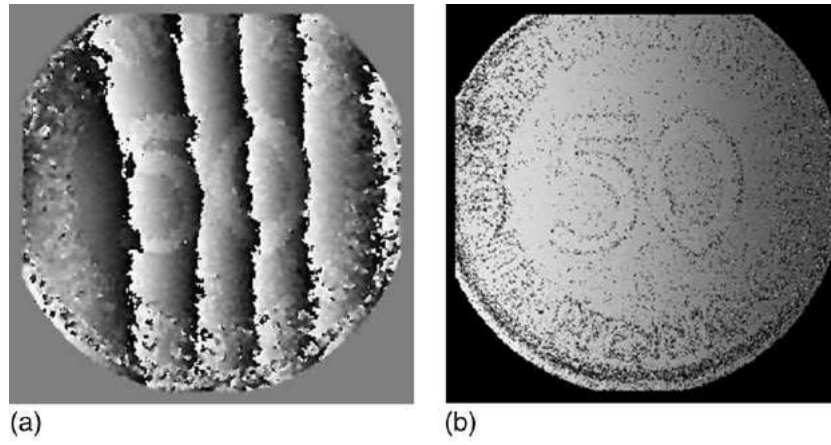


Fig. 8 Temporal unwrapping by digital holography. (a) wrapped phase of a thermal loaded coin, (b) temporally unwrapped phase.

- Digital holography also allows new approaches for the exploration and exploitation of scattering media (Singh *et al.*, 2017).
- Last but not least, the complex holographic sensor setups can be miniaturized drastically (Kolenovic *et al.*, 2003) and methods for remote comparative interferometry (Baumbach *et al.*, 2006) make digital holography to a versatile tool for the solution of numerous future inspection and measurement problems.

See also: Holography: Computer Generated Holograms

References

Baumbach, T., Osten, W., Kopylow, C., Jueptner, W., 2006. Remote metrology by comparative digital holography. Appl. Opt. 45, 925–934.

Butters, J.N., Leendertz, J.A., 1971. Holographic and video techniques applied to engineering measurement. J. Meas. Control 4, 349–354.

Coufal, H.J., Psaltis, D., Sincerbox, G.T., Glass, A.M., Cardillo, M.J. (Eds.), 2000. Holographic Data Storage. Berlin: Springer.

Creath, K., 1993. Temporal phase measurement methods. In: Robinson, D.W., Reid, G.T. (Eds.), Interferogram Analysis. Bristol and Philadelphia: IOP Publishing Ltd.

Cuche, E., Marquet, P., Depeursinge, C., 1999. Simultaneous amplitude-contrast and quantitative phase-contrast microscopy by numerical reconstruction of Fresnel off-axis holograms. Appl. Opt. 38 (34), 6994–7001.

DaneshPanah, M., Zwick, S., Schaal, F., *et al.*, 2010. 3D Holographic imaging and trapping for non-invasive cell identification and tracking. J. Display Technol. 6, 400–409.

Demetrakopoulos, T.H., Mitra, R., 1974. Digital and optical reconstruction of images from suboptical patterns. Appl. Opt. 133, 665–670.

Demetrakopoulos, T., Mittra, T., 1974. Digital and optical reconstruction of images from suboptical diffraction patterns. Appl. Opt. 13, 665–670.

Denisyuk, Yu.N., 1962. Photographic reconstruction of the optical properties of an object in its own scattered radiation field. Dokl. Akad. Nauk SSSR 144, 1275–1279.

Ferraro, P., De Nicola, S., Finizio, A., *et al.*, 2003. Compensation of the inherent wave front curvature in digital holographic coherent microscopy for quantitative phase-contrast imaging. Appl. Opt. 42, 1938–1946.

Frauel, Y., Naughton, T.J., Matoba, O., Tajahuerce, E., Javidi, B., 2006. Three-dimensional imaging and processing using computational holographic imaging. Proc. IEEE 94 (3), 636–653.

Gabor, D., 1949. Microscopy by reconstructed wave-fronts. Proc. Royal Soc., A 197, 454–487.

Gang, J., 2013. Three-dimensional display technologies. Adv. Opt. Photon. 5 (4), 456–535.

Goodman, J.W., 1996. Introduction to Fourier Optics. New York: McGraw Hill Comp. Inc..

Goodman, J.W., Lawrence, R.W., 1967. Digital image formation from electronically detected holograms. Appl. Phys. Lett. 11, 77–79.

Haist, T., Hasler, M., Osten, W., Baranek, M., 2014. Programmable Microscopy. In: Javidi, B., Tajahuerce, E., Andrés, P. (Eds.), Multidimensional Imaging. New York: John Wiley & Sons, pp. 153–174.

Haist, T., Osten, W., 2015a. Holography using pixelated spatial light modulators – Part 1: Theory and basic considerations. J. Micro/Nanolith MEMS MOEMS 14 (4), 041310.

Haist, T., Osten, W., 2015b. Holography using pixelated spatial light modulators – Part 2: Applications. J. Micro/Nanolith MEMS MOEMS 14 (4), 041311.

HAMAMTSU Photonics K.K., http://www.hamamatsu.com/jp/en/technology/innovation/lcos-slm/index.html.

Hariharan, P., 1984. Optical Holography. Cambridge: Cambridge University Press.

Hasegawa, S., Hayasaki, Y., Nishida, N., 2006. Holographic femtosecond laser processing with multiplexed phase Fresnel lenses. Opt. Lett. 31, 1705–1707.

HOLOEYE Photonic AG, http://holoeye.com/.

Huang, T., 1971. Digital Holography. Proc. IEEE 599, 1335–1346.

Kim, M.K., 2011. Digital Holographic Microscopy. Berlin: Springer.

Kohler, C., Schwab, X., Osten, W., 2006. Optimally tuned spatial light modulators for digital holography. Appl. Opt. 45, 960–967.

Kolenovic, E., Osten, W., Klattenhoff, R., *et al.*, 2003. Miniaturized digital holography sensor for distal three-dimensional endoscopy. Appl. Opt. 42 (25), 5167–5172.

Kreis, T., 2005. Handbook of Holographic Interferometry – Optical and Digital Methods. Weinheim: Wiley-VCH.

Kreis, Th, Jüptner, W., 1997. The suppression of the dc-term in digital holography. Opt. Eng. 36, 2357–2360.

Kronrod, M.A., Merzlyakov, N.S., Yaroslavsky, L.P., 1972. Reconstruction of holograms with a computer. Sov. Phys. Tech. Phys. 17, 333–334.

Kujawinska, M., 1993. Spatial phase measurement methods. In: Robinson, D.W., Reid, G.T. (Eds.), Interferogram Analysis.. Bristol and Philadelphia: IOP Publishing Ltd.

Lazarev, G., Hermerschmidt, A., Krüger, S., Osten, S., 2012. LCOS Spatial Light Modulators: Trends and Applications. In: Osten, W., Reingand, N. (Eds.), Optical Imaging and Metrology: Advanced Technologies. Weinheim: Wiley-VCH, pp. 1–29.

Lee, W.-H., 1978. Computer generated holograms: Techniques and applications. In: Wolf, E. (Ed.) Progress in Optics, XVI. Amsterdam: North-Holland, pp. 121–230.

Lee, B., Kim, Y., 2012. Three-dimensional display and imaging: Status and prospects. In: Osten, W., Reingand, N. (Eds.), Optical Imaging and Metrology: Advanced Technologies. Weinheim: Wiley-VCH, pp. 31–56.

Leith, E.N., Upatnieks, J., 1961. New techniques in wavefront reconstruction. JOSA 51, 1469–1473.

Liesener, J., Reicherter, M., Tiziani, H.J., 2004. Determination and compensation of aberrations using SLMs. Opt. Commun. 233, 161–166.

Marquet, P., Depeursinge, C., 2014. Digital Holographic Microscopy: a New Imaging Technique to Quantitatively Explore Cell Dynamics with Nanometer Sensitivity. In: Javidi, B., Tajahuerce, E., Andrés, P. (Eds.), Multidimensional Imaging. New York: John Wiley & Sons, pp. 197–212.

Maurer, C., Jesacher, A., Bernet, S., Ritsch-Marte, M., 2010. What spatial light modulators can do for optical microscopy. Laser Photonics Rev. 5, 81–101.

Onural, L., Yaras, F., Kang, H., 2011. Digital holographic three-dimensional video displays. Proc. IEEE 99 (4), 576–589.

Osten, W., 2003. Active metrology by digital holography. Proc. SPIE 4933, 96–110.

Osten, W., Baumbach, Th., Jueptner, W., 2002. Comparative Digital Holography. Opt. Letts. 27, 1764–1766.

Osten, W., Faridian, A., Gao, P., *et al.*, 2014. Recent advances in digital holography. Appl. Opt 53 (27), G44–G63. doi:10.1364/AO.53.000G44.

Osten, W., Ferraro, P., 2007. Digital Holography and IST application in MEMS/MOEMS inspection. In: Osten, W. (Ed.), Optical Inspection of Microsystems. Boca Raton, FL: Taylor and Francis, pp. 351–425.

Osten, W., Jüptner, W., Seebacher, S., 1998. The qualification of large scale approved measurement techniques for the testing of microcomponents. Proceedings 18th Symposium on Seebacher, Experimental Mechanics of Solids, Jachranka, pp. 43–55.

Osten, W., Kujawinska, M., 2000. Active phase measuring metrology. In: Rastogi, P.K., Inaudi, D. (Eds.), Trends in Optical Nondestructive Testing and Inspection. Elsevier Science B.V, pp. 45–69.

Osten, W., Wilke, M., Pedrini, G., 2013. Remote laboratories for optical metrology: From the lab to the cloud. Opt. Engn. 52 (10), 101914. [Article Number: 101914].

Powell, R.L., Stetson, K., 1965. Interferometric vibration analysis of three-dimensional objects by wavefront reconstruction. JOSA 55, 1593–1598.

Reicherter, M., Liesener, J., Haist, T., Tiziani, H.J., 1999. Optical particle trapping with computer-generated holograms written in a liquid crystal display. Opt. Lett. 9, 508–510.

Schnars, U., Jüptner, W., 1994. Direct recording of holograms by a CCD target and numerical reconstruction. Appl. Opt. 33 (2), 179–181.

Schnars, U., 1994. Direct phase determination in hologram interferometry with use of digitally recorded holograms. J. Opt. Soc. Am. A 11, 2011–2015.

Seebacher, S., Osten, W., Baumbach, Th., Jüptner, W., 2001. The determination of material parameters of microcomponents using digital holography. Opt. Las Eng. 36 (2), 103–126.
Seebacher, S., Osten, W., Jüptner, W., 1997. 3-D deformation analysis of micro-components using digital holography. Proc. SPIE 3098, 382–391.
Seebacher, S., 2001. Application of digital holography for 3d-shape and deformation measurement of micro components. PhD Thesis, University of Bremen.
Singh, A., Pedrini, G., Takeda, M., Osten, W., 2017. Exploiting scattering media for exploring 3D objects. Light: Science & Applications 6, e16219. doi:10.1038/lsa.2016.219.
Steel, W.H., 1983. Interferometry. Cambridge: Cambridge University Press.
Sutkowski, M., Kujawinska, M., 2000. Application of liquid crystal (LC) devices for optoelectronic reconstruction of digitally stored holograms. Opt. Laser Eng. 33, 191–201.
Takeda, M., Taniguchi, K., Hirayama, T., Kogho, H., 1996. Single transform Fourier-Hartley fringe analysis for holographic interferometry. In: Füzessy, Z., Jüptner, W., Osten, W. (Eds.), Simulation and Experiment in Laser Metrology. Berlin: Akademie Verlag, pp. 67–73.
Wagner, C., Osten, W., Seebacher, S., 2000. Direct shape measurement by digital wavefront reconstruction and multi-wavelength contouring. Opt. Eng. 39 (1), 79–85.
Wagner, C., Seebacher, S., Osten, W., Jüptner, W., 1999. Digital recording and numerical reconstruction of lensless Fourier holograms in optical metrology. Appl. Opt. 38 (22), 4812–4820.
Yamaguchi, I., Zhang, T., 1979. Phase-shifting digital holography. Opt. Lett. 22, 1268–1270.
Yaroslavsky, L., 2004. Digital Holography and Digital Image Processing: Principles, Methods, Algorithms. Dordrecht: Kluwer Academic Publishers.
Yu, X., Hong, J., Liu, C., Kim, M.K., 2014. Review of digital holographic microscopy for three-dimensional profiling and tracking. Opt. Eng. 53, 112306.
Zhang, Y., Pedrini, G., Osten, W., Tiziani, H., 2003. Image reconstruction for in-line holography with the Yang-Gu algorithm. Appl. Opt. 42 (32), 6452–6457.
Zwick, S., Haist, T., Warber, M., Osten, W., 2010. Dynamic holography using pixelated light modulators. Appl. Opt. 49, F47–F58.

Overview: Holography

C Shakher and AK Ghatak, Indian Institute of Technology, New Delhi, India

Major Milestones

1948: Essential concept for holographic recording by Dennis Gabor.
1960: The first successful operation of a laser device by Theodore Maiman.
1962: Off-axis technique of holography by Leith and Upatnieks.
1962: Yu N Denisyuk suggested the idea of three-dimensional holograms based on thick photoemulsion layers. His holograms can be reconstructed in ordinary sunlight. These holograms are called Lippmann–Bragg holograms.
1964: Leith and Upatnieks pointed out that a multicolor image can be produced by a hologram recorded with three suitably chosen wavelengths.
1969: S A Benton invented 'Rainbow Holography' for display of holograms in white light. This was a vital step to make holography suitable for display applications.

Holography is the science of recording an entire optical wavefront, both amplitude and phase information, on appropriate recording material. The record is called a hologram. Unlike conventional photography, which records a three-dimensional scene in a two-dimensional format, holography records true three-dimensional information about the scene.

Holography was invented by Gabor in 1948 and his first paper introduced the essential concept for holographic recording – the reference beam. Holography is based on the interference between waves and, it provides us with a way of storing all the light information arriving at the film in such a way that it can be regenerated later.

The use of the reference beam is utilized because the physical detectors and recorders are sensitive only to light intensity. The phase is not recorded but is manifest only when two coherent waves of the same frequency are simultaneously present at the same location. In that case, the waves combine to form a single wave whose intensity depends not only on intensities of the two individual waves, but also on the phase difference between them. This is key to holography. The film record, or hologram, can be considered as a complicated diffraction grating. Holograms bear no resemblance to conventional photographs in that an image is not actually recorded. In fact, the interferometric fringes which are recorded on the recording material are not visible to an unaided eye because of extremely fine interfringe spacing ($\sim$0.5 micrometer). The fringes which are visible on the recording material are the result of dust particles in the optical system used to produce the hologram.

Gabor's original technique is now known as in-line holography. In this arrangement, the coherent light source as well as the object, which is a transparency containing small opaque details on a clear background, is located along the axis normal to the photographic plate. With such a system, an observer focusing on one image observes it superposed on the out-of-focus twin image as well as a strong coherent background. This constitutes the most serious problem of Gabor's original technique.

The first successful technique for separating the twin images was developed by Leith and Upatnieks in 1962. This is called the off-axis or side band technique of holography. We shall consider mainly the off-axis technique here.

Basic Holography Principle

Light from a laser or any light beam is characterized by spatial coherence (l_c) and temporal coherence (τ_c), which is discussed in all textbooks on optics. For typical laboratory hologram recording, the required degree of coherence of laser radiation is determined by the type of object and geometrical arrangement being used for the recording. The following condition must hold:

$$L \ll c\tau_c \tag{1}$$

where τ_c is the coherence time, and L the maximum path length difference between the two waves chosen to record the hologram. To begin recording, two wavefronts are derived from the same laser source. One wavefront is used to illuminate the object and another is used as a reference wavefront. The wavefront derived by illuminating the object is superimposed with the reference wave and the interference pattern is recorded. These object and reference waves must:

- be coherent (derived from the same laser); and
- have a fixed relative phase at each point on the recording medium.

If the above conditions are not met, the fringes will move during the exposure and the holographic record will be smeared. Therefore, to record the hologram one should avoid air currents and vibrations. The recording medium must have sufficient resolution to record the fine details in the fringe pattern. This will be typically of the order of the wavelength. To create a three-dimensional image from the holographic process one then has to record and reconstruct the hologram. The object and reference waves are derived from the same laser.

Holograms record an interference pattern formed by interference of object and reference wavefronts, as explained above. We may mention here that for the two plane waves propagating at an angle θ between them, the spacing of the interference fringes is given by

$$\alpha = \frac{\lambda_0}{2\sin(\theta/2)} \tag{2}$$

where λ_0 is the free space wavelength.

Processing of the hologram (developing, fixing, and washing of the recording material) yields a plate with alternating transparent and opaque parts, variation of refractive index, or variation of height corresponding to intensity variation in the fringe pattern. Such a plate can be regarded as a complicated diffraction grating. In this process, the hologram is illuminated with monochromatic, coherent light. The hologram diffracts this light into wavefronts which are essentially indistinguishable from the original waves which were diffracted from the object. These diffracted waves produce all the optical phenomena that can be produced by the original waves. They can be collected by a lens and brought into focus, thereby forming an image of the original object, even though the object has since been removed. If the reconstructed waves are intercepted by the eye of an observer, the effect is exactly as if the original waves were being observed; the observer sees what appears to be the original object in true three-dimensional form. As the observer changes his viewing position, the perspective of the image scene changes; parallactic effects are evident and the observer must refocus when the observation point is changed from a near to a distant object in the scene. Assuming that both the construction and reconstruction of the hologram are made with the same monochromatic light source, there is no visual test which can be made to distinguish between the real object and the reconstructed image of the object. It is as if the hologram were a window through which the apparent object is viewed.

Hologram of a Point Object

Consider a hologram recorded with a collimated reference wave normal to the recording plate and a point object inclined at a certain angle. If the hologram is illuminated once again with the same collimated reference wave, it reconstructs two images, one virtual true image and the other real image. However, the two images differ in one very important respect.

While the virtual image is located in the same position as the object and exhibits the same parallax properties, the real image is formed at the same distance from the hologram but in front of it. Corresponding points on the real and virtual images are located at equal distances from the plane of the hologram; the real image has the curious property that its depth is inverted. Such an image is not formed with a normal optical system; it is therefore called a pseudo image as opposed to a normal or orthoscopic image.

This depth inversion results in conflicting visual clues, which make viewing of the real image psychologically unsatisfactory. Thus, if O_1 and O_2 are two elements in the object field, and if O_1 blocks the light scattered by O_2 at a certain angle, the hologram records information only on the element O_1 at this angle and records no information about this part of O_2. An observer viewing the real image from the corresponding direction then cannot see this part of O_2, which, contrary to normal experience, is obscured by O_1, even though O_2 is in front of O_1.

Production of an Orthoscopic Real Image

An orthoscopic real image of an object can be produced by recording two holograms in succession. In the first step, a hologram is recorded of the object with a collimated reference beam. When this hologram is illuminated once again with the collimated reference beam that was used to record it, it reconstructs two images of the object at unit magnification, one of them being an orthoscopic virtual image, while the other is a pseudoscopic real image. A second hologram is then recorded of this real image with a second collimated reference beam.

When the second hologram is illuminated with a collimated beam it reconstructs a pseudoscopic virtual image located in the same position as the real image formed by the second hologram is an orthoscopic image. Since a collimated reference beam is used throughout, the final real image is the same size as the original object and free from aberrations.

In addition to these characteristic linked intensities with the three-dimensional nature of the reconstruction, the holographic recording has several other properties. Each portion of the hologram can reproduce the entire image scene. If a smaller and smaller portion of the hologram is used for reconstruction, there is loss of image intensity and reconstruction. When a hologram is reversed, such as in contact printing processes, it will still reconstruct a positive image indistinguishable from the image produced by the original hologram.

Simple Mathematical Description of Holography

Let the object and reference waves be given by

$$U_o = O(x,y)e^{i\phi(x,y)} \tag{3}$$

and

$$U_r = R\, e^{ik\gamma \sin\theta} \tag{4}$$

U_r describes the reference (plane wave) propagating at angle θ to the z-axis.

The intensity at the hologram plane is

$$\begin{aligned} I &= |U_r + U_o|^2 \\ &= |U_r|^2 + |U_o|^2 + U_r^* U_o + U_r U_o^* \end{aligned} \tag{5}$$

As can be seen from the above equation, the amplitude and phase of the wave are encoded as the amplitude and phase modulation of a set of interference fringes. The material used to record the patterns of interference fringes described above is assumed to provide linear mapping of the intensity incident during the reconstruction process into amplitude transmitted by or reflected from the recorded material. Usually both light detection and wavefront modulation are performed by photographic plate/film. We assume the amplitude transmission properties of the plate/film after processing to be described by

$$T = T_o - b(E - E_o) \tag{6}$$

where the exposure E, at the film is $E = I\tau$; here τ is the exposure time. T_0 is the transmittance of the unexposed plate.

If the hologram is reilluminated by the reference wave, the transmitted wave amplitude will be

$$U_t = U_r(T_o - b\tau(|U_o|^2 + U_r^* U_o + U_r U_o^*)) \tag{7}$$

The first term is reference wave times a constant. The second term is the reference wave modulated by $|U_o|^2 = O(x, y)^2$; it gives small-angle scattering about the reference wave direction. The third term is proportional to U_o It is the same as the original object wave (note this is only so if the reconstruction wave is identical to the reference wave). The fourth term is

$$-b\tau U_r^2 U_0^* = -b\tau R^2 e^{ik\gamma 2\sin\theta} O(x, y) e^{-i\phi(x,y)} \tag{8}$$

This is essentially a wave traveling in a direction $\sin^{-1}(2\sin\theta)$ to the z-axis, with the correct object amplitude modulation but its phase reversed in sign, producing a conjugate wave.

Types of Holograms

Primarily, the holograms are classified as thin and thick, based on the thickness of the recording medium. When the thickness of the recording medium is small compared with the average spacing of the interference fringes, the hologram can be treated as a thin hologram. Such holograms are characterized by spatially varying complex amplitude transmittance:

$$\mathbf{t}(x, y) = [\mathbf{t}(x, y)]\exp[-i\phi(x, y)] \tag{9}$$

Thin holograms can be further categorized as thin amplitude holograms or thin phase holograms. If amplitude transmittance of the hologram is such that $\phi(x, y)$ is constant while $\mathbf{t}(x, y)$ varies over the hologram, the hologram is termed an amplitude hologram. For a lossless phase hologram, $|\mathbf{t}(x, y)| = 1$, so that the complex amplitude transmittance is caused by variation in phase.

When the thickness of the hologram recording material is large compared to the fringe spacing of the interference fringes, the holograms may be considered as volume holograms. These may be treated as a three-dimensional system of layers corresponding to a periodic variation of absorption coefficient or refractive index, and the diffracted amplitude is at a maximum when Bragg's condition is satisfied. In general, the behavior of thin and thick holograms corresponds to Raman Nath and Bragg diffraction regimes. The distinctions between two regimes is made on the basis of a parameter $\mathcal{Q}$, which is defined by the relation:

$$\mathcal{Q} = \frac{2\pi\lambda_o d}{n_o \Lambda^2} \tag{10}$$

where

Λ = spacing of fringe on hologram, measured normal to the surface;
d = thickness of recording medium;
n_o = mean refractive index of the medium; and
λ_o = wavelength of light.

Small values of $\mathcal{Q}$ ($\mathcal{Q} < 1$) correspond to thin gratings, while large values of $\mathcal{Q}$ ($\mathcal{Q} > 10$) correspond to volume gratings. For values of $\mathcal{Q}$ between 1 and 10, intermediate properties are exhibited. However, this criterion is not always adequate. The boundaries between the thin and thick holograms are given by the value of

$$P = \frac{\lambda_o^2}{\Lambda^2 n_o n_1} \tag{11}$$

where P^{-2} is the relative power diffracted into higher orders, and $P < 1$ for thin holograms and $P > 10$ for thick holograms. For P having values between 1 and 10, the hologram may be thin or thick, depending on the other parameters involved.

Holograms can also be classified based on whether they are reconstructed in transmitted light or reflected light. For transmission holograms, during the recording stage, two interfering wavefronts make equal but opposite angles to the surfaces of

recording medium and are incident on it from the same side. Reflection holograms reconstruct images in reflected light. They are recorded such that the interfering wavefronts are symmetrical with respect to the surface of the recording medium but are incident on it from opposite sides. When the angle between the interfering wavefronts is maximum (180°), the spacing between the fringe planes of the recording medium is minimum. Under such conditions, reflection holograms may have wavelength sensitivity high enough to be reconstructed, even with white light.

An important development in the field of holography was the invention of rainbow holograms by Benton in 1969. This provides a method for utilizing white light for illumination when viewing the holograms. The technique does so by minimizing the blur introduced by color dispersion in transmission holograms, at the price of giving up parallax information in one dimension.

Recording Materials

An ideal recording medium for holography should have a well matching spectral sensitivity corresponding to available laser wavelengths, linear transfer characteristics, high resolution, and low noise. It should also be either inexpensive or indefinitely recyclable. Toward achieving the above-mentioned properties, several materials have been studied, but none has been found so far that meets all the requirements. Materials investigated include: silver halide photographic emulsion; dichromated gelatin plates/films; silver halide sensitized gelatin plates/films; photo-resists; photo polymer systems; photochromics; photo thermoplastics; and ferro-electric crystals. Recently, the use of storage and processing capabilities of computers, together with CCD cameras, has been used for recording holograms.

Silver halide photographic emulsions are a commonly used recording material for holography, mainly because of relatively high sensitivity and easy availability. Manufacture, use, and processing of these emulsions have been well standardized. The need for wet processing and drying may constitute a major drawback. Dichromated gelatin can be considered an almost ideal recording material for volume phase holograms, as it has a large refractive index modulation capability, high resolution, low absorption and scattering.

Because of these features, dichromated gelatin has been extensively investigated. The most significant disadvantage is its comparatively low sensitivity. Pennington *et al.* developed an alternative technique, which combines the advantage of silver halide emulsions (high sensitivity) with that of dichromated gelatin (low absorption scattering and high stability). It involves exposing silver halide photographic emulsion and then processing it, so as to obtain a volume phase hologram consisting solely of hardened gelatin.

Thin phase holograms can be recorded on photoresists, which are light-sensitive organic films yielding relief image after exposure and development. They offer advantages of easy replication using thermoplastics but are slow in response and undergo nonlinear effects at diffraction efficiency greater than ~ 0.05. Shipley AZ-1350 is a widely used photoresist, with maximum sensitivity in the ultraviolet, dropping rapidly for longer wavelengths towards the blue.

Photopolymers are being keenly investigated because they offer advantages such as ease of handling, low cost, and real-time recording for the application of holography and nonlinear optics. However, they have low sensitivity and short shelf life.

Thin phase surface relief holograms can be recorded in a thin layer of thermoplastics. These materials have a reasonably high sensitivity over the whole visible spectrum, fairly high diffraction efficiency, and do not require wet processing. Their application in holographic interferometry, optical information processing, and in making compact holographic devices has been widely reported.

Photorefractive crystals such as $Bi_{12}SiO_{20}$, $LiNbO_3$, $Bi_{12}GeO_{20}$, $BaTiO_3$, $LiTaO_3$, etc., as recording materials, offer high-angular sensitivity and provide capability to read and write volume holographic data in real time. Besides the materials discussed here, several other materials have been investigated for recording holograms; these include photocromics, elastomere devices, magneto-optic materials, etc.

Application of Holography

Holography can be constructed not only with the light waves of lasers, but also with sound waves, microwaves, and other waves in the electromagnetic spectrum of radiation. Holograms made with ultraviolet light or X-rays can record images of objects/particles smaller than the wavelength of visible light, e.g., atoms or molecules. Acoustical holography uses sound waves to see through solid objects. Holography has a vast scope of practical applications, which have been classified into two major categories:

1. applications requiring three-dimensional images for visual perception; and
2. applications in which holography is used as a measuring tool.

The unique ability of holography – to record and reconstruct both electromagnetic and sound waves – makes it a valuable tool for education, science, industry, and business. Below are some of the important applications:

1. Holographic interferometry (HI) is one of the most powerful measuring tools. In HI, two states of an object, i.e., initial and deformed state are recorded on the same photographic plate. After reconstruction of the light wave corresponding to two states of an object, interferences and deformations are displayed in terms of the interference pattern. The change of distance of one tenth of a micron, or lower, can be resolved. HI provides scientists/engineers with crucial data for design of critical machine

parts of power-generating equipment, in the aircraft industry, automobile industry, and nuclear installations (say, for example, in the design of containers used to transport nuclear materials, improve the design of aircraft wings and turbine blades, etc.). Presently, HI is being widely used in mechanical engineering, acoustics, and aerodynamics, for nondestructive testing, to investigate oscillation in diaphragms and flow around various objects, respectively.

2. Microwave holography can detect objects deep within spaces, by recording the radio waves they emit.
3. Another important application of holography is the design of optical elements, which possess special properties. A holographic recording of a concave mirror behaves in much the same way as the mirror itself, i.e., it can focus the light. In some cases chromatism can be introduced in the design of elements so that a location of the point, where the beams are focused, depends on the wavelength. This can be achieved by accurately choosing the recording arrangement of the focusing elements, these elements are, in fact, diffraction gratings. These can have low noise levels, freedom from astigmatism, and have other useful properties. Holographic optical elements have found applications in supermarket scanners to read barcodes, headup displays in fighter aircraft to observe critical cockpit instruments, etc.
4. A telephone credit card used in Europe and other developed countries has embossed surface holograms which carry a monetary value. When the card is inserted into the telephone, a card reader discerns the amount due and deducts the appropriate amount to cover the cost of the call.
5. Holography is having applications in analog and digital computers, offering remarkable opportunities to realize various logical operators, devices for identifying images based on matched filtering, and in computer memory units. The basic advantage of holographic memory is that a relatively large volume of information can be stored and that there are a limited number of ways to change the record. The arrival of the first prototype of optical computers, which use holograms as storage material for data, could have a dramatic impact on the holographic market. The yet-to-be-unveiled optical computers will be able to deliver trillions of bits of information faster than the current generation of computers.
6. Optical tweezers are going to become an important tool for the study of micro-organisms/bacteria, etc.
7. Holograms can be used to locate and retrieve information without knowing its location in the storage medium, only needing to know some of its content.
8. The links between computer science and holography are now well-established and are developing, with at least two aspects making computer-generated holograms extremely interesting. First, such holograms enable us to obtain visual 3-dimensional reconstructions of imagined objects. For example, one can reconstruct three dimensions of a model of an object still in the design stage. Second, computer-generated holograms can be used to reconstruct lightwaves with specified wave fronts. This means specially computed and manufactured holograms may function as optical elements that transform the incident light wave into desired wavefronts.
9. Another important applications of holography is its utilization to compensate for the distortion that occurs when viewing objects through optically heterogeneous mediums. It can be achieved based on the principle of beam reversal by the hologram.
10. Finally, let us consider the application of holography to art. The development of holography gives very effective ways of creating qualitative three-dimensional images. Thus, a new independent area of holographic creative work representational/artistic holography has appeared. The art of holographic depiction has developed along two major routes. The first is creation of the view hologram, used as holograms of natural objects that are to be displayed in exhibitions and museums; these are also known as artistic holograms. Portrait holography is also classified under this category. The progress in portrait holography is hampered partly because of imperfection of pulsed lasers and partly because of the deterioration of photographic material when exposed to pulsed electromagnetic radiation.

    The principle behind the creation of elusion by using composite holograms is also very convincing for the display of objects. To synthesize the composite holograms, the photographs of various aspects of a scene are printed onto the photographic plate. The synthesis techniques for preparing the composite hologram are very complicated, while the images created by holograms are still far from perfect. However, there is no doubt that composite holography opens holography up as an artistic technique. Recently rainbow holography has been very popular to display such objects.

See also: Fraunhofer Diffraction. Fresnel Diffraction

Further Reading

Benton, S.A., 1969. On a method for reducing the information content of holograms. Journal of the Optical Society of America 59, 1545.
Denisyuk, Y.N., 1962. Photographic reconstruction of the optical properties of an object in its own scattered radiation field. Sov. Phys.-Dokl. 7, 543.
Denisyuk, Y.N., 1963. On the reproduction of the optical properties of an object by the wave field of its scattered radiation, Pt. I. Optical Spectroscopy 15, 279.
Denisyuk, Y.N., 1965. On the reproduction of the optical properties of an object by the wave field of its scattered radiation, Pt II. Optical Spectroscopy 18, 152.
Gabor, D., 1949. Microscopy by reconstructed wavefronts. Proceedings of the Royal Society A197, 454.
Gabor, D., 1951. Microscopy by reconstructed wavefronts: II. Proceedings of the Physical Society B64, 449.
Gabor, S.D., 1948. A new microscopic principle. Nature 161, 777.
Hariharan, P., 1983. Optical Holography: Principle, Techniques and Applications. Cambridge, UK: Cambridge University Press.
Leith, E.N., Upatnieks, J., 1962. Reconstructed wavefronts and communication theory. Journal Optical of the Society of America 52, 1123.

Leith, E.N., Upatnieks, J., 1963. Wavefront reconstruction with continuous-tone objects. Journal of the Optical Society of America 53, 1377.
Leith, E.N., Upatnieks, J., 1964. Wavefront reconstruction with diffused illumination and three-dimensional objects. Journal of the Optical Society of America 54, 1295.
Pennington, K.S., Lin, L.H., 1965. Multicolor wavefront reconstruction. Applied Physics Letters 7, 56.
Stroke, G.W., Labeyrie, A.E., 1966. White-light reconstruction of holographic images using the Lippmann-Bragg diffraction effect. Physics Letters 20, 368.

The Fractional Order Fourier Transform and Fresnel Diffraction

Pierre Pellat-Finet, University of Southern Brittany, Lorient, France
Yezid Torres Moreno, Industrial University of Santander, Bucaramanga, Colombia

Introduction

The fractional order Fourier transform is an extension of the standard Fourier transform. Let α be a complex number and let f be a function of two real variables. The 2-dimensional fractional Fourier transform of order α of f is defined by

$$\mathcal{F}_\alpha[f]\left(\vec{\sigma}\right) = \frac{\mathrm{i}e^{-\mathrm{i}\alpha}}{\sin\alpha}\exp\left(-\mathrm{i}\pi\sigma^2\cot\alpha\right)\int_{\mathbb{R}^2}\exp\left(-\mathrm{i}\pi\rho^2\cot\alpha\right)\exp\left(\frac{2\mathrm{i}\pi\vec{\sigma}\cdot\vec{\rho}}{\sin\alpha}\right)f(\vec{\rho})\,d\vec{\rho} \tag{1}$$

where $\vec{\rho}$ and $\vec{\sigma}$ are 2-dimensional real vectors. The standard Fourier transform is $\mathcal{F}_{\pi/2}$, and $\mathcal{F}_0[f] = f$ for every function f. Fractional order Fourier transforms compose according to $\mathcal{F}_\beta \circ \mathcal{F}_\alpha = \mathcal{F}_{\beta+\alpha}$.

The integral expression of $\mathcal{F}_\alpha$ in Eq. (1) is similar to the integral expression of Fresnel diffraction, so that it is possible to express Fresnel diffraction phenomenae in the form of fractional order Fourier transforms and to then develop a method for analyzing and solving some problems of diffraction and light propagation.

Diffraction-Propagation From a Spherical Emitter to a Spherical Receiver

If $\mathcal{A}$ is a spherical emitter or receiver (a spherical segment) with vertex Ω_A and curvature center C_A, its radius of curvature is $R_A = \overline{\Omega_A C_A}$ (**Fig. 1**). We choose Cartesian coordinates x, y on $\mathcal{A}$ with origin at Ω_A, and use the spatial vector $\vec{r} = (x, y)$ to localize a point on $\mathcal{A}$. We denote $r = ||\vec{r}|| = (x^2 + y^2)^{1/2}$ and $\mathrm{d}\vec{r} = \mathrm{d}x\,\mathrm{d}y$. We consider monochromatic waves whose wavelength is λ in the propagation medium, which is assumed to be isotropic and homogeneous.

In the framework of a scalar theory of diffraction, the field transfer from a spherical emitter $\mathcal{A}$ (radius R_A, coordinates $\vec{r}$) to a spherical receiver $\mathcal{B}$ (radius R_B, coordinates $\vec{s}$) at a distance D (taken from vertex to vertex, **Fig. 1**) is expressed by

$$U_B(\vec{s}) = \frac{\mathrm{i}}{\lambda D}\exp\left[-\frac{\mathrm{i}\pi}{\lambda}\left(\frac{1}{R_B}+\frac{1}{D}\right)s^2\right]\int_{\mathbb{R}^2}\exp\left[-\frac{\mathrm{i}\pi}{\lambda}\left(\frac{1}{D}-\frac{1}{R_A}\right)r^2\right]\exp\left(\frac{2\mathrm{i}\pi}{\lambda D}\vec{r}\cdot\vec{s}\right)U_A(\vec{r})\mathrm{d}\vec{r} \tag{2}$$

where U_A is the complex (electric) field amplitude on $\mathcal{A}$, and U_B on $\mathcal{B}$. A constant multiplicative phase factor, of the form $\exp(-2\mathrm{i}\pi D/\lambda)$, has been omitted in Eq. (2).

Eq. (2) generally corresponds to a Fresnel diffraction phenomenon. If $D = R_A$, we obtain a Fraunhofer phenomenon; moreover if $R_B = -R_A$, the field amplitude U_B is the spatial standard Fourier transform of U_A and the sphere $\mathcal{B}$ is called the Fourier sphere of $\mathcal{A}$, denoted by $\mathcal{F}$ (see **Fig. 2**).

Mathematical Expression of Diffraction Through a Fractional Order Fourier Transform

The similarity between Eqs. (1) and (2) should be clear. To express Eq. (2) by using a fractional order Fourier transform we proceed as follows. Let K be such that

$$K = \left(1 - \frac{D}{R_A}\right)\left(1 + \frac{D}{R_B}\right) \tag{3}$$

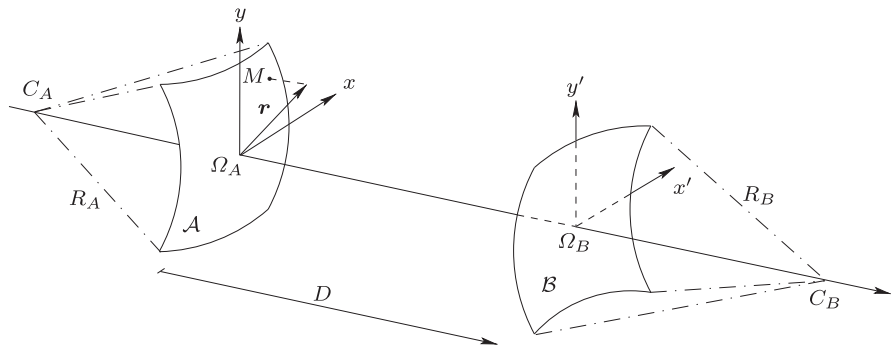


Fig. 1 Describing the field transfer by diffraction-propagation from the spherical emitter $\mathcal{A}$ to the spherical receiver $\mathcal{B}$ at a distance $D = \overline{\Omega_A\Omega_B}$. Generally we have a Fresnel diffraction phenomenon.

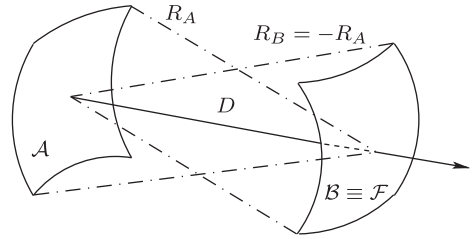


Fig. 2 If $D = R_A = -R_B$, light propagates from $\mathcal{A}$ to its Fourier sphere $\mathcal{F}$ according to a Fraunhofer diffraction phenomenon.

For the sake of simplicity we assume $0 \leq K \leq 1$, and we define α (a real number) by

$$\cos^2 \alpha = K, \quad -\pi \leq \alpha \leq \pi, \quad \alpha D \geq 0 \tag{4}$$

An equivalent definition, that is useful in some derivations, is

$$\cot^2 \alpha = \frac{(D + R_B)(R_A - D)}{D(D - R_A + R_B)}, \quad -\pi \leq \alpha \leq \pi, \quad \alpha D \geq 0 \tag{5}$$

We then define auxiliary parameters ε_A and ε_B by

$$\varepsilon_A = \frac{D}{R_A - D} \cot \alpha, \quad \varepsilon_A R_A > 0 \tag{6}$$

$$\varepsilon_B = \frac{D}{R_B + D} \cot \alpha \tag{7}$$

(Then, necessarily $\varepsilon_B R_B > 0$.)

Eventually we choose scaled spatial variables $\vec{\rho}$ on $\mathcal{A}$ and $\vec{\sigma}$ on $\mathcal{B}$ according to

$$\vec{\rho} = \frac{\vec{r}}{\sqrt{\lambda \varepsilon_A R_A}}, \quad \vec{\sigma} = \frac{\vec{s}}{\sqrt{\lambda \varepsilon_B R_B}} \tag{8}$$

We use scaled field amplitudes V_A on $\mathcal{A}$ and V_B on $\mathcal{B}$ such that

$$V_A\left(\vec{\rho}\right) = \sqrt{\frac{\varepsilon_A R_A}{\lambda}}\, U_A\left(\sqrt{\lambda \varepsilon_A R_A}\, \vec{\rho}\right), \qquad V_B\left(\vec{\sigma}\right) = \sqrt{\frac{\varepsilon_B R_B}{\lambda}}\, U_B\left(\sqrt{\lambda \varepsilon_B R_B}\, \vec{\sigma}\right) \tag{9}$$

Then, Eq. (2) becomes

$$V_B(\vec{\sigma}) = \mathrm{e}^{\mathrm{i}z} \mathcal{F}_\alpha[V_A](\vec{\sigma}) \tag{10}$$

Eq. (10) expresses a Fresnel diffraction phenomenon in the form of a fractional Fourier transform whose order depends on the propagation distance and the radii of the emitter and the receiver. Fraunhofer diffraction is obtained for $\alpha = \pi/2$.

If $K > 1$ or $K < 0$, the order α becomes a complex number and also corresponds to some diffraction phenomenon, depending on propagation distance and radii of curvature.

In order to not repeat definitions of parameters everytime, we will write

$$\left(R_A, R_B, D, \vec{r}, \vec{s}, U_A, U_B, \lambda\right) \leftrightsquigarrow \left(\alpha, \varepsilon_A, \varepsilon_B, \vec{\rho}, \vec{\sigma}, V_A, V_B\right) \tag{11}$$

to indicate the correspondence between physical parameters and variables (on the left side) and fractional parameters, scaled variables and scaled field amplitudes (on the right).

Representing diffraction through a fractional order Fourier transform is in accordance with two basic features of light propagation:

1. Continuity of light propagation. Fractional order Fourier transforms are such that $\mathcal{F}_\alpha[f] \to \mathcal{F}_\beta[f]$ if $\alpha \to \beta$, for every function f. Near the emitter, we have $\alpha = 0$ and $\mathcal{F}[V_A] = V_A$; there is no diffraction. Near the Fourier sphere ($D = R_A$) we have a Fourier transform $\mathcal{F}_{\pi/2}$ and a Fraunhofer diffraction. Physically there is a continuous succession of diffraction phenomenae, from no diffraction (near the emitter) to a Fraunhofer diffraction (near the Fourier sphere), passing by intermediate Fresnel type phenomenae (see **Fig. 3**).
2. The composition law of fractional order Fourier transforms is in accordance with the Huygens-Fresnel principle that establishes that diffraction from $\mathcal{A}_1$ to $\mathcal{A}_2$ (represented by $\mathcal{F}_\alpha$) can be seen as the succession of two diffraction phenomenae : from $\mathcal{A}_1$ to $\mathcal{A}_3$ ($\mathcal{F}_{\alpha_1}$) and from $\mathcal{A}_3$ to $\mathcal{A}_2$ ($\mathcal{F}_{\alpha_2}$), where $\mathcal{A}_3$ is an intermediate spherical segment ; then $\mathcal{F}_\alpha = \mathcal{F}_{\alpha_2} \circ \mathcal{F}_{\alpha_1}$ and $\alpha = \alpha_1 + \alpha_2$ (see **Fig. 4**).

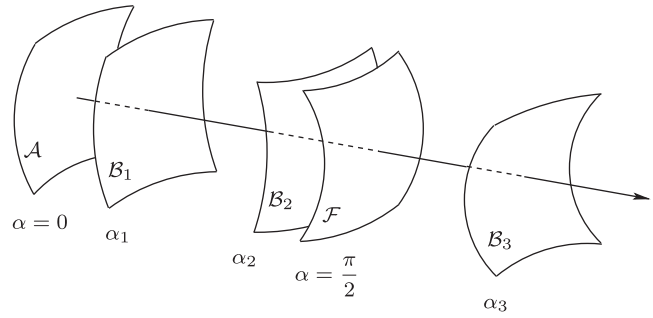


Fig. 3 Various values of α for various receivers: $0 < \alpha_1 < \alpha_2 < \pi/2 < \alpha_3$. The parameter α varies continuously when the receiver moves along the propagation axis.

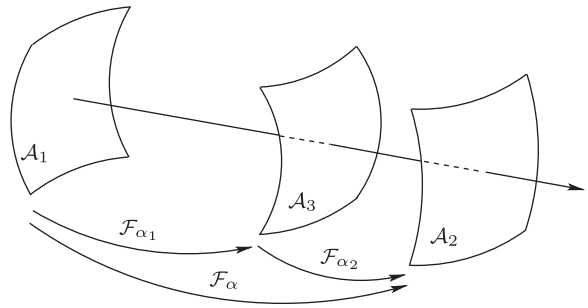


Fig. 4 Illustrating the accordance between the composition law of fractional order Fourier transforms and the Huygens-Fresnel principle: $\alpha = \alpha_1 + \alpha_2$, so that $\mathcal{F}_\alpha = \mathcal{F}_{\alpha_2} \circ \mathcal{F}_{\alpha_1}$.

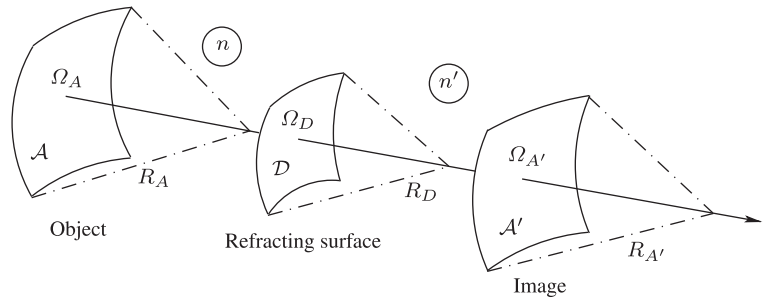


Fig. 5 The refracting surface, separating two propagation media with refractive indices n and n'.

Working With Fractional Orders

The basic idea is that problems in the scalar theory of diffraction can be solved by only manipulating fractional orders without explicitely writing corresponding integral expressions. We illustrate the method by considering imaging through a refracting spherical surface (vertex Ω_D, radius R_D, see **Fig. 5**). Let $\mathcal{A}$ be the object (spherical emitter), in the object space (with refractive index n), and let $\mathcal{A}'$ be a receiver in the image space (refractive index n'). We denote $d = \overline{\Omega_D \Omega_A}$ and $d' = \overline{\Omega_D \Omega_{A'}}$, as usual in geometrical optics (d is opposite to the distance of propagation from $\mathcal{A}$ to $\mathcal{D}$).

According to the Huygens-Fresnel principle, the field transfer from $\mathcal{A}$ to $\mathcal{A}'$ is the composition of two transfers: from $\mathcal{A}$ to $\mathcal{D}$ and from $\mathcal{D}$ to $\mathcal{A}'$. The corresponding mathematical expression could be obtained by the composition of two integrals that are like the one in Eq. (2). This is a cumbersome method that can be greatly simplified if we use the fractional Fourier transforms associated with ligth propagation.

The field transfer from $\mathcal{A}$ to $\mathcal{D}$ is described by

$$\left(R_A, R_D, -d, \vec{r}, \vec{s}, U_A, U_D, \lambda\right) \leftrightsquigarrow \left(\alpha, \varepsilon_A, \varepsilon_D, \vec{\rho}, \vec{\sigma}, V_A, V_D\right) \tag{12}$$

and leads to

$$V_D = \mathrm{e}^{\mathrm{i}\alpha} \mathcal{F}_\alpha[V_A] \tag{13}$$

The field transfer from $\mathcal{D}$ to $\mathcal{A}'$ corresponds to

$$\left(R_D, R_{A'}, d', \vec{s}, \vec{r'}, U_D, U_{A'}, \lambda'\right) \leftrightsquigarrow \left(\alpha', \varepsilon'_D, \varepsilon'_A, \vec{\sigma'}, \vec{\rho'}, V'_D, V'_{A'}\right) \tag{14}$$

so that

$$V'_{A'} = \mathrm{e}^{\mathrm{i}\alpha'} \mathcal{F}_{\alpha'}\left[V'_D\right] \tag{15}$$

The composition of fractional transforms in Eqs. (13) and (15) is possible only if scaled variables on $\mathcal{D}$, $\vec{\sigma}$ and $\vec{\sigma}'$, are identical for both transfers, in order to obtain $V'_D = V_D$. This happens if $\lambda\varepsilon_D = \lambda'\varepsilon'_D$, that is, if

$$\frac{-\lambda d}{R_D - d}\cot\alpha = \frac{\lambda' d'}{R_D - d'}\cot\alpha' \tag{16}$$

For every function f we have $\mathcal{F}_0[f] = f$ and $\mathcal{F}_\pi[f](\vec{\sigma}) = f(-\vec{\sigma})$, so that the field amplitude on $\mathcal{A}'$ is the image of the field amplitude on $\mathcal{A}$ if $\alpha + \alpha' = 0$ or if $\alpha' + \alpha = \pi$ (inverted image). Then $\cot\alpha = -\cot\alpha'$, and since $n\lambda = n'\lambda'$, we deduce from Eq. (16)

$$\frac{n'}{d'} = \frac{n}{d} + \frac{n' - n}{R_D} \tag{17}$$

Eq. (17) is the relation of conjugation of the refracting surface. It has been deduced from a scalar theory of diffraction without writing any integral, but adequately managing parameters of the involved fractional order Fourier transforms.

The imaging between $\mathcal{A}$ and $\mathcal{A}'$ is expressed by $V'_{A'}(\vec{\sigma}') = V_A(\vec{\sigma}')$, if $\alpha + \alpha' = 0$, and by $V'_{A'}(\vec{\sigma}') = V_A(-\vec{\sigma}')$, if $\alpha + \alpha' = \pi$.

We now examine the curvature of the receiver. We denote $q = \overline{\Omega_A C_A} = d + R_A$ and $q' = \overline{\Omega_{A'} C_{A'}} = d' + R_{A'}$. Since $\cot\alpha = -\cot\alpha'$, we use Eq. (5) and write

$$\frac{(R_D - d)(R_A + d)}{d(d + R_A - R_D)} = \frac{(d' + R_{A'})(R_D - d')}{d'(d' - R_D + RA')} \tag{18}$$

We use Eq. (16) with $\cot\alpha' = -\cot\alpha$, and deduce from Eq. (18) the relation

$$\frac{q}{n(q - R_D)} = \frac{q'}{n'(q' - R_D)} \tag{19}$$

that is

$$\frac{n'}{q'} = \frac{n}{q} + \frac{n' - n}{R_D} \tag{20}$$

Eq. (20) is the conjugaison relationship written for the curvature centers of $\mathcal{A}$ y $\mathcal{A}'$. We conclude that $\mathcal{A}'$ is the image of $\mathcal{A}$ if, and only if, the vertices of $\mathcal{A}$ and $\mathcal{A}'$ are conjugated and their centers are also conjugated.

Image Formation by a Lens

The use of fractional Fourier transforms in diffraction theory leads us to generalize one of the most important results in modern optics: the field amplitude of an image formed by a lens is the convolution of the object field amplitude with the Fourier transform of the pupil function of the lens. More precisely, we refer to **Fig. 6** where a lens images the spherical emitter $\mathcal{A}$ (vertex Ω_A) on the spherical receiver $\mathcal{A}'$ (vertex $\Omega_{A'}$). If g_ν denotes the transversal magnification at points Ω_A and $\Omega_{A'}$ we consider first the geometrical image field amplitude U_{GA} on A', defined by

$$U_{GA}\left(\vec{r}'\right) = \frac{1}{g_\nu} U_A\left(\frac{\vec{r}'}{g_\nu}\right) \tag{21}$$

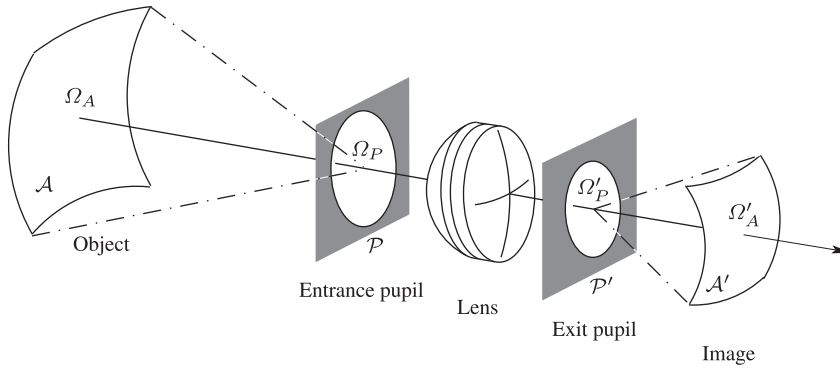


Fig. 6 Image formation by a lens, corresponding to Eq. (23).

We introduce the entrance pupil of the lens, say $\mathcal{P}$, located at Ω_p, and the exit pupil $\mathcal{P}'$ at $\Omega_{p'}$, which is the image of the entrance pupil through the lens. The pupil function of the lens is p such that $p(\vec{s}) = 1$, if the point $\vec{s}$ lays inside the pupil, and $p(\vec{s}) = 0$, if not. Let $d = \overline{\Omega_P \Omega_A}$ and let h be the function such that

$$h\left(\vec{r}\right) = \frac{1}{\lambda g_v{}^2 d^2} \mathcal{F}_{\pi/2}[p]\left(\frac{\vec{r}}{\lambda g_v d}\right) \tag{22}$$

Then the field amplitude on $\mathcal{A}'$ is given by

$$U_{A'} = h * U_{GA} \tag{23}$$

Although very important, Eq. (23) rigorously holds, indeed, only if $\mathcal{A}$ is centered on the entrance pupil $\mathcal{P}$, as shown in **Fig. 6** (It can be shown that $\mathcal{A}'$ is centered on the exit pupil.). It does not hold for an arbitrary emitter, not centered on the pupil; nor if the objet is the entrance pupil of the lens (whose image is the exit pupil).

By using the fractional Fourier transform representation of diffraction, the previous result can be extended to emitters that are not centered on the entrance pupil. For that purpose we first define the fractional convolution product of functions f and g by

$$f *^{\alpha} g = \mathcal{F}_{-\alpha}[\mathcal{F}_{\alpha}[f]\mathcal{F}_{\alpha}[g]] \tag{24}$$

We have $f *^{0} g = fg$ and $f *^{\pi/2} g = f * g$, where * denotes the standard convolution product.

Let α be the order of the fractional Fourier transform associated with the field transfer from $\mathcal{A}$ to $\mathcal{P}$. We introduce scaled variables and amplitudes V_A on $\mathcal{A}$ and $V_{A'}$ on $\mathcal{A}'$. If p_s is the scaled pupil function and $h_\alpha = \mathcal{F}_\alpha[p_s]$ then

$$V_{A'} = h_\alpha *^{\alpha} V_{GA} \tag{25}$$

Eq. (25) is a general expression for image formation by a lens, taking into account the effect of the lens aperture. It holds whatever the position of the object and of its curvature center. If $\mathcal{P}$ is located on the Fourier sphere of $\mathcal{A}$, since $\alpha = \pi/2$, we obtain Eq. (23) once more. If $\mathcal{A}$ is located on the pupil, we obtain $\alpha = 0$ and the result is compatible with the fact that the field on the exit pupil is identical to the field on the entrance pupil (since pupils are conjugated).

Optical Resonator Stability

An optical resonator is made up of two spherical mirrors, say $\mathcal{M}_1$ (radius R_1) and $\mathcal{M}_2$ (radius R_2). Expressing propagation from a mirror to the other by using a fractional order Fourier transform leads us to a complete theory of optical resonators, based on a scalar diffraction theory. Once more, the main interest of the method is deduce results without explicitly writing integrals, but suitably manipulating fractional orders of involved fractional Fourier transforms.

The field transfer from a mirror to the other can be expressed by a fractional Fourier transform whose order α depends on the distance between the mirrors and their radii. This holds true for the reverse transfer, so that a round trip is expressed by the composition of two fractional order Fourier transforms. It can be shown that the orders are the same for both transfers, so that a round trip from $\mathcal{M}_1$ to $\mathcal{M}_1$ is expressed in the form

$$V_1 = \mathrm{e}^{-4\mathrm{i}\pi D/\lambda} \mathrm{e}^{2\mathrm{i}\alpha} \mathscr{F}_{2\alpha}[V_1] \tag{26}$$

where V_1 is the scaled field amplitude on $\mathcal{M}_1$ and where the phase factor that was omitted in Eq. (2) has been reintroduced.

A distinction should be made according to whether α is a real or a complex number. If α is real, there is a phase shift in the round trip, and the resonator is said to be stable. If α is a complex number, there is an attenuation of the field amplitude, which is due to diffraction losses; the resonator is said to be unstable.

We state: an optical resonator is stable if, and only if, the field transfer from a mirror to the other is expressed by a fractional Fourier transform whose order is a real number. According to Eqs. (3) and (4) we conclude that an optical resonator is stable if, and only if,

$$0 \le \left(1 - \frac{D}{R_1}\right)\left(1 - \frac{D}{R_2}\right) \le 1 \tag{27}$$

(The sign of R_2 in Eq. (27) is opposite to the sign of R_B in Eq. (3), because light propagation direction is changed after reflexion on the corresponding mirror).

Eq. (27) is a very well known condition for a resonator to be stable.

Another consequence of Eq. (26) is that field amplitudes of waves that propagate inside the resonator are eigenfunctions of fractional Fourier transforms (whatever the order). These eigenfunctions are Hermite-Gauss functions, of the form

$$\varphi_{m,n}(\xi, \eta) = H_m\left(\sqrt{2\pi}\xi\right) H_n\left(\sqrt{2\pi}\eta\right) \exp\left[-\pi\left(\xi^2 + \eta^2\right)\right] \tag{28}$$

where H_m is the Hermite polynomial of integer order m. Hence the notion of Hermite-Gauss modes of an optical resonator.

Eq. (26) also leads to find longitudinal modes of a resonator. For every integer m, we have

$$\mathscr{F}_{\alpha}[H_m] = e^{2\mathrm{i}m\alpha} H_m \tag{29}$$

so that for $V_1 = \varphi_{m,n}$ we deduce from Eq. (26)

$$\varphi_{m,n} = e^{-4i\pi D/\lambda} e^{2i\alpha} e^{2i(m+n)\alpha}\, \varphi_{m,n} \tag{30}$$

and necessarily

$$\frac{2\pi D}{\lambda} - (1 + m + n)\alpha = k\pi \tag{31}$$

where k is an integer. Only waves whose wavelengths satisfy Eq. (31) can propagate in the resonator. They are the longitudinal modes of the resonator.

Gaussian Beams and Gouy Phase

A laser with a stable optical cavity (or resonator) generates Gaussian beams, which are modes of the laser resonator and are described by Hermite-Gauss functions. It can be shown that a Gaussian beam admits a minimal area, called the beam waist and located on a plane (while other wave surfaces are spherical segments). At a distance d from the waist, the Gaussian beam wave surface is a sphere whose curvature radius is R and the field propagation from the beam waist to the previously described surface is expressed by a fractional Fourier transform whose order is α_d with

$$\tan \alpha_d = \frac{\lambda d}{\pi w_0^2} \tag{32}$$

If the phase origine is taken on the waist plane, the field amplitude at a distance d is

$$U_{m,n}(x, y, d) = U_0 \frac{w_0}{w_d} H_m\left(\frac{\sqrt{2}}{w_d} x\right) H_n\left(\frac{\sqrt{2}}{w_d} y\right) \exp\left[-\frac{x^2 + y^2}{w_d^2} - \frac{2i\pi d}{\lambda} + i(m + n + 1)\alpha_d\right] \tag{33}$$

and $(m + n + 1)\alpha_d$ is no more than the Gouy phase.

Uncoupling Variables

A 2-dimensional fractional Fourier transform can be seen like the composition of two 1-dimensional transforms, a fact that can be adequately applied in optics.

We denote by $\mathcal{F}_\alpha^{[1]}$ the 1-dimensional Fourier transform of order α (a real number), defined by

$$\mathcal{F}_\alpha^{[1]}[f](\eta) = \frac{\exp\left[i\left(\frac{\pi}{4} s(\alpha) - \frac{\alpha}{2}\right)\right]}{\sqrt{|\sin\alpha|}} \exp\left(-i\pi\eta^2 \cot\alpha\right) \int_{\mathbb{R}} \exp\left(-i\pi\xi^2 \cot\alpha\right) \exp\left(\frac{2i\pi\xi\eta}{\sin\alpha}\right) f(\xi) d\xi \tag{34}$$

where $s(\alpha)$ denotes the sign of α.

The Fubini theorem leads us to compose two 1-dimensional fractional order Fourier transforms according to the following diagram, where $\mathcal{F}_{\alpha_1}^{[1]}$ operates on the variable ξ_1, and $\mathcal{F}_{\alpha_2}^{[1]}$ on ξ_2,

$$\begin{array}{ccc} f(\xi_1, \xi_2) & \xrightarrow{\mathcal{F}_{\alpha_1}^{[1]}} & \mathcal{F}_{\alpha_1}^{[1]}[f](\eta_1, \xi_2) \\ \downarrow \mathcal{F}_{\alpha_2}^{[1]} & & \downarrow \mathcal{F}_{\alpha_2}^{[1]} \\ \mathcal{F}_{\alpha_2}^{[1]}[f](\xi_1, \eta_2) & \xrightarrow{\mathcal{F}_{\alpha_1}^{[1]}} & \mathcal{F}_{\alpha_1,\alpha_2}[f](\eta_1, \eta_2) \end{array} \tag{35}$$

where $\mathcal{F}_{\alpha_1,\alpha_2}$ is a 2-dimensional integral. We remark that $\mathcal{F}_{\alpha,\alpha}$ is no more than the 2-dimensional fractional order Fourier transform defined in Eq. (1).

The previous decomposition is useful when dealing with optical emitters and receivers that are segments of tori, that is, have two radii of curvature (see **Fig. 7**). We consider a toric emitter $\mathcal{A}$ (coordinates x_1 and x_2, taken along principal sections) and a toric receiver $\mathcal{B}$ (coordinates x_1' and x_2') respectively at a distance D. We assume that principal sections of $\mathcal{A}$ and $\mathcal{B}$ are along parallel axes: $x_1' \| x_1$ and $x_2' \| x_2$ (see **Fig. 7**). Let R_1 and R_2 be the principal radii of $\mathcal{A}$ and R_1' and R_2' those of $\mathcal{B}$.

We consider a first 1-dimensional fractional Fourier transform for the transfer from $\mathcal{A}$ to $\mathcal{B}$ but only for the x_1 and x_1' variables. We define α_1 by

$$\cos^2\alpha_1 = \left(1 - \frac{D}{R_1}\right)\left(1 + \frac{D}{R_1'}\right) \tag{36}$$

and then

$$\varepsilon_1 = \frac{D}{R_1 - D} \cot\alpha_1, \quad \varepsilon_1' = \frac{D}{R_1' + D} \cot\alpha_1 \tag{37}$$

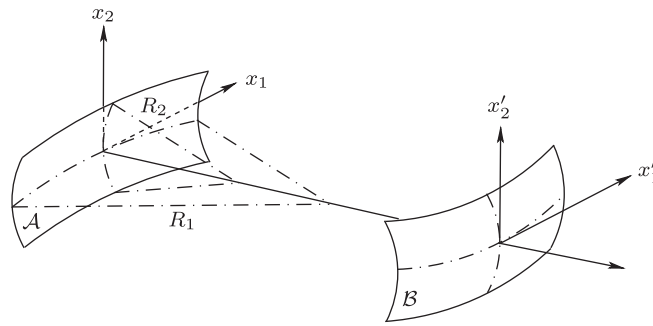


Fig. 7 Describing the field transfer by diffraction-propagation from a toric emitter *A* to a toric receiver *B*.

Scaled variables are

$$\xi_1 = \frac{x_1}{\sqrt{\lambda\varepsilon_1 R_1}}, \quad \xi'_1 = \frac{x'_1}{\sqrt{\lambda\varepsilon'_1 R'_1}} \tag{38}$$

and scaled functions

$$V_1(\xi_1, x_2) = \sqrt[4]{\frac{\varepsilon_1 R_1}{\lambda}} U_A\left(\sqrt{\lambda\varepsilon_1 R_1}\,\xi_1, x_2\right), \quad V'_1(\xi'_1, x_2) = \sqrt[4]{\frac{\varepsilon'_1 R'_1}{\lambda}} U_B\left(\sqrt{\lambda\varepsilon_1 R_1}\,\xi'_1, x'_2\right) \tag{39}$$

We then obtain

$$V'_1(\xi'_1, x_2) = \mathrm{e}^{\mathrm{i}\alpha_1/2}\mathcal{F}^{[1]}_{\alpha_1}[V_1](\xi'_1, x_2) \tag{40}$$

Then, we consider a second 1-dimensional Fourier transform for the transfer from $\mathcal{A}$ to $\mathcal{B}$ but for the x_2 and x'_2 variables. We define α_2, ε_2, ε'_2, ξ_2 and ξ'_2 as α_1, ε_1, ε'_1, ξ_1 and ξ'_1, after changing R_1 into R_2 and R'_1 into R'_2 in Eqs. (36)–(38). Scaled field amplitudes on $\mathcal{A}$ and $\mathcal{B}$ are

$$V_A(\xi_1, \xi_2) = \sqrt[4]{\frac{\varepsilon_2 R_2}{\lambda}} V_1\left(\xi_1, \sqrt{\lambda\varepsilon_2 R_2}\,\xi_2\right) \tag{41}$$

$$V_B(\xi'_1, \xi'_2) = \sqrt[4]{\frac{\varepsilon'_2 R'_2}{\lambda}} V'_1\left(\xi'_1, \sqrt{\lambda\varepsilon_2 R_2}\,\xi'_2\right) \tag{42}$$

so that

$$V_B(\xi'_1, \xi'_2) = \mathrm{e}^{\mathrm{i}\alpha_2/2}\mathcal{F}^{[1]}_{\alpha_2}[V'_1](\xi'_1, \xi'_2) = \mathrm{e}^{\mathrm{i}(\alpha_1+\alpha_2)/2}\mathcal{F}^{[1]}_{\alpha_2}\left[\mathcal{F}^{[1]}_{\alpha_1}[V_1]\right](\xi'_1, \xi'_2) = \mathrm{e}^{\mathrm{i}(\alpha_1+\alpha_2)/2}\mathcal{F}_{\alpha_1,\alpha_2}[V_A](\xi'_1, \xi'_2) \tag{43}$$

Eq. (43) is a generalization of Eq. (10).

The method can be applied to deal with toric (or astigmatic) refractive surfaces and astigmatic lenses. It also provides a method for analyzing propagation and imaging of Gaussian beams with elliptical waists, like those produced by some laser diodes.

See also: Fraunhofer Diffraction. Fresnel Diffraction

Further Reading

Goodman, J.W., 2005. Introduction to Fourier Optics. Englewood: Robert & Company.
Namias, V., 1980. The fractional order Fourier transform and its application to quantum mechanics. Journal of the Institute of Mathematics and its Applications 25, 241–265.
Ozaktas, H.M., Zalevsky, Z., Kutay, M.A., 2001. The Fractional Fourier Transform, with Applications in Optics and Signal Processing. Chichester: John Wiley & Sons.
Pellat-Finet, P., 2009. Optique de Fourier, Théorie métaxiale et fractionnaire. Paris: Springer.
Pellat-Finet, P., Fogret, É., 2011. A fractional Fourier transform theory of optical resonators. In: Emersone, P.S. (Ed.), Progress in Optical Fibers. New York: Nova Science Publisher, pp. 299–351.
Pellat-Finet, P., Torres, Y., 1997. Image formation with coherent light: The fractional fourier transform approach. Journal of Modern Optics 44, 1581–1594.

Ambiguity Function in Optics

JP Guigay, European Synchrotron Radiation Facility (ESRF), Grenoble, France

Introduction

The ambiguity function (AF) associated to a time signal $s(t)$ is defined as the following two-dimensional time-frequency function

$$A(\tau,f)=\int dt\, s^*(t)s(t-\tau)\exp(-i2\pi ft) \tag{1}$$

It has been introduced by Woodward (1953) in the theory of signal processing of radar or sonar measurements, τ being the delay time related to the distance of the target and f the frequency shift due to its velocity. It bears in its name the idea that it is impossible to perform arbitrarily accurate measurements of both the distance and the velocity of the target. The well-known reason is that the width of a signal in the time domain is inversely proportional to its width in the frequency domain (a narrower signal has a wider spectrum and inversely). From the mathematical point of view, this is just a property of the Fourier transformation; from the physical point of view, this is a common property for waves of any kind and is the basis of the well-known uncertainty relations in quantum physics.

It was shown by Papoulis (1974) that the AF concept can be used in Fourier optics, considering the propagation of a spatially coherent quasi-monochromatic optical field $\psi_z(x)$ transmitted through slit apertures and thin lenses, in the conditions of paraxial propagation in the z-direction; the considered AF is then defined as the following Fourier transform with respect to x:

$$A_z(a,f)=\int dx\, \psi_z^*(x-a/2)\psi_z(x+a/2)\exp(-i2\pi fx) \tag{2}$$

In the present paper, it will be shown that the concept of AF in optics can be considered as an extension of the spectral analysis of images; the images are two-dimensional intensity patterns $I(x,y)$ which can often be analyzed in a convenient way by considering their Fourier transform (intensity spectrum):

$$\tilde{I}(u,v)=\iint dxdyI(x,y)\exp[-i2\pi(ux+vy)]\ \text{ or }\ \tilde{I}\left(\vec{f}\right)=\int d\vec{x}I(\vec{x})\exp\left[-i2\pi\vec{x}.\vec{f}\right] \tag{3}$$

where two-dimensional vectors are used; the variables (u, v) are the spatial frequencies conjugate to the coordinates (x, y).

Definition in Terms of the Mutual Intensity

Besides the intensity distribution, the phase correlation between an arbitrary pair of points must be included for a complete description of the optical field. For this purpose, it seems natural to generalize $I(\vec{x})$ by the mutual intensity $\rho(\vec{x},\vec{x'})=\sqrt{I(\vec{x})I(\vec{x'})}\gamma(\vec{x},\vec{x'})$, where $\gamma(\vec{x},\vec{x'})$ is the complex degree of coherence associated to the pair of points $(\vec{x},\vec{x'})$ (see Born and Wolf (1999)). The mutual intensity is reduced to the usual intensity $I(\vec{x})$ when these two points coincide. The ambiguity is defined (Guigay, 1978), in the frame of Phase-Space Optics (PSO), as the Fourier transform with respect to $\vec{x}$ of the mutual-intensity written as $\rho(\vec{x}+\vec{a}/2,\vec{x}-\vec{a}/2)$:

$$A\left(\vec{f},\vec{a}\right)=\int d\vec{x}\,\exp\left(-i2\pi\vec{x}.\vec{f}\right)\rho(\vec{x}+\vec{a}/2,\vec{x}-\vec{a}/2) \tag{4}$$

$A\left(\vec{f},\vec{a}\right)$ is defined in the phase-space which is a four-dimensional space, since $\vec{f}$ and $\vec{a}$ are two-dimensional vectors. It represents a complete description of the partially coherent field optical field. The equivalent representation

$$A\left(\vec{f},\vec{a}\right)=\int d\vec{m}\,\exp(i2\pi\vec{m}.\vec{a})\tilde{\rho}\left(\vec{m}+\vec{f}/2,\vec{m}-\vec{f}/2\right) \tag{5}$$

is obtained by replacing the mutual-intensity by its Fourier expansion with respect to $\vec{x}$

$$\rho(\vec{x}+\vec{a}/2,\vec{x}-\vec{a}/2)=\iint d\vec{g}d\vec{h}\,\exp\left\{i2\pi\left[\vec{g}.(\vec{x}+\vec{a}/2)-\vec{h}.(\vec{x}-\vec{a}/2)\right]\right\}\tilde{\rho}\left(\vec{g},\vec{h}\right) \tag{6}$$

Formula (4) shows that $A\left(\vec{f},0\right)$ represents the intensity spectrum defined in (3), which, as well as $I(\vec{x})$, represents the experimental data recorded by a digital detector. Formula (5) shows that $A(0,\vec{a})$ is the inverse Fourier transform of the intensity distribution $\tilde{\rho}(\vec{m},\vec{m})$in Fourier space.

A particularly important phase-space function is the Wigner Distribution Function (WDF) which was introduced in optics by Bastiaans (1978) (see Testorf *et al.* (2010) for a precise account of its properties). The WDF is defined as the Fourier transform of $\rho(\vec{x}+\vec{a}/2,\vec{x}-\vec{a}/2)$ with respect to $\vec{a}$, instead of $\vec{x}$:

$$W(\vec{x},\vec{g})=\int d\vec{a}\,\exp(-i2\pi\vec{a}.\vec{g})\rho(\vec{x}+\vec{a}/2,\vec{x}-\vec{a}/2) \tag{7}$$

 doi:10.1016/B978-0-12-803581-8.09375-9

It can also be expressed as

$$W(\vec{x},\vec{g}) = \int d\vec{m}\, \exp(i2\pi\vec{m}.\vec{x})\tilde{\rho}(\vec{g}+\vec{m}/2,\vec{g}-\vec{m}/2) \tag{8}$$

The mutual intensity can be obtained from the AF or from the WDF as

$$\rho\left(\vec{x}+\frac{\vec{a}}{2},\vec{x}-\frac{\vec{a}}{2}\right) = \int d\vec{f}\, \exp\left(i2\pi\vec{f}.\vec{x}\right)A\left(\vec{f},\vec{a}\right) = \int d\vec{g}\, \exp(i2\pi\vec{g}.\vec{a})W(\vec{x},\vec{g}) \tag{9}$$

The AF and the WDF are related to each-other by double Fourier transformation over the position and frequency variables:

$$A\left(\vec{f},\vec{a}\right) = \int d\vec{x} \int d\vec{g}\, \exp\left[i2\pi\left(\vec{a}.\vec{g}-\vec{f}.\vec{x}\right)\right]W(\vec{x},\vec{g}) \tag{10}$$

$$W(\vec{x},\vec{g}) = \int d\vec{a} \int d\vec{f}\, \exp\left[i2\pi\left(\vec{f}.\vec{x}-\vec{a}.\vec{g}\right)\right]A\left(\vec{f},\vec{a}\right) \tag{11}$$

The property $\rho\left(\vec{x}',\vec{x}\right) = \left(\rho(\vec{x},\vec{x}')\right)^*$ of the mutual intensity shows that the AF, in general complex, satisfies the relation $A\left(-\vec{f},-\vec{a}\right) = \left[A\left(\vec{f},\vec{a}\right)\right]^*$. It also shows that the WDF is real (but can be negative as well as positive).

In the exit plane of an object of transmittance $T(\vec{x})$, under uniform coherent illumination, the mutual intensity is equal to the product $T(\vec{x}+\vec{a}/2)T^*(\vec{x}-\vec{a}/2)$ which is often referred to as the "product-space-representation" of the signal $T(\vec{x})$; The corresponding AF and WDF, which are referred to as the AF and the WDF associated to $T(\vec{x})$, are redundant representations of $T(\vec{x})$.

PSO representations are useful tools to characterize the performances of optical systems. They provide elegant approaches to the description and processing of optical signals or images. It has been shown by Nugent (2007) that the concept of AF can be used to unify the various noninterferometric approaches to phase retrieval.

In order to simplify the formulation of the following sections, we shall often consider one-dimensional fields, in which case the two-dimensional vectors are replaced by scalars, without a real lost of generality, because the extension to the general case is usually straightforward.

Intensity Spectrum of a Fresnel Diffraction Pattern Under Coherent Illumination

General Formulation

For simplicity, let us consider a plane wave, of wavelength λ, incident along the z-direction on a thin object of transmittance $T(\vec{x})$ in the plane $z=0$. In the conditions of Fresnel diffraction, the wave-function in a plane $z=D$ is

$$\psi_D(\vec{x}) = |\lambda D|^{-1/2} \exp\left(-i\frac{\pi}{4}\right) \int d\vec{\eta}\, \exp\left[i\pi\frac{(\vec{\eta}-\vec{x})^2}{\lambda D}\right]T(\vec{\eta}) \tag{12}$$

The corresponding intensity spectrum can be expressed by the multiple integral

$$\tilde{I}_D\left(\vec{f}\right) = \int d\vec{x}\, \exp\left(-i2\pi\vec{x}.\vec{f}\right) \iint \frac{d\vec{\eta}d\vec{\eta}'}{\lambda D} \exp\left[i\pi\frac{(\vec{\eta}-\vec{x})^2-(\vec{\eta}'-\vec{x})^2}{\lambda D}\right]T(\vec{\eta})T^*(\vec{\eta}')$$

The integration over $\vec{x}$ results in the appearance of a two-dimensional delta-function

$$\int d\vec{x}\, \exp\left[-i2\pi\vec{x}.\left(\vec{f}+\frac{\vec{\eta}-\vec{\eta}'}{\lambda D}\right)\right] = \lambda D\delta\left(\lambda D\vec{f}+\vec{\eta}-\vec{\eta}'\right) \tag{13}$$

and the intensity spectrum can therefore be reduced to a single integration (Guigay, 1977a,b):

$$\tilde{I}_D\left(\vec{f}\right) = \exp\left(-i\pi\lambda D\vec{f}^2\right) \int d\vec{\eta}\, \exp\left(-i2\pi\vec{f}.\vec{\eta}\right)T(\vec{\eta})T^*\left(\vec{\eta}+\lambda D\vec{f}\right) \tag{14}$$

This last formula can be expressed in symmetrical form:

$$\tilde{I}_D\left(\vec{f}\right) = \int d\vec{x}\, \exp\left(-i2\pi\vec{f}.\vec{x}\right)T\left(\vec{x}-\frac{\lambda D\vec{f}}{2}\right)T^*\left(\vec{x}+\frac{\lambda D\vec{f}}{2}\right) \tag{15}$$

Similar expressions also exist in terms of $\tilde{T}(\vec{f})$:

$$\begin{aligned}\tilde{I}_D\left(\vec{f}\right) &= \exp\left(-i\pi\lambda D\vec{f}^2\right) \int d\vec{h} \exp\left(-i2\pi\lambda D\vec{h}.\vec{f}\right)\tilde{T}\left(\vec{h}+\vec{f}\right)\tilde{T}^*\left(\vec{h}\right) \\ &= \int d\vec{h} \exp\left(-i2\pi\lambda D\vec{h}.\vec{f}\right)\tilde{T}\left(\vec{h}+\vec{f}/2\right)\tilde{T}^*\left(\vec{h}-\vec{f}/2\right)\end{aligned} \tag{16}$$

It is interesting to note that the AF associated with $T(x)$ is apparent in this formulation if the intensity spectrum is formally considered as a function of f and $a=\lambda Df$.

Derivation of the Transport of Intensity Equation

By linearization of $T\left(\vec{x}-\frac{\lambda D\vec{f}}{2}\right)T^*\left(\vec{x}+\frac{\lambda D\vec{f}}{2}\right)$ as $T(\vec{x})T^*(\vec{x})+\lambda D\vec{f}.\left[T\vec{\partial}T^*-T^*\vec{\partial}T\right]$, where $\vec{\partial}T$ denotes the two-dimensional gradient of the function $T(\vec{x})=\sqrt{I_0(\vec{x})}\exp[i(\vec{x})]$ written in terms of its modulus and its phase, we obtain the following approximation of formula (15) in the case of small values of $|\lambda D\vec{f}|$:

$$\tilde{I}_D\left(\vec{f}\right)=\tilde{I}_0\left(\vec{f}\right)-i\lambda D\vec{f}.\int d\vec{x}\,\exp\left(-i2\pi\vec{f}.\vec{x}\right)I_0\left(\vec{x}\right)\vec{\partial}\left(\vec{x}\right)$$

By inverse Fourier transformation of this last formula, we obtain the well-known transport of intensity equation (Teague, 1983):

$$I_D\left(\vec{x}\right)=I_0\left(\vec{x}\right)-\frac{\lambda D}{2\pi}\vec{\partial}.\left[I_0(\vec{x})\vec{\partial}(\vec{x})\right] \tag{17}$$

which is widely used for phase retrieval.

Application to Simple Objects

This formulation can provide interesting results for some typical Fresnel diffraction patterns; for instance, in the one-dimensional case of a slit of full-width w, we obtain (Guigay, 1977b):

$$\tilde{I}_D(f)=\int_{-(w-|\lambda Df|)/2}^{(w-|\lambda Df|)/2}dx e^{-i2\pi fx}=\frac{\sin[\pi f(w-|\lambda Df|)]}{\pi f}\ \text{for}\ |f|\le\left|\frac{w}{\lambda D}\right|,\ \ 0\ \text{otherwise} \tag{18}$$

Note that this expression is analytically much simpler than the intensity distribution $I(x)$ expressed in terms of Fresnel integrals represented geometrically by the Cornu spiral (see Born and Wolf (1999)).

Contrast Transfer Functions

Considering $T(\vec{x})=\exp[-B(\vec{x})+i(\vec{x})]$, where $\exp[-B(\vec{x})]$ and $(\vec{x})$ are the absorption and the phase modulations of the object, we can introduce in formula (11) the approximation

$$T^*\left(\vec{x}+\frac{\lambda D\vec{f}}{2}\right)T\left(\vec{x}-\frac{\lambda D\vec{f}}{2}\right)1-B\left(\vec{x}+\frac{\lambda D\vec{f}}{2}\right)-B\left(\vec{x}-\frac{\lambda D\vec{f}}{2}\right)+i\left[\left(\vec{x}+\frac{\lambda D\vec{f}}{2}\right)-\left(\vec{x}-\frac{\lambda D\vec{f}}{2}\right)\right] \tag{19}$$

which should be valid under the conditions $0B(\vec{x})1$ (weak absorption) and $|(\vec{x})-(\vec{x}-\lambda D\vec{f})|1$. This last condition is the slowly-varying phase condition (Testorf *et al.*, 2010) which is less restrictive than the weak-phase condition $|(\vec{x})|1$. Under such conditions, the intensity spectrum takes a simple linear form (Guigay, 1977a)

$$\tilde{I}_D\left(\vec{f}\right)=\delta\left(\vec{f}\right)-2\cos\left(\pi\lambda D\vec{f}^{\,2}\right)\tilde{B}\left(\vec{f}\right)+2\sin\left(\pi\lambda D\vec{f}^{\,2}\right)\sim\left(\vec{f}\right) \tag{20}$$

The factors of $\tilde{B}(\vec{f})$ and $\sim(\vec{f})$ are named as the absorption-transfer function (ATF) and the phase-transfer function (PTF) respectively.

Formula (20) can be generalized to the case of an imaging system with aberrations others than defocusing. In electron microscopy, for which primary spherical aberration characterized by the coefficient C_s is unavoidable, the following formula is to be used (Hanszen, 1972; Wade, 1974):

$$\tilde{I}_D(\vec{f})=\delta(\vec{f})-2\cos\left[\omega(\vec{f})\right]\tilde{B}(\vec{f})+2\sin\left[\omega(\vec{f})\right]\sim(\vec{f}) \tag{21}$$

$$\text{with}\quad \omega(\vec{f})=\pi\left(\lambda D\vec{f}^{\,2}+C_S\lambda^3\vec{f}^{\,4}/2\right) \tag{22}$$

Partial Talbot Effect

According to the Talbot effect, in the case of a periodic object of period a, a Fresnel diffraction amplitude identical to the object amplitude $T(x)$ is obtained at the so-called Talbot distance $D_T=a^2/\lambda$. It was shown in Guigay (1971) and Arrizón and Ojeda-Castañeda (1992) that the amplitude distribution at any intermediate distance equal to a rational fraction $(n'/n)D_T$ of the Talbot distance is a linear combination of n laterally shifted copies of $T(x)$; this is called the Fractional Talbot effect.

The intensity of any Fresnel diffraction pattern is itself a periodic function of $\vec{x}$ (the letter), also with periodicity a; according to (15), its Fourier coefficient of order m (m integer) is

$$\tilde{I}_D\left(\frac{m}{a}\right)=\frac{1}{a}\int_{-a/2}^{a/2}dx\ \exp\left(-i2\pi x\frac{m}{a}\right)T^*\left(x+\frac{\lambda Dm}{2a}\right)T\left(x-\frac{\lambda Dm}{2a}\right) \tag{23}$$

This last formula shows that $\tilde{I}_D(pm/a) = \tilde{I}_0(m/a)$, with p being an integer, if the distance D is equal to D_T/m. This result may be named as the "the partial Talbot effect," because it represents the conservation of a part of the intensity spectrum of the periodic object.

Propagation Through a Paraxial Optical System in Terms of AF

Propagation in Free Space

Let us consider the propagation in free space, with mean direction along the z-axis, of a partially coherent beam. The mutual intensity in the $z=D$ plane is given in terms of the mutual intensity in the $z=0$ plane as:

$$\rho_D\left(\vec{x}+\frac{\vec{a}}{2},\vec{x}-\frac{\vec{a}}{2}\right) = \frac{1}{\lambda D}\int d\vec{\eta}\,\exp\left[i\pi\frac{(\vec{\eta}-\vec{x})^2}{\lambda D}\right]\int d\vec{\xi}\,\exp\left[-i\pi\frac{(\vec{\xi}-\vec{x})^2}{\lambda D}\right]\rho_0\left(\vec{\eta}+\frac{\vec{a}}{2},\vec{\xi}-\frac{\vec{a}}{2}\right)$$
$$= \frac{1}{\lambda D}\int d\vec{\eta}\int d\vec{\xi}\,\exp\left[i\pi\frac{\eta^2-\xi^2}{\lambda D}-2i\pi\vec{x}.\frac{\vec{\eta}-\vec{\xi}}{\lambda D}\right]\rho_0\left(\vec{\eta}+\frac{\vec{a}}{2},\vec{\xi}-\frac{\vec{a}}{2}\right)$$

After insertion of this expression in (4), it can be seen that the integration over x results in the delta-function $\lambda D\delta\left(\vec{\eta}-\vec{\xi}+\lambda D\vec{f}\right)$. The multiple integral is thus reduced to a single integral.

$$A_D\left(\vec{f},\vec{a}\right) = \int d\vec{\eta}\,\exp\left(-i2\pi\vec{f}.\vec{\eta}\right)\rho_0\left(\vec{\eta}+\frac{\vec{a}-\lambda D\vec{f}}{2},\vec{\eta}-\frac{\vec{a}-\lambda D\vec{f}}{2}\right)$$
$$\text{Therefore}\quad A_D\left(\vec{f},\vec{a}\right) = A_0\left(\vec{f},\vec{a}-\lambda D\vec{f}\right) \tag{24}$$

This represents a translation of amplitude $\lambda D\vec{f}$ of $A_0\left(\vec{f},\vec{a}\right)$ on the second variable.

This result can be more readily obtained by recalling that the Fresnel diffraction integral is the convolution of the input function $T(\vec{x})$ by the propagator $G(\vec{x}) = \exp\left(i\pi\vec{x}^2/\lambda D - i\pi/4\right)/\sqrt{\lambda D}$; the output spectrum is therefore the product of the input spectrum $\tilde{T}(f)$ by the spectrum of $G(x)$ which is $\tilde{G}(f) = \exp(-i\pi\lambda D f^2)$; this is translated in terms of mutual intensity as

$$\tilde{\rho}_D\left(\vec{m}+\vec{f}/2,\vec{m}-\vec{f}/2\right) = \exp\left\{-i\pi\lambda D\left[\left(\vec{m}+\vec{f}/2\right)^2-\left(\vec{m}-\vec{f}/2\right)^2\right]\right\}\tilde{\rho}_0\left(\vec{m}+\vec{f}/2,\vec{m}-\vec{f}/2\right)$$
$$= \exp\left(-i2\pi\lambda D\vec{m}.\vec{f}\right)\tilde{\rho}_0\left(\vec{m}+\vec{f}/2,\vec{m}-\vec{f}/2\right) \tag{25}$$

Inserting this expression in (5), we indeed obtain directly:

$$A_D\left(\vec{f},\vec{a}\right) = \int d\vec{m}\exp\left[i2\pi\vec{m}.\left(\vec{a}-\lambda D\vec{f}\right)\right]\tilde{\rho}_0\left(\vec{m}+\frac{\vec{f}}{2},\vec{m}-\frac{\vec{f}}{2}\right) = A_0\left(\vec{f},\vec{a}-\lambda D\vec{f}\right)$$

As shown in Bastiaans (1978), a similar formula also exists for the WDF:

$$W_D(\vec{x},\vec{g}) = W_0(\vec{x}-\lambda D\vec{g},\vec{g}) \tag{26}$$

According to formulas (24) and (26), the AF and the WDF propagate in a uniform medium without a change of their functional forms; only the variables are linearly transformed. This is an elegant representation of Fresnel diffraction phenomena.

Transmission Through a Thin Object

In this case, the incident mutual-intensity $\rho_{inc}(\vec{x},\vec{x'})$ is multiplied by $T(\vec{x})T^*(\vec{x'})$, where $T(\vec{x})$ is the object transmittance; The AF of the incident beam is then to be convoluted with the object-AF $A_T(\vec{f},\vec{a})$, as follows

$$A\left(\vec{f},\vec{a}\right) = \int d\vec{h}A_{inc}\left(\vec{h},\vec{a}\right)A_T\left(\vec{f}-\vec{h},\vec{a}\right) \tag{27}$$

where $A_T\left(\vec{f},\vec{a}\right) = \int d\vec{x}\,\exp\left(-i2\pi\vec{x}.\vec{f}\right)T(\vec{x}+\vec{a}/2)T^*(\vec{x}-\vec{a}/2)$. The transmission by a thin lens of focal length F is of special interest; this lens behaves as an object of transmittance $\exp(-i\pi x^2/\lambda F)$. With $\rho_{inc}(\vec{x}+\vec{a}/2,\vec{x}-\vec{a}/2)$ being the mutual intensity in the lens entrance surface, the AF in the lens exit surface is calculated as

$$A_{exit}\left(\vec{f},\vec{a}\right) = \int d\vec{x}\,\exp\left[-i2\pi\vec{x}.\vec{f}-i\pi\frac{(\vec{x}+\vec{a}/2)^2-(\vec{x}-\vec{a}/2)^2}{\lambda F}\right]\rho_{inc}(\vec{x}+\vec{a}/2,\vec{x}-\vec{a}/2)$$
$$= \int d\vec{x}\,\exp\left[-i2\pi\vec{x}.\left(\vec{f}+\frac{\vec{a}}{\lambda F}\right)\right]\rho_{inc}(\vec{x}+\vec{a}/2,\vec{x}-\vec{a}/2)$$

$$\text{Therefore}\quad A_{exit}\left(\vec{f},\vec{a}\right) = A_{inc}\left(\vec{f}+\vec{a}/\lambda F,\vec{a}\right) \tag{28}$$

This represents a translation of amplitude $(-\vec{a}/\lambda F)$ of $A_0\left(\vec{f},\vec{a}\right)$ on the first variable.

The corresponding formula for the WDF is easily found as:

$$W(\vec{x},\vec{g}) = W_{inc}(\vec{x},\vec{g}+\vec{x}/\lambda F) \tag{29}$$

Propagation in a Paraxial Optical System

Consider the following process: propagation from an input plane to a thin lens over a distance D_1, then transmission by this lens of focal length F and finally propagation to the output plane over a distance D_2. Performing the corresponding transformations successively, according to (24, 28), it is easy to obtain the output AF in terms of the input AF as:

$$A_{out}\left(\vec{f},\vec{a}\right) = A_{in}\left(\vec{f}-\vec{f}D_2/F+\vec{a}/\lambda F,\ \vec{a}-\vec{a}D_1/F-\vec{f}\lambda(D_1+D_2-D_1D_2/F)\right) \tag{30}$$

This represents a translation of amplitude $\vec{f}D_2/F-\vec{a}/\lambda F$ on the first variable and a translation of amplitude $\vec{a}D_1/F+\vec{f}\lambda(D_1+D_2-D_1D_2/F)$ on the second variable.

This phase-space method is thus a convenient and elegant tool to describe the propagation of a coherent or partially coherent beam in any system comprising coaxial lenses and it is possible to use the WDF instead of the AF. The phase-space method is much simpler than the method based on the propagation of mutual-intensity which would involve convolutions integrals or Fourier transformations.

The AF in Space-Invariant (Isoplanatic) Imaging

The mutual intensity $\rho_{im}(\vec{x},\vec{x'})$ in the image plane is given in terms of the mutual intensity $\rho_{ob}(\vec{x},\vec{x'})$ in the object plane (for convenience, the magnification is set equal to 1) as:

$$\rho_{im}\left(\vec{x},\vec{x'}\right) = \iint d\vec{\eta}d\vec{\eta'}\,G(\vec{x}-\vec{\eta})G^*\left(\vec{x'}-\vec{\eta'}\right)\rho_{ob}\left(\vec{\eta},\vec{\eta'}\right) \tag{31}$$

where $G(\vec{x})$ is the coherent point-spread-function (PSF) of the imaging system. The image-AF is therefore

$$A_{im}\left(\vec{f},\vec{a}\right) = \int d\vec{x}\,\exp\left(-i2\pi\vec{x}.\vec{f}\right)\int d\vec{\eta}G(\vec{x}+\vec{a}/2-\vec{\eta})\int d\vec{\eta'}\,G^*\left(\vec{x}-\vec{a}/2-\vec{\eta'}\right)\rho_{ob}\left(\vec{\eta},\vec{\eta'}\right) \tag{32}$$

By introducing new variables $\vec{\sigma}=(\vec{\eta}+\vec{\eta'})/2$, $\vec{\tau}=\vec{\eta}-\vec{\eta'}$, $\vec{t}=\vec{x}-\vec{\sigma}$ in this integral expression, we obtain directly (Guigay, 1978; Dutta and Goodman, 1977; Ojeda-Castañeda and Sicre, 1984)

$$\begin{aligned} A_{im}\left(\vec{f},\vec{a}\right) &= \iint d\vec{t}d\vec{\sigma}\,\exp\left[-i2\pi\vec{f}.(\vec{t}+\vec{\sigma})\right]\int d\vec{\tau}\,\mathrm{G}\left(\vec{t}+\frac{\vec{a}-\vec{\tau}}{2}\right)\mathrm{G}^*\left(t-\frac{\vec{a}-\vec{\tau}}{2}\right)\rho_{ob}\left(\vec{\sigma}+\frac{\vec{\tau}}{2},\vec{\sigma}-\frac{\vec{\tau}}{2}\right) \\ &= \int d\vec{\tau}A_G\left(\vec{f},\vec{a}-\vec{\tau}\right)A_{ob}\left(\vec{f},\vec{\tau}\right) \end{aligned} \tag{33}$$

This is the convolution integral, with respect to the variable $\vec{a}$, of the AF $A_{ob}(\vec{f},\vec{\tau})$ in the object plane with the function

$$A_G\left(\vec{f},\vec{a}\right) = \int d\vec{x}\,\exp\left(-i2\pi\vec{x}.\vec{f}\right)\mathrm{G}\left(\vec{x}+\frac{\vec{a}}{2}\right)\mathrm{G}^*\left(\vec{x}-\frac{\vec{a}}{2}\right) = \int d\vec{g}\,\exp(i2\pi\vec{g}.\vec{a})\tilde{G}\left(\vec{g}+\frac{f}{2}\right)\tilde{G}^*\left(\vec{g}-\frac{\vec{f}}{2}\right) \tag{34}$$

$\tilde{G}(\vec{g})$ is known as the pupil-function of the imaging system. $A_G(\vec{f},\vec{a})$ is to be considered equivalently as the AF of the coherent PSF or as the pupil-AF.

The image intensity spectrum, which is a quantity of special interest, is obtained by setting a equal to 0 in formula (33):

$$\tilde{I}_{im}(\vec{f}) = A_{im}\left(\vec{f},\vec{0}\right) = \int d\vec{\tau}A_G\left(\vec{f},-\vec{\tau}\right)A_{ob}\left(\vec{f},\vec{\tau}\right) \tag{35}$$

Formula (34) shows that, in the particular case of free-space propagation for which $\tilde{G}(\vec{f}) = \exp(-i\pi\lambda Df^2)$, the pupil-AF is a delta-function $\delta\left(\vec{a}-\lambda D\vec{f}\right)$. The pupil-AF of an ideal (stigmatic) system is equal to $\delta(\vec{a})$.

The AF of the Image of an Incoherent Source

The AF and the WDF of a point-source represented by the Dirac-function $\delta(\vec{x}-\vec{x}_0)$ are respectively

$$A\left(\vec{f},\vec{a}\right) = \int d\vec{x}\exp\left(-i2\pi\vec{x}.\vec{f}\right)\delta(\vec{x}-\vec{x}_0+\vec{a}/2)\delta(\vec{x}-\vec{x}_0-\vec{a}/2) = \delta(\vec{a})\exp\left(-i2\pi\vec{x}_0.\vec{f}\right) \tag{36}$$

$$W(\vec{x},\vec{g}) = \int d\vec{a}\,\exp(-i2\pi\vec{a}.\vec{g})\delta(\vec{x}-\vec{x}_0+\vec{a}/2)\delta(\vec{x}-\vec{x}_0-\vec{a}/2) = \delta(\vec{x}-\vec{x}_0) \tag{37}$$

The AF and the WDF of an extended incoherent source, considered as a continuous distribution $S(\vec{x}_0)$ of mutually incoherent point-sources, are obtained by summing the AF and the WDF of these mutually incoherent point-sources. They are respectively:

$$A\left(\vec{f},\vec{a}\right) = \int d\vec{x}_0 S(\vec{x}_0)\delta(\vec{a})\exp\left(-i2\pi\vec{x}_0.\vec{f}\right) = \tilde{S}\left(\vec{f}\right)\delta(\vec{a}) \tag{38}$$

$$W(\vec{x},\vec{g}) = \int d\vec{x}_0 S(\vec{x}_0)\delta(\vec{x}-\vec{x}_0) = S(\vec{x}) \tag{39}$$

Derivation of the Zernike-Van Cittert Theorem from the Propagation of the AF

According to formula (24) and (38), at the distance L from the incoherent source, the AF is therefore $\tilde{S}\left(\vec{f}\right)\delta\left(\vec{a}-\lambda L\vec{f}\right)$ and the mutual-intensity can be obtained from formula (9) as

$$\rho(\vec{x}+\vec{a}/2,\vec{x}-\vec{a}/2) = \int d\vec{f}\,\exp\left(i2\pi\vec{x}.\vec{f}\right)\tilde{S}\left(\vec{f}\right)\delta\left(\vec{a}-\lambda L\vec{f}\right) = \tilde{S}\left(\frac{\vec{a}}{\lambda L}\right)\exp\left(\frac{i2\pi\vec{x}.\vec{a}}{\lambda L}\right)(\lambda L)^{-1} \tag{40}$$

This result, which is the expression of the Van Cittert-Zernike theorem (Born and Wolf, 1999), can also be obtained from the WDF which is equal to $S\left(\vec{x}-\lambda L\vec{f}\right)$, since the WDF is easily found to be equal to $S(\vec{x})$ in the plane of the incoherent source.

Partial Coherence Properties in the Image of an Incoherent Source (Guigay, 1978)

The image of an incoherent source (or equivalently in the image of an object under incoherent illumination) by a non-ideal optical system shows some degree of coherence because the light from each point of the primary source is spread over a finite area in the image; it is convenient to introduce the image-AF, which characterizes this image completely, including its coherence properties, and has a simple expression:

$$A_{im}\left(\vec{f},\vec{a}\right) = \int d\vec{\tau} A_G\left(\vec{f},\vec{a}-\vec{\tau}\right)\tilde{S}\left(\vec{f}\right)\delta(\vec{\tau}) = A_G\left(\vec{f},\vec{a}\right)\tilde{S}\left(\vec{f}\right) \tag{41}$$

The Pupil-AF as a Generalization of the OTF

In the case $\vec{a}=0$, formula (41) gives the image intensity spectrum as

$$\tilde{I}_{im}\left(\vec{f}\right) = A_{im}\left(\vec{f},\vec{0}\right) = A_G\left(\vec{f},\vec{0}\right)\tilde{S}\left(\vec{f}\right) \tag{42}$$

This formula shows that $A_G\left(\vec{f},\vec{0}\right)$ is identical to the well-known Optical Transfer Function (OTF); a detailed account of the OTF can be found in Born and Wolf (1999). Since formula (41) is obviously a generalization of formula (42), the pupil-AF is to be considered as a generalization of the OTF.

The Pupil-AF and the OTF with Defocusing

According to (24), if the image is recorded at a distance D from its normal position (defocusing), $\vec{a}$ is to be replaced by $\vec{a}-\lambda D\vec{f}$ in (41). This means that the defocused pupil-AF is $A_G\left(\vec{f},\vec{a}-\lambda D\vec{f}\right)$. The defocused OTF is consequently $A_G\left(\vec{f},-\lambda D\vec{f}\right)$. Therefore, the pupil-AF contains the values of the OTF for any value of the defocusing distance. More precisely, as first pointed out in Brenner *et al.* (1983), the pupil-AF can be seen as a polar display of the OTF for variable defocusing distance: the OTF is displayed, as represented schematically in **Fig. 1**, along lines going through the origin of coordinates in the $\left(\vec{f},\vec{a}\right)$ representation (see chapter 5 of Testorf *et al.* (2010) for details).

This connection between the OTF and the pupil-AF has been used (Ojeda-Castañeda and Berriel-Valdos, 1988; Dowski and Cathey, 1995; Fitzgerrell *et al.*, 1997; Castro and Ojeda-Castañeda, 2004; Ojeda-Castañeda *et al.*, 2005; Castro *et al.*, 2006; Ojeda-Castañeda and Gómez-Sarabia, 2015) for designing pupil phase-masks (phase apodizers) which increase the depth of focus without losing lateral resolution and light-gathering power. Furthermore, various effects such as the behaviour of the Strehl ratio and the sensitivity to spherical aberration (Ojeda-Castañeda *et al.*, 1987), or the focal shift (Sheppard and Larkin, 2000), have been studied by considerations based on the pupil-AF. These applications are detailed in chapter 10 of Testorf *et al.* (2010).

Phase-Space Tomography

The idea of phase-space tomography (Raymer *et al.*, 1994; Tu and Tamura, 1998; Dragoman *et al.*, 2002; Tran *et al.*, 2005; Liu and Brenner, 2003; Liu *et al.*, 2008) is to reconstruct the AF (or the WDF) in the plane $z=0$ from a set of intensity measurements $I_D(\vec{x})$ in planes at different distances $z=D_n$. The mutual intensity can then be derived from the reconstructed AF (or WDF) by using formula (9). This is of interest for the characterization of the optical field in the plane $z=0$. If an object of transmittance $T(\vec{x})$ is placed in this plane, we obtain

$$\rho_{inc}\left(\vec{x}+\frac{\vec{a}}{2},x-\frac{\vec{a}}{2}\right)T\left(x+\frac{\vec{a}}{2}\right)T^*\left(\vec{x}-\frac{\vec{a}}{2}\right) = \int d\vec{f}\,\exp\left(i2\pi\vec{f}.\vec{x}\right)A\left(\vec{f},\vec{a}\right) \tag{43}$$

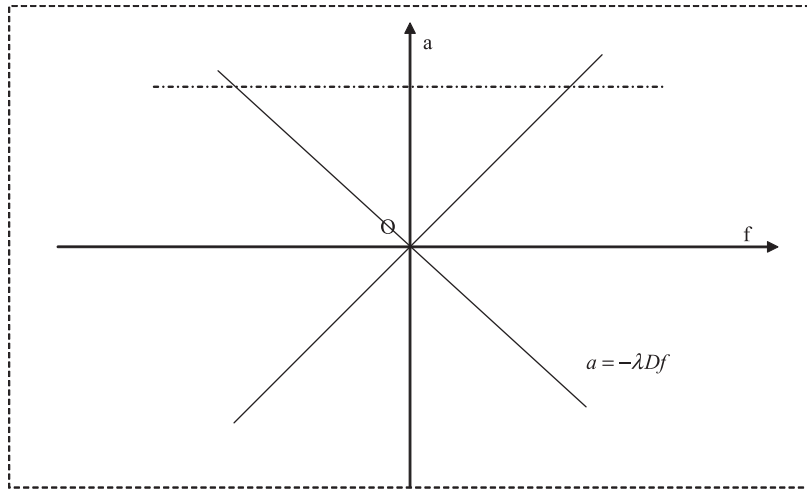


Fig. 1 Schematic PSO representation. The AF along the line $a = -\lambda Df$ correspond to the intensity spectra $\tilde{I}_D(f)$ at distance D from the reference plane. The integration of formula (44) is to be performed along lines parallel to the f-axis.

where ρ_{inc} is the mutual intensity of the incident beam; with $x=a/2$, this is reduced to

$$\rho_{inc}\left(\vec{a},\vec{0}\right)T(\vec{a})T^*\left(\vec{0}\right) = \int d\vec{f}\,\exp\left(i\pi\vec{f}.\vec{a}\right)A\left(\vec{f},\vec{a}\right) \tag{44}$$

As $\rho_{inc}\left(\vec{a},\vec{0}\right)$ and the modulus of $T(\vec{0})$ can be measured independently, this last formula allows the determination (up to a constant phase factor) of the complex function $T(\vec{a})$ from the AF.

The WDF tomographic reconstruction is based on the formula

$$I_D(\vec{x}) = \int d\vec{f}\,W\left(\vec{x}-\lambda D\vec{f},\vec{f}\right) \tag{45}$$

which shows that $I_D(\vec{x})$ is the projection of the WDF in the $(\vec{x},\vec{f})$ space along a direction which can be varied by the position $z=D$ of the recording plane. The operation which allows the WDF reconstruction is an inverse Radon transform (Testorf *et al.*, 2010; Tu and Tamura, 1998). The feasibility of the tomographic WDF reconstruction has been discussed in Raymer *et al.* (1994), Tu and Tamura (1998), Dragoman *et al.* (2002), and Tran *et al.* (2005).

The AF reconstruction is considered (Tu and Tamura, 1998; Dragoman *et al.*, 2002; Tran *et al.*, 2005; Liu and Brenner, 2003; Liu *et al.*, 2008) to be simpler, because there is no need of inverse Radon transform (the term phase-space tomography may be considered as inappropriate in this case); we only need to perform the Fourier transformation of the measured $I_D(\vec{x})$; the relation $\tilde{I}_D(\vec{f}) = A\left(\vec{f},-\lambda D\vec{f}\right)$ shows that the intensity spectra represent the variations of the AF along the radial lines $\vec{a} = -\lambda D\vec{f}$ in the $\left(\vec{f},\vec{a}\right)$ space, as depicted schematically in **Fig. 1**. In order to sample the AF over the complete $\left(\vec{f},\vec{a}\right)$ space, it is necessary to use negative, as well as positive, values of the distance D; this was not possible previously in X-ray optics because appropriate lenses were not available, but this situation may be changing now, because of the development of the so-called compound refractive X-ray lenses (see Simons *et al.* (2017), Terentyev *et al.* (2017) and references therein).

The process of AF reconstruction was first considered in the case of one-dimensional (1-dim) structures (Tu and Tamura, 1998). A 2-dim structure can be considered as an ensemble of 1-dim y-structures $T(x_0,y)$; the corresponding AF are $A(x_o;f_y,a_y)$, from which $T(x_0,y)$ can be derived as a function of y according to (44). For this purpose, a convenient set-up, actually a 1-Dim propagator system, has been proposed by Liu and Brenner (2003) (see also Liu *et al.* (2008)): a cylindrical lens of focal length F produces an exact image (corresponding to no effective propagation) in the x-direction, while there is propagation over the object-image distance in the y-direction; this distance can be varied by using several cylindrical lenses.

A Possible Approach to AF Reconstruction

Let us consider here the case of a one-dimensional object. The AF in the exit plane of an object illuminated by a tilted plane wave of tilting angle α is

$$\begin{aligned}\int dx\,\exp(-i2\pi f.x)T\left(x+\frac{a}{2}\right)\exp\left[i2\pi\frac{\alpha}{\lambda}\left(x+\frac{a}{2}\right)\right]T^*\left(x-\frac{a}{2}\right)\exp\left[-i2\pi\frac{\alpha}{\lambda}\left(x-\frac{a}{2}\right)\right]\\ = A_{ob}(f,a)\exp\left(i2\pi\frac{\alpha a}{\lambda}\right)\end{aligned} \tag{46}$$

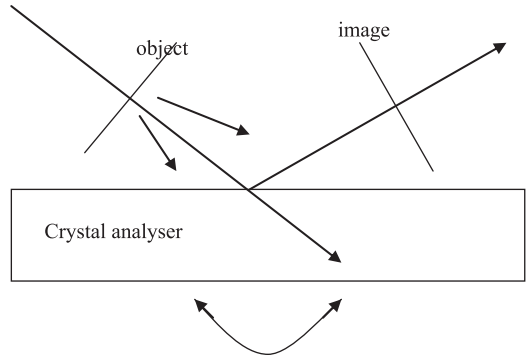


Fig. 2 Principle of the X-ray analyzer-based imaging system. A quasi-parallel and quasi-monochromatic beam is diffracted, after transmission through the object, by a plate of perfect silicon crystal. The arrangement is such that Bragg diffraction occurs, corresponding to reflecting planes parallel to the crystal surface. The Bragg-diffraction process is highly sensitive to the direction of the rays. Images are recorded across the diffracted beam for different angular settings of the crystal.

where $A_{ob}(f,a)$ is the AF associated to the transmittance $T(x)$. Setting $\omega=\alpha/\lambda$, the intensity spectrum of the image delivered by an imaging system with pupil-AF $A_G(f,\tau)$ is

$$\tilde{I}_{im}(f,\omega)=\int d\tau A_G(f,-\tau)A_{ob}(f,\tau)\exp(i2\pi\omega\tau) \tag{47}$$

which is a Fourier transform. Consequently:

$$A_{ob}(f,\tau)A_G(f,-\tau)=\int d\omega\exp(-i2\pi\omega\tau)\tilde{I}_{im}(f,\omega) \tag{48}$$

Supposing $A_G(f,-\tau)$ to be known and $\tilde{I}_{im}(f,\omega)$ to be measured as a function of ω, this last formula provides the possibility to obtain the object-AF.

This approach has been suggested (Guigay *et al.*, 2007b) in the context of X-ray analyzer-based imaging (ABI) which is a Schlieren-type technique (see **Fig. 2**) based on X-ray Bragg reflection by a perfect crystal plate (analyzer) placed downstream of the object and acting as an angle selector, therefore as a filter in Fourier space; the Fourier transform of the image amplitude is $\tilde{T}(f)\tilde{G}(f)$, where $\tilde{T}(f)$ is the Fourier transform of the object transmittance and $\tilde{G}(f)$ is the complex reflectivity of the crystal for an incident plane wave having angular shift $\delta\theta=\theta-\theta_B=\lambda f$ with respect to the exact Bragg angle θ_B. $\tilde{G}(f)$ is the pupil-function of the imaging system.

The angular range of Bragg reflection is typically a few microradians. For practical reasons, the beam is fixed and the crystal is rotated. When the crystal is rotated by an angle $\delta\theta$ from its peak position in the incident beam, this pupil-function is changed into $\tilde{G}(f-\delta\theta/\lambda)$. It is easily shown from formula (34) that the pupil-AF $A_G(f,a)$ is then changed into $A_G(f,a)\exp(i2\pi a\delta\theta/\lambda)$; we then obtain for the image intensity spectrum the same formula as (47), with $\omega=\delta\theta/\lambda$.

This technique is sensitive to the object structure in one dimension (sensitive to angular deviations in the Bragg diffraction plane). In order to overcome this limitation, it should be just necessary, for each angular position of the crystal-analyzer, to perform a 90-rotation of the object in its plane, in order to obtain finally two-dim information.

Propagation-Based Holographic Phase Retrieval From Several Images

For the sake of simplicity, let us consider the case of a one-dimensional object in the present section.

Fresnel Diffraction Images as In-Line Holograms

The holographic features of Fresnel diffraction images are clearly shown by considering an object transmittance in the form $T(\eta)=1+\psi(\eta)$, which is $\tilde{T}(f)=\delta(f)+\tilde{\psi}(f)$ in Fourier space. Formula (16) can be written as:

$$\begin{aligned}\exp\left(i\pi\lambda Df^2\right)\tilde{I}_D(f)&=\int dh\exp(-i2\pi\lambda Dhf)\left[\delta(h+f)+\tilde{\psi}(h+f)\right]\left[\delta(h)+\tilde{\psi}^*(h)\right]\\ \exp\left(i\pi\lambda Df^2\right)\tilde{I}_D(f)&=\delta(f)+\tilde{\psi}(f)+\exp\left(2i\pi\lambda Df^2\right)\tilde{\psi}^*(-f)+\int dh\exp(-i2\pi\lambda Dhf)\tilde{\psi}(h+f)\tilde{\psi}^*(h)\end{aligned} \tag{49}$$

This describes the process of holographic reconstruction. The two first terms corresponds to the reconstructed object; the next term corresponds to the out-of-focus image (at distance $2D$) of the conjugate object and the integral term is negligible if $|\psi(x)|1$ (weak object). The importance of these perturbation terms can be strongly reduced by performing the following summation based on N images recorded at different distances D_n:

$$\frac{1}{N}\sum_{n=1}^{N}\tilde{I}_{D_n}(f)\exp\left(i\pi\lambda D_nf^2\right)=\delta(f)+\tilde{\psi}(f)+\frac{\tilde{\psi}^*(-f)}{N}\sum_{n=1}^{N}\exp\left(2i\pi\lambda D_nf^2\right)+N^{-1}\int dh\tilde{\psi}(h+f)\tilde{\psi}^*(h)\sum_{n=1}^{N}\exp(-i2\pi\lambda D_nhf) \tag{50}$$

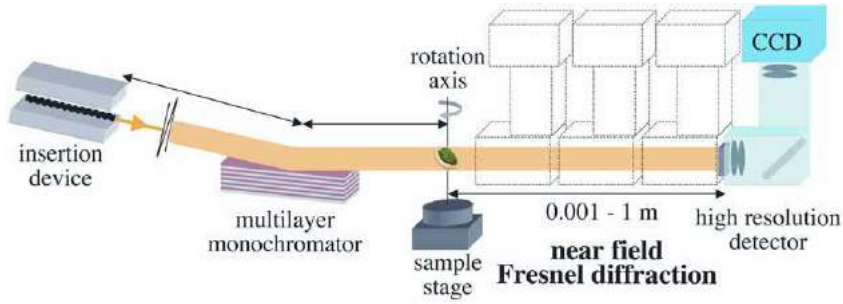


Fig. 3 Scheme of the set-up for X-ray holotomography operated on the ID19 beamline of the ESRF. The Synchrotron X-ray beam is monochromatised by a crystal monochromator or a multilayer, the energy used being typically around 15 keV (wavelength around 0.08 nm). The sample is mounted on a rotating table. The detector ensemble is a scintillator screen coupled by light optics to a CCD camera. This detector can be moved in order to record images close to the object or at different distances from it.

The quantity on the left-hand side may be calculated from digitally recorded images; this allows a good reconstruction of the object if the perturbation terms are nearly cancelled by this summation procedure.

Application to Phase Retrieval and X-ray Holotomography

In the case of a phase object, such that $T(\eta) = \exp[i(\eta)]$, we obtain

$$\tilde{I}_{D_n}(f) = \delta(f) + 2\sin(\pi\lambda D_n f^2)\sim(f) + (NLD)_n \tag{51}$$

Neglecting the nonlinear term (the integral term) denoted as $(NLD)_n$ in the left-hand side of this equation, the following estimation of the phase spectrum (Cloetens *et al.*, 1999) can be obtained, from the experimentally known $\tilde{I}_{D_n}(f)$ by a least-squares fitting as

$$\sim(f) = \frac{\sum_n \sin(\pi\lambda D_n f^2)\tilde{I}_{D_n}(f)}{2\sum_n \sin^2(\pi\lambda D_n f^2)} \tag{52}$$

From this result, it is possible to calculate the nonlinear terms in order to check whether they could indeed be neglected. If necessary, it is possible to take them into account recursively: the calculated $(NLD)_n$ can be subtracted from the experimentally known $\tilde{I}_{D_n}(f)$ to obtain a new estimate of the phase spectrum

$$\sim(f) = \frac{\sum_n \sin(\pi\lambda D_n f^2)\left[\tilde{I}_{D_n}(f) - NLT_{D_n}\right]}{2\sum_n \sin^2(\pi\lambda D_n f^2)} \tag{53}$$

and this process can be continued recursively.

If a single image ($N=1$) was used in formula (43), the phase spectrum could not be obtained for the spatial frequencies f such that transfer function $\sin(\pi\lambda D f^2)$ is close to 0. Using several images (typically 4 or 5 images) allows to eliminate this defect and to reduce the influence of the nonlinear terms.

This phase retrieval approach , which has some similarity with the focus variation method used in electron microscopy (Op de Beeck *et al.*, 1996), has been implemented in Synchrotron X-ray optics (see **Fig. 3**) to provide two-dimensional phase maps, with micrometer resolution, of objects showing a nearly uniform absorption but introducing an important phase modulation; advantage is taken from the high degree of spatial coherence (due to the small lateral size of the source and the long source-specimen-distance) and the good monochromaticity available on modern Synchrotron beamlines using crystal or multilayer monochromators. The phase maps obtained for different orientations of the object are used as input for a tomographic reconstruction of the three-dimensional distribution of the electron density in the sample. This technique named holotomography (Guigay *et al.*, 2007a; Cloetens *et al.*, 1999; Zabler *et al.*, 2005) is of particular interest in the case of objects opaque to visible light. It has been successfully applied to a large variety of objects of interest in biological or material sciences. For a particular recent application, see Mocella *et al.* (2015).

Conclusion

The AF shares with the WDF the ability to describe the propagation of a partially coherent beam in free space and through a paraxial optical system in a simple and elegant way. The AF is a generalization of the intensity spectra which are the basis of important phase retrieval methods. The images given by a space-invariant system are conveniently described in terms of AF of the object and of the pupil-AF which is a generalization of the OTF and has important applications in the design of phase apodizers.

Phase-space tomography is a growing field of research, in which the AF tomographic reconstruction may be more practical than the WDF tomographic reconstruction.

References

Arrizón, V., Ojeda-Castañeda, J., 1992. Irradiance at Fresnel planes of a phase grating. J. Opt. Soc. Am. A 9, 1801–1806.
Bastiaans, M.J., 1978. The Wigner distribution function applied to optical signals and systems. Opt. Commun. 25 (1), 274–278.
Born, M., Wolf, E., 1999. Principles of Optics, seventh ed. Cambridge university press. [Chapter X].
Brenner, K.-H., Lohmann, A., Ojeda-Castañeda, J., 1983. The ambiguity function as a display of the OTF. Opt. Commun. 44 (5), 323–326.
Castro, A., Ojeda-Castañeda, J., 2004. Asymmetric phase masks for extended depth of field. Appl. Opt. 43 (17), 3474–3479.
Castro, A., Ojeda-Castañeda, J., Lohmann, A., 2006. Bow-tie effect: Differential operator. Appl. Opt. 45 (30), 7878–7884.
Cloetens, P., Ludwig, W., Baruchel, J., *et al*., 1999. Holotomography: Quantitative phase tomography with micrometer resolution using hard synchrotron radiation x rays. Appl. Phys. Lett. 75, 2912–2914.
Dowski, E.R., Cathey, W.T., 1995. Extended depth of field through wave-front coding. Appl. Opt. 34 (11), 1865.
Dragoman, D., Dragoman, M., Brenner, K.H., 2002. Tomographic amplitude and phase recovery of vertical-cavity surface-emitting lasers by use of the ambiguity function. Opt. Lett. 27 (17), 1519–1521.
Dutta, K., Goodman, J.W., 1977. Reconstruction of images of partially coherent objects from samples of mutual intensity. J. Opt. Soc. Am. 67 (6), 796–803.
Fitzgerrell, A.R., Dowski, E.R., Cathey, W.T., 1997. Defocus transfer function for circularly symmetric pupils. Appl. Opt. 36 (23), 5796–5804.
Guigay, J.P., 1971. On Fresnel diffraction by one-dimensional periodic objects, with application to structure determination of phase objects. Opt. Acta 18 (9), 677–682.
Guigay, J.P., 1977a. Fourier transform analysis of Fresnel diffraction patterns and in-line holograms. Optik 49, 121–125.
Guigay, J.P., 1977b. Analyse spectrale (fréquences spatiales) d'une image de diffraction de Fresnel. C. R. Acad. Sc. Paris 284 B. 193–196.
Guigay, J.P., 1978. The ambiguity function in diffraction and isoplanatic imaging by partially coherent beams. Opt. Comm. 26, 136–138.
Guigay, J.P., Langer, M., Boistel, R., Cloetens, P., 2007a. Mixed transfer function and transport of intensity approach for phase retrieval in the Fresnel region. Opt. Lett. 32 (12), 1617–1619.
Guigay, J.P., Pagot, E., Cloetens, P., 2007b. Fourier Optics approach to X-ray analyser-based imaging. Opt. Comm. 270, 180–188.
Hanszen, K.J., 1972. In-line-holographische Erfahrungenmit Radialgittern als Testobjekten in lichtoptischen Modellanordungen für das Elektronenmikroskop. Optik 36, 41–54.
Liu, X., Brenner, K.H., 2003. Reconstruction of two-dimensional complex amplitudes from intensity measurements. Opt. Commun. 225, 19–30.
Liu, X., Hruscha, C., Brenner, K.H., 2008. Efficient reconstruction of two-dimensional complex amplitudes utilizing the redundancy of the ambiguity function. Appl. Opt. 47 (22), E1–E7. (Feature issue on Phase-space representations in Optics).
Mocella, V., Brun, E., Ferrero, C., Delattre, D., 2015. Revealing letters in rolled Herculaneum papyri by X-ray phase contrast imaging. Nat. commun. 5895.
Nugent, K.A., 2007. X-ray noninterferometric phase imaging: A unified picture. J. Opt. Soc. Am. A 24, 536–547.
Ojeda-Castañeda, J., Andrés, P., Montes, E., 1987. Phase-space representation of the Strehl ratio: Ambiguity function. J. Opt. Soc. Am. 4 (2), 313–317.
Ojeda-Castañeda, J., Berriel-Valdos, L.R., 1988. Ambiguity function as a design tool for high focal depth. Appl. Opt. 27, 790–795.
Ojeda-Castañeda, J., Gómez-Sarabia, C.M., 2015. Tuning field depth at high resolution by pupil engineering. Adv. Opt. Photonics 7 (4).
Ojeda-Castañeda, J., Landgrave, J.E.A., Escamilla, H.M., 2005. Annular phase-only mask for high focal depth. Opt. Lett. 30, 1647–1649.
Ojeda-Castañeda, J., Sicre, E., 1984. Bilinear systems: Wigner distribution function and ambiguity function representations. Opt. Acta 31 (3), 255–260.
Op de Beeck, M., Van Dyck, D., Coene, W., 1996. Wave-function reconstruction in HRTEM: The parabola method. Ultramicroscopy 64, 167–183.
Papoulis, A., 1974. Ambiguity function in Fourier optics. J. Opt. Soc. Am. 64, 779–788.
Raymer, M.G., Beck, M., McAlister, D.F., 1994. Complex Wave-Field Reconstruction using Phase-Space Tomography. Phys. Rev. Lett. 72 (8), 1137–1140.
Sheppard, C.J.R., Larkin, K.G., 2000. Focal shift, optical transfer function and phase-space representations. J. Opt. Soc. Am. A 17 (4), 772–779.
Simons, H., Ahl, Sonja Rosenlund, Poulsen, H.F., Detlefs, C., 2017. Simulating and optimizing compound refracted-based X-ray microscopes. J. Synchrotron Radiat. 24, 392–401.
Teague, M.R., 1983. Deterministic phase retrieval: A Green's function solution. J. Opt. Soc. Am. 73, 1434.
Terentyev, S., Polikarpov, M., Snigireva, I., *et al*., 2017. Linear parabolic single-crystal diamond refractive lenses for synchrotron X-ray sources. J.Synchrotron Radiat. 24, 103–109.
Testorf, M., Hennelly, B., Ojeda-Castañeda, J., 2010. Phase-Space Optics, Fundamentals and Applications. New York, NY: Mc Graw-Hill.
Tran, C.Q., Peele, A.G., Roberts, A., *et al*., 2005. X-ray imaging: A generalized approach using phase-space tomography. J. Opt. Soc. Am 22 (8), 1691–1700.
Tu, J., Tamura, S., 1997. Wave field determination using tomography of the ambiguity function. J. Opt. Soc. Am. 55 (2), 1946–1949. (1998. Analytic relation for recovering the mutual intensity by means of intensity information. J. Opt. Soc. Am. 15 (1), 202–206).
Wade, R.W., 1974. Spectral analysis of holograms and reconstructed images. Optik 40, 201.
Woodward, P.M., 1953. Probability and Information Theory With Application to Radar. New York, NY: Pergamon.
Zabler, S., Cloetens, P., Guigay, J.P., Baruchel, J., Schlenker, M., 2005. Optimization of phase contrast imaging using hard x rays. Rev. Sci. Instrum. 76, 073705.

Further Reading

Brenner, K.H., Ojeda-Castañeda, J., 1984. Ambiguity function and Wigner distribution function applied to partially coherent imagery. Opt. Acta 31 (2), 213–223.
Semichaevsky, A., Testorf, M., 2004. Phase-space interpretation of deterministic phase retrieval. J. Opt. Soc. Am. A21, 2173–2179.

Phase Space Tomography in Optics

Tatiana Alieva and José A Rodrigo, Complutense University of Madrid, Madrid, Spain
Antonio Picón, University of Salamanca, Salamanca, Spain

Introduction

Tomographic techniques in a wide sense consist in the reconstruction of a $n-$dimensional (nD) object from a set of data of lower, usually $(n-1)$, dimension. They are widely applied in medicine and technology for noninvasive inspection of internal structure of macro- and microorganisms and other 3D samples. There exist different approaches developed in optics to obtain this valuable information: computed tomography (CT), optical diffraction tomography, optical coherence tomography, to name the most known ones. Phase space tomography (PST) is intrinsically different from the mentioned tomographic modalities, because it does not directly serve for the tomographic analysis of the sample structure. Meanwhile it is focused on the tomographic study of the light beam itself. From mathematical point of view the PST is closer to the CT technique since both of them are based on the rotation of the object under inspection and measuring a set of its projections. However, the object of reconstruction in the CT and the PST are completely different. In the case of the CT it is the distribution of the absorption coefficient of a physically existed 3D object. While in the case of the PST the object is the Wigner distribution function (WD) which describes coherence properties of classical light (Bastiaans, 2008) or defines a quantum state (density matrix) of light source in quantum systems (D'Ariano *et al.*, 2003). Neither the light coherence characteristics nor density matrix can be measured directly. This information, required for the development of imaging techniques, telecommunication protocols, quantum computing methods, etc., is possible to recover applying the PST tools. The WD is a real function but it can take negative values that impedes its direct measurements. Nevertheless, the projections of the WD are always non-negative and are associated with measurable quantities. This fact together with rotation of the WD after light propagation through certain optical systems described by canonical integral transforms form the basis of the PST (Raymer *et al.*, 1994). Measuring a proper set of the WD projections and using the inverse Radon transform, or other tomography reconstruction algorithm, the WD, and therefore the related beam characteristics are obtained. Below the application of the PST for the analysis of spatial structure of optical beams, short pulses and quantum states are considered.

Light Beam Characterization

Let us consider a linearly polarized optical beam propagated in z direction. The only quantity which can be measured directly is its time-averaged intensity distribution, because the oscillations of the electromagnetic field are too fast for any currently available detectors. However, from the knowledge of the intensity distribution $I(\mathbf{r})$ in a certain plane $z=z_1$ (here we have defined a 2D position vector $\mathbf{r}=[x,y]^t$ at the xy plane transverse to the propagation direction z, where the superscript t denotes the transposition operation) it is impossible to predict the value of $I(\mathbf{r})$ in other plane $z=z_2$ or to describe the beam interaction with a sample under study. In order to do this in the case of quasi-monochromatic coherent beam we need to know its complex field amplitude $f(\mathbf{r})=a(\mathbf{r})\exp[i\varphi(\mathbf{r})]$, where $a(\mathbf{r})$ and $\varphi(\mathbf{r})$ refereed to as amplitude and phase are real valued. In this case $I(\mathbf{r})=|f(\mathbf{r})|^2=a^2(\mathbf{r})$ and therefore the amplitude $a(\mathbf{r})$ can be easily measured $a(\mathbf{r})=\sqrt{I(\mathbf{r})}$ using, for example, a digital camera while the determination of the phase $\varphi(\mathbf{r})$ requires some special interferometric or iterative techniques. However, in the most cases light is not fully coherent and its phase is stochastically varying. The description of such partially coherent light requires the application of statistical methods. Often it is assumed that light follows Gaussian statistics and therefore is described by two-point correlation function $\Gamma(\mathbf{r}_1,\ \mathbf{r}_2)=\langle f(\mathbf{r}_1)f^*(\mathbf{r}_2)\rangle$ known as mutual intensity (MI) (Goodman, 2000). Here $\langle\cdot\rangle$ stands for ensemble averaging which for ergodic process can be substituted by time averaging. The MI is a non-negative definite Hermitian function of four variables $\Gamma(\mathbf{r}_1,\ \mathbf{r}_2)=\Gamma^*(\mathbf{r}_2,\ \mathbf{r}_1)$. Note that only the values $\Gamma(\mathbf{r},\mathbf{r})=\langle|f(\mathbf{r})|^2\rangle$, corresponding to the intensity distribution, can be directly measured. For the polychromatic light, similar expression known as cross-spectral density, corresponding to the temporal Fourier transform of the mutual coherence function is used for beam description. The MI gauges the field correlation at the points $\mathbf{r}_1$ and $\mathbf{r}_2$ via complex coherence factor (Goodman, 2000): $\gamma(\mathbf{r}_1,\mathbf{r}_2)=\Gamma(\mathbf{r}_1,\mathbf{r}_2)/\sqrt{\Gamma(\mathbf{r}_1,\mathbf{r}_1)\Gamma(\mathbf{r}_2,\mathbf{r}_2)}$.

Instead of the MI, its Fourier transform (FT) with respect to the position difference $\mathbf{r}'=\mathbf{r}_1-\mathbf{r}_2$, known as Wigner distribution (WD) (Bastiaans, 2008)

$$W(\mathbf{r},\mathbf{p})=\frac{1}{\sigma^2}\int \mathrm{d}\mathbf{r}'\Gamma\left(\mathbf{r}-\frac{\mathbf{r}'}{2},\ \mathbf{r}+\frac{\mathbf{r}'}{2}\right)\exp\left(-\frac{i2\pi}{\sigma^2}\mathbf{p}^t\mathbf{r}'\right) \tag{1}$$

or the FT with respect to $\mathbf{r}=(\mathbf{r}_1+\mathbf{r}_2)/2$ known as Ambiguity function (AF)

$$A(\mathbf{r}',\mathbf{p})=\frac{1}{\sigma^2}\int \mathrm{d}\mathbf{r}\Gamma\left(\mathbf{r}-\frac{\mathbf{r}'}{2},\mathbf{r}+\frac{\mathbf{r}'}{2}\right)\exp\left(-\frac{i2\pi}{\sigma^2}\mathbf{p}^t\mathbf{r}\right) \tag{2}$$

 doi:10.1016/B978-0-12-803581-8.09376-0

can be used for beam description. Here $\mathbf{p}=[u,v]^t=\sigma^2\mathbf{k}_\perp$ is a vector proportional to the transverse projection of the spatial frequency vector, $\mathbf{k}_\perp$, and σ is a convenient constant with units of length that is commonly used in Fourier optics. The WD of 2D field is a real valued function of four Cartesian coordinates (x, y, u, v), which form so called phase space. It follows from the Eqs. (1–2) that the MI can be recovered from the WD (or the AF) applying the inverse FT

$$\begin{aligned}\Gamma(\mathbf{r}_1,\mathbf{r}_2)&=\frac{1}{\sigma^2}\int \mathrm{d}\mathbf{p}W\left(\frac{\mathbf{r}_1+\mathbf{r}_2}{2},\,\mathbf{p}\right)\exp\left(\frac{i2\pi}{\sigma^2}\mathbf{p}^t(\mathbf{r}_1-\mathbf{r}_2)\right)\\&=\frac{1}{\sigma^2}\int \mathrm{d}\mathbf{p}A(\mathbf{r}_1-\mathbf{r}_2),\,\mathbf{p})\exp\left(\frac{i2\pi}{\sigma^2}\mathbf{p}^t(\mathbf{r}_1+\mathbf{r}_2)\right)\end{aligned}\tag{3}$$

Moreover, if the light in completely coherent, $|\gamma(\mathbf{r}_1,\,\mathbf{r}_2)|=1$ and $\Gamma(\mathbf{r}_1,\,\mathbf{r}_2)=f(\mathbf{r}_1)f^*(\mathbf{r}_2)$, then the phase of the complex field amplitude up to the constant φ_0 can be obtained from the MI

$$\varphi(\mathbf{r})=\varphi_0+\arg[\Gamma(\mathbf{r},\mathbf{r}_0)]\tag{4}$$

where $\mathbf{r}_0$ is a point in which the field intensity is not zero. Note that the phase information is very demanded, for example, in optical metrology and quantitative microscopy.

PST Fundamentals

Actually developed detectors allows measuring only beam intensity distribution and therefore no one of the mentioned functions, MI, WD and AF, can be measured directly: the MI and AF are complex valued function, while the WD is real but takes non-negative values only in the case of coherent Gaussian beam. Nevertheless, the projection of the WD on the (x, y) − plane corresponds to the beam intensity distribution

$$I(\mathbf{r})=\frac{1}{\sigma^2}\int \mathrm{d}\mathbf{p}W(\mathbf{r},\mathbf{p})\tag{5}$$

and therefore is measurable. On the other side the intensity distribution is the FT of the AF slice

$$I(\mathbf{r})=\frac{1}{\sigma^2}\int \mathrm{d}\mathbf{p}A(0,\mathbf{p})\exp\left(\frac{i2\pi}{\sigma^2}\mathbf{p}^t\mathbf{r}\right)\tag{6}$$

The PST method based on the recovery of the AF is considered, for example, in Cámara *et al.* (2014), Tran *et al.* (2005), and Tu and Tamura (1997).

In order to apply the PST method for beam characterization the WD (or the AF) has to be rotated and a set of corresponding projections have to be measured. In paraxial approximation the beam propagation through a first-order optical system – an optical system composed by centered spherical and cylindrical lenses and/or mirrors separated by homogeneous medium – is described by the 2D linear canonical integral transform $f_{\mathbf{T}}(\mathbf{r}_o)=\int K_{\mathbf{T}}(\mathbf{r}_i,\,\mathbf{r}_o)f(\mathbf{r}_i)\mathrm{d}\mathbf{r}_i$ where $f(\mathbf{r}_i)$ and $f_{\mathbf{T}}(\mathbf{r}_o)$ are the complex field amplitude in the input and output planes of the system and

$$K_{\mathbf{T}}(\mathbf{r}_i,\mathbf{r}_o)=\frac{1}{\sigma^2\sqrt{\det(i\mathbf{B})}}\exp\left[\frac{i\pi}{\sigma^2}\left(\mathbf{r}_o^t\mathbf{D}\mathbf{B}^{-1}\mathbf{r}_o-2\mathbf{r}_i^t\mathbf{B}^{-1}\mathbf{r}_o+\mathbf{r}_i^t\mathbf{B}^{-1}\mathbf{A}\mathbf{r}_i\right)\right]\tag{7}$$

for $\det\mathbf{B}\neq 0$ and

$$K_{\mathbf{T}}(\mathbf{r}_i,\mathbf{r}_o)=\frac{1}{\sqrt{|\det\mathbf{A}|}}\exp\left[\frac{i\pi}{\sigma^2}\mathbf{r}_o^t\mathbf{C}\mathbf{A}^{-1}\mathbf{r}_o\right]\delta(\mathbf{r}_i-\mathbf{A}^{-1}\mathbf{r}_o)\tag{8}$$

for $\mathbf{B}=\mathbf{0}$ Here $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$ are 2×2 sub-matrices of the 4×4 dimensionless ray transformation matrix $\mathbf{T}$, which provides the relation between the position $\mathbf{r}$ and direction (proportional to $\mathbf{p}$) of the ray in the input ($\mathbf{r}_i$ and $\mathbf{p}_i$) and output ($\mathbf{r}_o$ and $\mathbf{p}_o$) planes of an optical system:

$$\begin{bmatrix}\mathbf{r}_o\\ \mathbf{p}_o\end{bmatrix}=\begin{bmatrix}\mathbf{A}&\mathbf{B}\\ \mathbf{C}&\mathbf{D}\end{bmatrix}\begin{bmatrix}\mathbf{r}_i\\ \mathbf{p}_i\end{bmatrix}=\mathbf{T}\begin{bmatrix}\mathbf{r}_i\\ \mathbf{p}_i\end{bmatrix}\tag{9}$$

Note, that for the matrix describing the linear ray propagation in free space

$$\mathbf{T}_z=\begin{bmatrix}\mathbf{I}&\frac{z}{\lambda}\mathbf{I}\\ \mathbf{0}&\mathbf{I}\end{bmatrix}\tag{10}$$

where $\mathbf{I}$ is a unity matrix, λ is the wavelength and z is the distance between the input and output planes, the expression Eq. (7) with $\sigma=\lambda$ is reduced to the kernel of the Fresnel transform. Meanwhile the matrix

$$\mathbf{T}_L=\begin{bmatrix}\mathbf{I}&\mathbf{0}\\ \mathbf{C}&\mathbf{I}\end{bmatrix}\tag{11}$$

corresponds to spherical/cylindrical thin lens or mirror action yielding to quadratic phase modulation of the wavefield according to the Eq. (8) with $\sigma=\lambda$. For example, the sub-matrix $\mathbf{C}=\left[-\lambda f_x^{-1},0;0,-\lambda f_y^{-1}\right]$ describes the orthogonal superposition of two

convergent cylindrical lenses with front focal distances f_x and f_y, that for $f_x = f_y$ corresponds to the spherical lens. Two assembled cylindrical lenses with transverse axes forming any other angle, called *generalized lens*, is associated with symmetric sub-matrix $\mathbf{C} = [c_{xx}, c_{xy}; c_{xy}, c_{yy}]$. Using a set of cylindrical lenses and choosing appropriate distances between them optical system performing any 2D canonical integral transform can be constructed.

The WD, $W_{\mathbf{T}}(\mathbf{r},\mathbf{p})$, of the beam at the output plane of a first-order system described by the ray transformation matrix $\mathbf{T} = [\mathbf{A},\mathbf{B};\mathbf{C},\mathbf{D}]$, is related to the WD at the input plane, $W(\mathbf{r},\mathbf{p})$, by an affine transformation (Bastiaans, 2008)

$$W_{\mathbf{T}}(\mathbf{r},\mathbf{p}) = W(\mathbf{D}^t\mathbf{r} - \mathbf{B}^t\mathbf{p}, -\mathbf{C}^t\mathbf{r} + \mathbf{A}^t\mathbf{p}) \tag{12}$$

which, in general, includes rotation, shearing and scaling in phase space. Similar expression is valid for the AF

$$A_{\mathbf{T}}(\mathbf{r},\mathbf{p}) = A(\mathbf{D}^t\mathbf{r} - \mathbf{B}^t\mathbf{p}, -\mathbf{C}^t\mathbf{r} + \mathbf{A}^t\mathbf{p}) \tag{13}$$

In particular, the rotations in phase space, which are described by the ray transformation matrix $\mathbf{T}_R$ with parameters $\mathbf{A} = \mathbf{D}$ and $\mathbf{B} = -\mathbf{C}$, are needed to apply the PST method. There are two principle kind phase space rotators: the fractional Fourier transform (FrFT) (Ozaktas *et al.*, 2001) associated with the matrix $\mathbf{T}_{FrFT}(\alpha_x, \alpha_y)$ defined by

$$\begin{aligned} \mathbf{A}_{FrFT} = \mathbf{D}_{FrFT} &= \begin{bmatrix} \cos\alpha_x & 0 \\ 0 & \cos\alpha_y \end{bmatrix} \\ \mathbf{B}_{FrFT} = -\mathbf{C}_{FrFT} &= \begin{bmatrix} \sin\alpha_x & 0 \\ 0 & \sin\alpha_y \end{bmatrix} \end{aligned} \tag{14}$$

which describes the WD rotation in the planes xu and yv for angles $\alpha_{x,y}$, respectively, and image rotator (IR) associated with the matrix $\mathbf{T}_{IR}(\beta)$ defined by

$$\mathbf{A}_{IR} = \mathbf{D}_{IR} = \begin{bmatrix} \cos\beta & \sin\beta \\ -\sin\beta & \cos\beta \end{bmatrix} \text{ and } \mathbf{B}_{IR} = \mathbf{C}_{IR} = \mathbf{0} \tag{15}$$

which corresponds to the simultaneous WD rotation in xy and uv for angle β. Their combination defines any other phase space rotator associated with $\mathbf{T}_R(\gamma, \alpha_x, \alpha_y, \beta) = \mathbf{T}_{IR}(\gamma)\mathbf{T}_{FrFT}(\alpha_x, \alpha_y)\mathbf{T}_{IR}(\beta)$. Note that for the PST applications the parameter γ is irrelevant since it describes the rotation of the coordinate system in the output xy plane. The intensity distribution in the output plane of a phase space rotator corresponds to the WD projection given by

$$I_{\mathbf{T}_R}(\mathbf{r}) = \langle |f_{\mathbf{T}_R}(\mathbf{r})|^2 \rangle = \frac{1}{\sigma^2}\int \mathrm{d}\mathbf{p} W_{\mathbf{T}_R}(\mathbf{r},\mathbf{p}) = \frac{1}{\sigma^2}\int \mathrm{d}\mathbf{p}\, W\left(\mathbf{D}_R^t\mathbf{r} - \mathbf{B}_R^t\mathbf{p}, -\mathbf{C}_R^t\mathbf{r} + \mathbf{A}_R^t\mathbf{p}\right) \tag{16}$$

For complete WD recovery using the inverse Radon transform, or other tomography reconstruction algorithm a set of 2D projections $\{I_{\mathbf{T}_R}(\mathbf{r})\}$ for two independent angles (both covering a $\pi-$interval) associated with the rotation in two orthogonal planes of the phase space is needed. Initially the PST method was established for the projection set corresponding to the FRFT, $\mathbf{T}_R(0, \alpha_x, \alpha_y, 0) = \mathbf{T}_{FrFT}(\alpha_x, \alpha_y)$ (Raymer *et al.*, 1994) while in Cámara (2015) and Cámara *et al.* (2013) it has been shown that the phase space rotator described by $\mathbf{T}_R(0, 0, \alpha, \beta)$ or by $\mathbf{T}_R(0, \alpha, 0, \beta)$ with

$$\begin{aligned} \mathbf{A}_R(0,\alpha,0,\beta) &= \begin{bmatrix} \cos\alpha\cos\beta & \cos\alpha\sin\beta \\ -\sin\beta & \cos\beta \end{bmatrix} \\ \mathbf{B}_R(0,\alpha,0,\beta) &= \begin{bmatrix} \sin\alpha\cos\beta & \sin\alpha\sin\beta \\ -\sin\beta & \cos\beta \end{bmatrix} \end{aligned} \tag{17}$$

can be applied.

While the theoretical fundamentals of the PST method seem straightforward its experimental realization is rather difficult since real-world applications require fast acquisition and processing of the experimental data together with comprehensive analysis of the resulting four-dimensional functions associated to the light coherence information. Any a priori information about a studied beam can simplify the task and decrease the number of projections required for WD reconstruction. Some methods assume certain hypothesis about field model (Tian *et al.*, 2012), its symmetry (Agarwal and Simon, 2000; Alieva *et al.*, 2011; Cámara, 2015; Cámara *et al.*, 2014) or coherence homogeneity. Below two schemes for practical realization of the PST analysis of the beam coherence state in the absence of a priori information are considered.

PST for 1D Beam Characterization

The application of the PST for beam analysis meets inherent difficulty caused by the high dimension of the WD and related functions. It yields to the excessive computation effort to process the measured data (WD projections) into the requested result (the WD, AF or MI). One of solutions of this complex problem is to decompose a 2D beam into 1D signals and recover their WDs. In order to select a 1D signal a thin slit is usually used. This mask has to be moved and rotated in the xy input plane in order to

explore consequently the beam coherence properties along the corresponding lines. The WD of a selected 1D signal described by the complex field amplitude $f(x)$ is a 2D function, $W(x, u)$, that can be easily presented graphically and analyzed. In the associated 2D phase space xu there is only one phase space rotator: the 1D FrFT whose the kernel is written as

$$K_\alpha(x_i, x) = \frac{1}{\sigma\sqrt{i \sin\alpha}} \exp\left[\frac{i\pi}{\sigma^2}\frac{(x^2 + x_i^2)\cos\alpha - 2xx_i}{\sin\alpha}\right] \tag{18}$$

The 1D FrFT produces the rotation of the WD

$$W_\alpha(x, u) = W(x\cos\alpha - u\sin\alpha,\ x\sin\alpha + u\cos\alpha) \tag{19}$$

Measuring the WD projections

$$P_\alpha(x) = \frac{1}{\sigma}\int \mathrm{d}u W(x\cos\alpha - u\sin\alpha,\ x\sin\alpha + u\cos\alpha) \tag{20}$$

for angles covered a $\pi-$interval $\alpha\in[0,\pi]$ which corresponds to the Radon transform of the WD (also called as Radon-Wigner transform) the $W(x, u)$ can be reconstructed applying, for example, the filtered back-propagation algorithm

$$W(x, u) = \frac{1}{\sigma^2}\int_0^\pi \int \mathrm{d}\alpha\mathrm{d}\rho|\rho|P_\alpha(\rho)\exp\left[-\frac{2\pi}{\sigma^2}\rho(x\cos\alpha - u\sin\alpha)\right] \tag{21}$$

If the projections are available for the interval $\alpha\in[\alpha_0, \alpha_0+\pi]$ then the WD, $W(x\cos\alpha_0 - u\sin\alpha_0, x\sin\alpha_0 + u\cos\alpha_0)$ reconstructed by this expression (the integration limits in this case are from α_0 to $\alpha_0+\pi$) is rotated at angle α_0 with respect to one, $W(x, u)$, described the input signal. The MI is obtained correspondingly as

$$\Gamma(x_1, x_2) = \frac{1}{\sigma}\int \mathrm{d}u W\left(\frac{x_1 + x_2}{2}, u\right)\exp\left(\frac{i2\pi}{\sigma^2}u(x_1 - x_2)\right) \tag{22}$$

The principal advantage of this method is a possibility to registered in one shot all required projections of the $W(x,u)$ in the digital camera display. For this purpose a 1D signal is converted into a 2D one, which is its multiple-copy and varifocal lenses or Fresnel zone plates are applied to perform the FrFT of every copy for different angle. Several designs for such system, known as Radon-Wigner display (RWD), have been proposed (Cámara *et al.*, 2011; Granieri *et al.*, 1997; Mendlovic *et al.*, 1996). One of them is discussed below. While this method is suitable for analysis of 1D optical signals it has important drawbacks when it is applied for the 2D beams. The slit mask application introduces artifacts. The conversion of 1D signal into a 2D one produces significant power reduction of the analyzed field and therefore decreases the signal to noise ratio. For analysis of the beam coherence properties along another line of xy plane the displacement and/or rotation of the slit mask and the RWD are required.

PST for 2D Beam Characterization

Another approach for practical application of the PST is based on an appropriate choice of the projections set and the order of their acquisition. As it was mentioned above several projection sets can be used for the reconstruction of the 4D-WD. As well as for the tomographic analysis of horizontal slices of the 3D object it is beneficial rotate the object along the vertical axes, there exists the 4D-WD projection set which is more appropriate for the beam coherence analysis. This set, hereinafter referred as $\{P_{\alpha,\beta}(\mathbf{r})\}$, is obtained by rotating the WD in the $x\gamma$ and xu planes for angles β and α, respectively (phase space rotator described by the ray transformation matrix $\mathbf{T}_R(0,\alpha,0,\beta)$) (Cámara, 2015; Cámara *et al.*, 2013).

The projection set $\{P_{\alpha,\beta}(\mathbf{r})\}$, comprises several subsets, $\{P_{\alpha,\beta_0}(\mathbf{r})\}$, defined by fixing $\beta=\beta_0$, which are acquired consecutively. Each subset provides the information about field correlation at whichever points $\mathbf{r}_1$ and $\mathbf{r}_2$ that are contained in a line which forms an angle β_0 with axes x. The MI at such a line is defined as $\Gamma_{\beta_0}(\mathbf{r}_0, s) = \Gamma(\mathbf{r}_0, \mathbf{r}_0 + s\mathbf{n})$, where $\mathbf{r}_0$ is the reference point, $\mathbf{n}=[\cos\beta_0, \sin\beta_0]^t$ and s are the direction and running coordinate of the line, respectively. For example, from one slice of the projection subset $\{P_{\alpha,0}([x, \gamma_0]^t)\}$ for any value of γ_0 (see **Fig. 1**) the 2D-WD, $W(x,u;\gamma_0)$, of the 1D signal $f(x;\gamma_0)$ is obtained using the Eq. (21), where

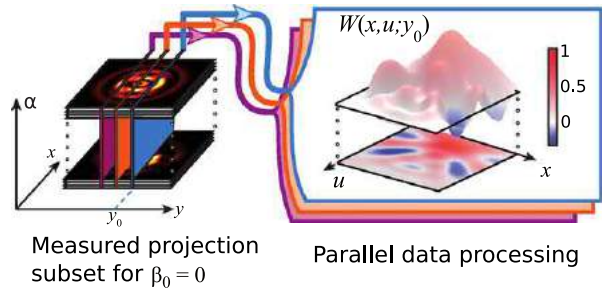


Fig. 1 Scheme explaining the parallel data processing for WD reconstruction. The Wigner distribution $W(x,u;y_0)$ of the field along a chosen line $y=y_0$ is reconstructed for each projection subset slice defined by both β_0 and y_0 (see, e.g., the blue slice). Note that the Wigner distribution takes positive and negative values (white color indicates zero level).

$P_\alpha(\rho)=P_{\alpha,0}([\rho,y_0]^t)$. The correlation function for the field at reference point $\mathbf{r}_0=[x_0,y_0]^t=\mathbf{r}_1$ and any other point $\mathbf{r}_2=\mathbf{r}_0+s[1,0]^t$ contained in the horizontal line $y=y_0$ is then calculated as

$$\Gamma_0(\mathbf{r}_0,s)=\frac{1}{\sigma}\int \mathrm{d}u W(x_0+s/2,u;y_0)\exp\left(\frac{\mathrm{i}2\pi}{\sigma^2}su\right) \tag{23}$$

Note that from the same subset $\{P_{\alpha,0}([x,y]^t)\}$ the MI for any two points contained in the same horizontal line ($y=$constant) is obtained. To find the coherence relations for two points with different y coordinates, one has to perform the rotation in the xy plane for the corresponding angle β_0 and to repeat the procedure described above using the projection subset $\{P_{\alpha,\beta_0}(\mathbf{r})\}$. Doing this, we recover the desired information about field correlation avoiding the reconstruction of the entire 4D-WD and its posterior processing as it needed, for example, if the projection set corresponding to 2D-FrFT ($\mathbf{T}_R(0,\ \alpha_x,\ \alpha_y,\ 0)=\mathbf{T}_{FrFT}(\alpha_x,\ \alpha_y)$) is used. Although the amount of data required for full beam characterization is the same as in other realizations of the PST, this choice of projection set, $\{P_{\alpha,\beta}(\mathbf{r})\}$, brings powerful benefits: (1) The data acquisition and processing tasks are performed simultaneously since every projection subset is an independent entity. (2) As one projection subset slice is independent of the rest, parallel computing can be used for the data processing. (3) Physically meaningful information is obtained from a reduced number of projections that allows starting the beam analysis before the complete projection set is acquired.

Optical Setups for PST

Several experimental phase space rotators described by ray transformation matrix $\mathbf{T}_R$ have been proposed. All of them are based on a combination of lens (or other optical elements producing quadratic phase modulation of wavefield) separated by the free space intervals. Since only the intensity distribution is measured then the phase in the output plane of the optical system applied for the PST can be arbitrary. This allows increasing a number of suitable setups and often simplifying their design. Indeed the intensity distributions in the output of the systems described by the ray transformation matrices $\mathbf{T}_L\mathbf{T}_R$ and $\mathbf{T}_R$ coincide and depend only on $\mathbf{T}_R$. The scanning of the rotation angles is achieved by modification of the lens powers or distances between them. However, the insert of a lens or change of the distance usually require the system alignment that is incompatible for a fast projection acquisition required for the practical application of the PST. Moreover, for accurate WD reconstruction it is recommended to avoid the projection re-scaling. All these conditions can be fulfilled if digital lenses implemented by spatial light modulator (SLM) are used. Indeed the phase-only SLM easily performs a quadratic phase wavefront modulation that is an action of a thin lens. This allows designing an optical setup with all elements located in fixed position and a programmable control of the projection parameters. In this case the speed of the projection acquisition is limited by the SLM and digital camera (usually video rate).

Different projections sets including ones discussed above (corresponding to $\mathbf{T}_R(0,\ \alpha_x,\ \alpha_y,\ 0)$ and $\mathbf{T}_R(0,\ \alpha,\ 0,\ \beta)$) can be automatically acquired using the programmable optical setup (Cámara *et al.*, 2013; Rodrigo *et al.*, 2009) shown in **Fig. 2**. The setup comprises two generalized lenses implemented by two SLMs and a digital camera. The distance z between every two consecutive elements is fixed. Specifically, for acquisition the projection set $\{P_{\alpha,\beta}(\mathbf{r})\}$ these lenses L_j ($j=1,2$) have the following transmission functions

$$L_j(x,y)=\exp\left[-i\pi\frac{(x\cos\beta+y\sin\beta)^2}{\lambda f_j}\right]\exp\left[-i\pi\frac{(-x\sin\beta+y\cos\beta)^2}{\lambda g_j}\right] \tag{24}$$

where the focal lengths are $f_1=2z/(2-\cot(\alpha/2))$, $f_2=z/(2-2\sin\alpha)$ and $g_1=z$, $g_2=z/2$, are given as a function of the transformation angle $\alpha\in[\pi/2,\ 3\pi/2]$. This setup is described by the matrix $\mathbf{T}_L\mathbf{T}_R(-\beta,\ \alpha,\ 0,\ \beta)$. Then the intensity distribution in its output plane is equal to one corresponding to the desired transformation $\mathbf{T}_R(0,\ \alpha,\ 0,\ \beta)$ except for a rotation at an angle $-\beta$. To compensate this effect and obtain the corresponding WD projection the image acquired by the digital camera has to be rotated an angle β that can be easily performed digitally. The number of the acquired projections which is one of the principle parameters defining the resolution of the WD reconstruction can easily reach a value 180 per one subset $\{P_{\alpha,\beta_0}(\mathbf{r})\}$.

The varifocal lens needed for RWD implementation can also be realized by the SLM (Cámara, 2015; Cámara *et al.*, 2011). An illustration of such system is displayed in **Fig. 3**. It consists in two SLMs separated a fixed distance z, and two cylindrical lenses of

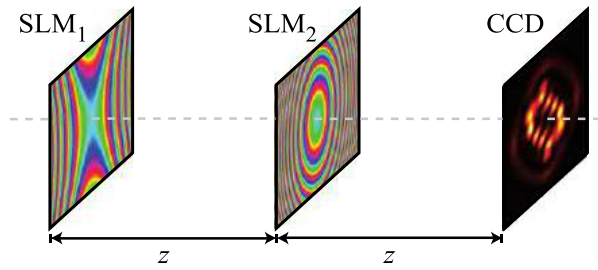


Fig. 2 Scheme of a programmable setup for complete WD projection set acquisition, required for the WD reconstruction. For every β and α the SLM_1 and SLM_2 implement generalized lenses L_1 and L_2, respectively, given by Eq. (24).

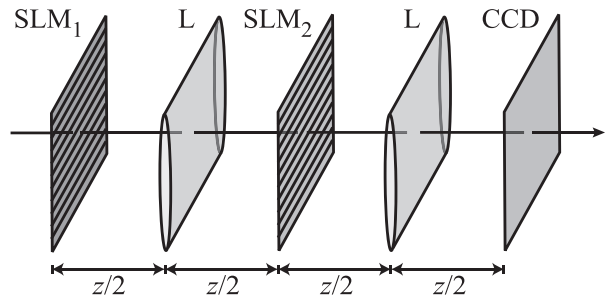


Fig. 3 Scheme of a programmable Radon Wigner display using spatial light modulators. SLMs implement N 1D lenses described by Eqs. (25)–(26).

focal length $z/4$ with phase modulation in the y direction also separated a fixed distance z. N copies of the input 1D signal $f(x)$ arranged in set of N vertical channels are projected into input plane of the RWD, which coincides with the position of the first SLM. The SLMs implement the multichannel phase masks. Each channel of the phase mask is programmed to perform the 1D-FrFT operation for an arbitrary angle in the range $\alpha \in [\pi/2, 3\pi/2]$. In order to acquire the full-range WD projections for N equidistant angles, the nth channel of the jth SLM implements the transmission function $L_j(x,n)$ given by:

$$L_1(x,n) = \exp\left[-i\pi\frac{x^2}{\lambda z}\left(1 - \frac{1}{2}\cot\frac{\alpha_n}{2}\right)\right] \tag{25}$$

$$L_2(x,n) = \exp\left[-\mathrm{i}2\pi\frac{x^2}{\lambda z}(1 - \sin\alpha_{N-n+1})\right] \tag{26}$$

where $n = 1,\ldots,N$ and $\alpha_n = \pi/2 + (n-1)/N\pi$ is the FrFT angle associated with the jth channel. While the SLMs perform the FrFT in the x direction, the glass cylindrical lenses compensate the beam propagation in the y direction by imaging the first SLM into the second, and the second into the output plane. Since the first cylindrical lens inverts the channel position, the nth channel of the second SLM implements the angle labeled by $N - n + 1$ in Eq. (26). At the output plane, placed at a distance $z/2$ from the last cylindrical lens, a digital camera registers in one shot all required WD projections of the input 1D signal. The number of channels and therefore the WD projections depend on the size of the SLM and camera displays and can easily reach $N = 300$ if the 2D expansion of the 1D signal is energetically resolved.

Chronocyclic Tomography for Optical Pulse Characterization

We have considered the application of the phase space tomography for characterization of the spatial light coherence, meanwhile it also can be used for the analysis of short optical pulses (Beck *et al.*, 1993; Dorrer and Kang, 2003). In this case the time t and temporal frequency v are coordinates of the phase space and the method of the corresponding WD recovery is called chronocyclic tomography. The electrical field of an optical pulse is described by

$$E(t) = |E(t)|\exp[i\phi_t(t)]\exp(-i2\pi v_0 t) \tag{27}$$

where $|E(t)|$ and $\phi_t(t)$ are the time-dependent pulse envelop and phase correspondingly and v_0 is the carrier frequency. The instantaneous power $|E(t)|^2$ can be measured only if the pulse duration is sufficiently larger than the response time of the photodetector. Alternatively the pulse can be described in frequency domain as $\tilde{E}(v) = |\tilde{E}(v)|\exp[i\phi_v(v)]$, where $\tilde{E}(v)$ is a FT of the $E(t)$, $\phi_v(v)$ is a spectral phase. The spectral density $|\tilde{E}(v)|^2$ can be easily measured by a spectrometer. Usually an ensemble of pulses are involved in the experiments and its statistical properties, in particular, two-time correlation function $\Gamma(t_1,t_2) = \langle E(t_1)E^*(t_2)\rangle$ or the related WD

$$W(t,v) = \eta\int \mathrm{d}t'\Gamma(t - t'/2, t + t'/2)\exp(-\mathrm{i}2\pi t' v) \tag{28}$$

are of interest. Here η is a convenient constant with units of frequency. If all pulses are identical then the amplitude and the phase of $E(t)$ up to the constant factor can be retrieved from $\Gamma(t_1,t_2)$.

As well as in the spatial PST case the devices which produce the rotation of the WD, $W(t,v)$, are needed for chronocyclic tomography. A quadratic temporal phase modulator and the propagation in dispersive medium or two-grating compressor (quadratic spectral phase modulation) are the analogues of lens and Fresnel propagator, respectively. After pulse propagation through a temporal fractional FT system constructed by such devices, whose parameters control the rotation angle α, the required number of the WD projections $P_\alpha(v) = \eta\int W(t\eta^2\cos\alpha + v\sin\alpha, v\cos\alpha - t\eta^2\sin\alpha)\mathrm{d}t$ can be registered by a spectrometer. Then the WD (and therefore the two-point correlation function $\Gamma(t_1,t_2)$) is reconstructed applying, for example, the inverse Radon transform.

PST for Quantum Light Characterization

In quantum optics, light is described by the concept of *photons* which are light particles with a particular energy (frequency), polarization, and wave vector (direction of propagation). In the classical textbooks where the second quantization is derived (D'Ariano *et al.*, 2003), photons have a plane-wave spatial structure. However, it is easy to imagine photons with a general transverse structure just by assuming a superposition of plane-wave photons (Calvo *et al.*, 2006). Hence, these quantum entities can be well localized both in space and time.

Many quantum information and quantum computation protocols exploit high-dimensional Hilbert spaces. Photons, which constitute the main carrier of information between nodes of quantum networks, can store high-dimensional quantum bits in their spatial degrees of freedom. These degrees of freedom can be tailored by resorting to the symplectic invariant approach based on lossless linear canonical transformations introduced before in the context of classical optics, Eqs. (7–8). These transformations enable one to manipulate the transverse structure of a single photon prepared in superpositions of paraxial modes. Before describing those transformations acting on photons, it is suitable at this stage to introduce in a nutshell the concept of a photon in the paraxial approximation.

In quantum electrodynamics the photon is interpreted as an excitation of the quantized mode field (D'Ariano *et al.*, 2003). Within this quantization, we can define a set of creation and annihilation operators: those that create a photon from the vacuum with a certain paraxial transverse profile, for example, with a well-defined Laguerre-Gaussian mode or a Hermite-Gaussian mode. The most general (paraxial) single-photon pure state can be described as (Calvo *et al.*, 2006)

$$|\psi\rangle = \sum_{\sigma,n_x,n_y} \int_0^\infty d\omega\, C_{\sigma,n_x,n_y}(\omega)\, \hat{b}^\dagger_{\sigma,n_x,n_y}(\omega)|vac\rangle \tag{29}$$

Here, $\hat{b}^\dagger_{\sigma,n_x,n_y}(\omega)$ denotes the bosonic creation operator of a photon with a Hermite-Gaussian mode (n_x,n_y), a linear polarization σ, and a frequency ω acting on the vacuum state $|vac\rangle$. The commutation relations read as $[\hat{b}_{\sigma,n_x,n_y}(\omega), \hat{b}^\dagger_{\sigma',n_x',n_y'}(\omega')] = \delta_{\sigma\sigma'}\delta_{n_x n_x'}\delta_{n_y n_y'}\delta(\omega-\omega')$. The complex coefficients $C_{\sigma,n_x,n_y}(\omega)$ can be interpreted as the probability amplitudes for finding a photon in the state $\hat{b}^\dagger_{\sigma,n_x,n_y}(\omega)|vac\rangle = |\sigma\rangle\otimes|n_x,n_y\rangle\otimes|\omega\rangle$. Here we will focus on the spatial part of Eq. (29).

In order to introduce the link between the Wigner formalism used in previous sections and the single-photon description, let us consider the connection between the cross-spectral density $w_c(\mathbf{r},\mathbf{r}')$ and the density matrix operator. The density matrix operator for a paraxial wave state $|\psi\rangle$ from Eq. (29) is $\hat{\rho} = |\psi\rangle\langle\psi|$. Here, we deal with coherent waves, but this formalism is more general, and can be extended to partially coherent waves. For coherent waves, the amplitude of the spatial distribution of the photon is defined by $\psi(\mathbf{r}) = \langle\mathbf{r}|\psi\rangle$. In contrast, partially coherent waves do not have a deterministic amplitude, but instead they are described by an ensemble of amplitudes with an associated probability, for example, consider the physical scenario of a phase that is stochastically varying. Within the density operator formalism, partially coherent waves can be described as $\hat{\rho} = \sum_k p_k\hat{\rho}_k = \sum_k p_k|\psi_k\rangle\langle\psi_k|$, where $\hat{\rho}_k$ is a coherent wave with an associated probability p_k. As it is expected for a probability distribution, $\sum_k p_k = 1$ must be satisfied. When $p_k = 1$ for a certain k we return to the coherent wave. Therefore, for both coherent and partially coherent waves, the cross-spectral density is defined as

$$w_c(\mathbf{r},\mathbf{r}') = \langle\mathbf{r}|\,\hat{\rho}\,|\mathbf{r}'\rangle \tag{30}$$

where $\hat{\rho}$ is the density operator of the wave. In the particular case of coherent waves, the cross-spectral density reads as $w_c(\mathbf{r},\mathbf{r}') = \psi(\mathbf{r})\psi^*(\mathbf{r}')$. As well as in the classical case (see Eq. (1)) the Wigner distribution related by the FT to the cross-spectral density is a faithful representation of quantum state.

Identically to previous sections in which light is considered to be a classical electromagnetic wave, we are interested in tomographic techniques to retrieve the WD and therefore the spatial structure of photons. The main difference between the quantum and the classical picture is the concept of measurement. When the square modulus of the amplitude of the photon is measured in a particular position, the quantum state collapses after the measurement and the information of the initial quantum state that we are interested in is lost. Due to this limitation, it is impossible to retrieve the quantum state of a single photon in a single shot. In general, all methods for characterizing quantum states of photons are based on an ensemble of photons, i.e., a set of many photons with the same quantum state. Then the strategy consists in measuring separately all photons and inferring the quantum state from the information obtained from these measurements.

A spatial structure of a photon propagating through a first-order optical system can be modified as explained in Eq. (12). Assuming that we have N identical photons, we send the photons one at the time through the optical system that performs the fractional FT and the image rotator. In each measurement, we measure a *click* in a particular transverse position of our detector. After repeating the experiment N times, we will obtain a distribution of *clicks* in our detector, related to the modulus square of the amplitude (at the position of the detector) of the quantum state. Using the same strategy than in the classical case, we implement the required transformations to retrieve the WD corresponding to a single photon by changing accordingly the optical system. Hence, the discussed PST techniques can be easily extended to the single-photon regime.

Note that the photon can also be described by a superposition of paraxial modes with different frequencies ω (Eq. (29)) and be localized in time. In that case, we can also apply the chronocyclic tomography strategies discussed for short pulses.

So far, the simple case of a single-photon state has been considered . However, there exist more complicated quantum states of light composed of different photons, and the number of photons not necessarily have to be well-defined (D'Ariano *et al.*, 2003).

Then the Eq. (29) given for single-photon state has to be modified in order to account for more than one photon. If there is well-defined number N of photons in the same mode, the quantum state corresponds to the Fock state given by

$$|\psi_F\rangle = \frac{\left[\hat{b}^{\dagger}_{\sigma,n_x,n_y}(\omega)\right]^N}{\sqrt{N!}}|\text{vac}\rangle \tag{31}$$

Other common states are the coherent and squeezed states that do not have a well-defined number of photons. For example, the coherent states can be described via the displacement operator as

$$|\psi_C\rangle = e^{\alpha \hat{b}^{\dagger}_{\sigma,n_x,n_y}(\omega)-\alpha^* \hat{b}_{\sigma,n_x,n_y}(\omega)}|\text{vac}\rangle \tag{32}$$

where α is related to the average photon number and might take a complex value. Note that by expanding the exponential a superposition of Fock states is obtained and then the photon number is not well-defined. In most applications of quantum tomography, there is some prior information about the system. For example, it might be the knowledge that the light is described by a coherent state and the problem is then reduced to the determination of the value of α.

Note that in the previous examples for the Fock and coherent states all photons are described by the same mode, which is the common situation in quantum optics experiments. However, there exist more complicated scenarios in which the quantum states of photons with different modes, including the spatial transverse structure, have to be retrieved. In this case the optical system used for quantum state characterization depends on the mode representation or encoding. For example, a mode representation could be a superposition of photons with different transverse spatial profile, or it could be a superposition of photons propagating in different directions, both being possible mode representations. In the latter, which is well-known in the context of linear optical quantum computing, the quantum state is represented by photons propagating in different directions in an optical circuit constituted by beam splitters, phase shifters, and mirror. In particular, it has been proven that any unitary transform in the Fock space for a specific mode can be implemented by using a set of beam splitters and phase shifters (Knill *et al.*, 2001). These transforms can be expressed by a similar mathematical formalism to the one introduced above for the PST characterization of the spatial structure of classical light.

The PST allows establishing a series of powerful techniques to retrieve both quantum and classical states of lights which are of relevance for a wide range of fields such as optical metrology, microscopy, astronomy, classical and quantum optics communication, to name a few.

Acknowledgements

T.A and J.A.R acknowledge the Spanish Ministerio de Economía y Competitividad for the grant TEC2014-57394-P, A.P. acknowledges the funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 702565.

References

Agarwal, G.S., Simon, R., 2000. Reconstruction of the Wigner transform of a rotationally symmetric two-dimensional beam from the Wigner transform of the beam's one-dimensional sample. Opt. Lett. 25 (18), 1379–1381.
Alieva, T., Cámara, A., Rodrigo, J.A., Calvo, M.L., 2011. Phase-space tomography of optical beams. In Optical and Digital Image Processing. Wiley-VCH.
Bastiaans, M.J., 2008. Applications of the Wigner distribution to partially coherent light beams. In Advances in Information Optics and Photonics. SPIE.
Beck, M., Raymer, M.G., Walmsley, I.A., Wong, V., 1993. Chronocyclic tomography for measuring the amplitude and phase structure of optical pulses. Opt. Lett. 18, 2041–2043.
Calvo, G.F., Picón, A., Bagan, E., 2006. Quantum field theory of photons with orbital angular momentum. Phys. Rev. A 73, 013805.
Cámara, A., 2015. Optical beam characterization via phase-space tomography. Springer.
Cámara, A., Alieva, T., Castro, I., Rodrigo, J.A., 2014. Phase-space tomography for characterization of rotationally symmetric beams. J. Opt. 16 (1), 015705.
Cámara, A., Alieva, T., Rodrigo, J.A., Calvo, M.L., 2011. Phase-space tomography with a programmable Radon–Wigner display. Opt. Lett. 36 (13), 2441–2443.
Cámara, A., Rodrigo, J.A., Alieva, T., 2013. Optical coherenscopy based on phase-space tomography. Opt. Express 21 (11), 13169–13183.
D'Ariano, G.M., Paris, M.G.A., Sacchi, M.F., 2003. Quantum tomography. Adv. Imag. Elect. Phys. 128, 205–308.
Dorrer, C., Kang, I., 2003. Complete temporal characterization on short optical pulses by simplified chronocyclic tomography. Opt. Lett. 28, 1481–1483.
Goodman, J.W., 2000. Statistical Optics, first ed. Wiley-Interscience.
Granieri, S., Furlan, W.D., Saavedra, G., Andres, P., 1997. Radon–Wigner display: A compact optical implementation with a single varifocal lens. Appl. Opt. 36 (32), 8363–8369.
Knill, E., Laflamme, R., Milburn, G.J., 2001. A scheme for efficient quantum computation with linear optics. Nature 409, 46–52.
Mendlovic, D., Dorsch, R.G., Lohmann, A.W., Zalevsky, Z., Ferreira, C., 1996. Optical illustration of a varied fractional Fourier-transform order and the Radon–Wigner display. Appl. Opt. 35 (20), 3925–3929.
Ozaktas, H.M., Zalevsky, Z., Kutay, M.A., 2001. The fractional Fourier transform with applications in optics and signal processing. New York, NY: Wiley.
Raymer, M.G., Beck, M., McAlister, D.F., 1994. Complex wave-field reconstruction using phase-space tomography. Phys. Rev. Lett. 72 (8), 1137–1140.
Rodrigo, J.A., Alieva, T., Calvo, M.L., 2009. Programmable two-dimensional optical fractional Fourier processor. Opt. Express 17 (7), 4976–4983.
Tian, L., Lee, J., Baek Oh, S., Barbastathis, G., 2012. Experimental compressive phase space tomography. Opt. Express 20 (8), 8296–8308.
Tran, C.Q., Peele, A.S., Roberts, A., *et al.*, 2005. X-ray imaging: A generalized approach using phase-space tomography. J. Opt. Soc. Am. A 22 (8), 1691–1700.
Tu, J., Tamura, S., 1997. Wave field determination using tomography of the ambiguity function. Phys. Rev. E 55 (2), 1946–1949.

Coordinate Transformations and the Hough Transform

Filippus S Roux, University of Witwatersrand, Johannesburg, South Africa and National Metrology Institute of South Africa, Pretoria, South Africa

Introduction

An optical coordinate transformation is a linear optical implementation of a class of point transforms that is applied to the intensity distribution of an optical beam. Such a coordinate transformation can be implemented with the aid of a thin optical element having a specially designed phase-only transmission function. The optical element is placed in the input plane where it modulates the input optical field in such a way that the subsequent linear optical system produces the required transformed optical field in the output plane. Another optical element with a phase-only transmission function is sometimes placed in the output plane to remove the residual phase, induced by the transformation, so that the transformed amplitude distribution would remain intact upon subsequent propagation (apart from normal diffraction).

The initial concept for the implementation of optical coordinate transformations was proposed in Bryngdahl (1974). However, for optical elements with continuous phase functions, this procedure can only implement geometrical transformations that obey a continuity condition (Cederquist and Tai, 1984). Subsequently, a number of techniques have been proposed to overcome this restriction. One way is to produce a multifaceted optical element with a discontinuous phase function (Case *et al.*, 1981). Another approach is to separate the geometrical transformation into two cascaded geometrical transformations, each of which satisfies the continuity condition (Stuff and Cederquist, 1990). A third method (Roux, 1994) is to allow the phase function to contain phase singularities, which would produce optical vortices (Nye and Berry, 1974) in the propagating optical field.

Optical coordinate transformations have been used to implement, among others: a log-polar transform for rotation and scale invariant feature extraction (Saito *et al.*, 1983); a rotationally symmetric beam-shaping system (Han *et al.*, 1983); a ring-to-point transform (Cederquist and Tai, 1984); a rotation transformation (Roux, 1993a); and a polar formatting transform to compensate for distortion in synthetic-aperture radar data (Roux, 1995). Recently, a refractive optical implementation of the log-polar transform was proposed for the efficient measurement of the azimuthal index of Laguerre-Gauss modes (Berkhout *et al.*, 2010).

A different kind of point transform is where the input and output coordinates are related by an implicit transformation equation. The Hough transform is an example of such a implicit transformation. It maps lines in the input plane to individual points in the output plane. After considering the optical implementation of coordinate transformation, we'll discuss the optical implementation of the Hough transform.

Coordinate Transformations

A coordinate transformation that can be implemented optically maps the points on a two-dimensional input plane $\mathbf{x}=(x,y)$ onto the points on a two-dimensional output plane $\mathbf{u}=(u,v)$. Being continuous and one-to-one, such a coordinate transformation is defined by a pair of equations, expressing the output coordinates in terms of functions of the input coordinates $u(\mathbf{x})$ and $v(\mathbf{x})$. These relationships are called the transformation equations, which can be represented as a vector field

$$\mathbf{u}(\mathbf{x}) = u(\mathbf{x})\hat{x} + v(\mathbf{x})\hat{y} \tag{1}$$

The optical implementation employs a phase-only transmission function to direct the light from each point in the input plane to the required location on the output plane. Such a phase-only transmission function can be physically realized as a thin refractive optical element or as a diffractive optical element. It can also be implemented with a programmable spatial light modulator.

Bryngdahl Method

First, we'll consider an optical system comprising a 2-f system with a phase-only transmission function located in the front focal plane, which serves as the input plane. Such a system is shown diagrammatically in **Fig. 1**. The output amplitude distribution in the back focal plane (the output plane) of a 2-f system is given by the Fourier transform of the amplitude distribution in the input plane. For the case of the system in **Fig. 1**, the amplitude distribution in the input plane is the input optical field times the phase-only transmission function. To implement the required geometrical transformation, the phase-only transmission function modulates the light at each point in the input plane with an appropriate spatial frequency so that the Fourier transforming action of the lens will place the light from that input point at the required location in the output plane.

Bryngdahl used the stationary-phase approximation to derive a set of differential equations that relates the transformation equations to the gradient of the phase of the transmission function in the input plane of the 2-f system. In vector notation, this

 doi:10.1016/B978-0-12-803581-8.09377-2

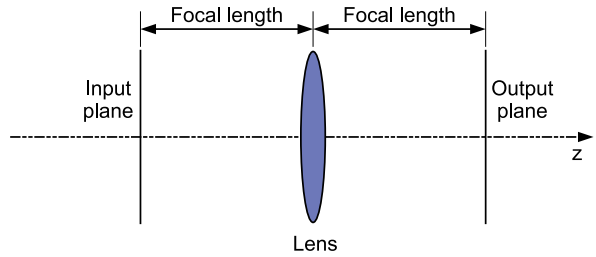


Fig. 1 Optical configuration of a 2-f system.

differential equation is given by

$$\nabla\theta(\mathbf{x}) = \frac{k}{f}\mathbf{u}(\mathbf{x}) \tag{2}$$

where $\theta(\mathbf{x})$ is the phase of the transmission function, k is the wave number and f is the focal length of the lens. Since we'll only work on the two-dimensional input and output planes, all vectors are assumed to be two-dimensional and the gradient operator only operates on the two-dimensional transverse plane $\nabla \equiv \hat{x}\partial_x + \hat{y}\partial_y$.

Note that, Eq. (2) only applies for the case of a 2-f system. One can also use an optical system without a lens, where the input and output planes are separated by a distance L that is large enough to allow the output to be given by Fresnel diffraction. In such a case, the differential equation is

$$\nabla\theta(\mathbf{x}) = \frac{k}{L}[\mathbf{u}(\mathbf{x}) - \mathbf{x}] \tag{3}$$

By solving the differential equation in either Eq. (2) or Eq. (3), depending on the system under consideration, one can obtain the required phase function for the transmission function in the input plane. The transmission function with this phase function can then be implemented with the aid of a refractive optical element or a diffractive optical element or simply a spatial light modulator to realize an optical system that would perform the required coordinate transform.

Continuity Condition

If the required phase function $\theta(\mathbf{x})$ in Eq. (2) is a continuous function, then $\nabla \times \nabla\theta = 0$. It would then imply that $\nabla \times \mathrm{u} = 0$ for both the 2-f system and the lens-less system. In other words, the transformation equations, expressed as a vector field, must be non-rotational. One can also express this requirement as a continuity equation, given by

$$\partial_y u(\mathbf{x}) = \partial_x v(\mathbf{x}) \tag{4}$$

By implication, the two transformation equations are not independent.

There are many useful transformations that do not obey the continuity condition in Eq. (4). A simple example is where the distribution is rotated by an angle α around the optical axis. The transformation equations for such a rotation transformation are

$$\begin{aligned} u(\mathbf{x}) &= \cos(\alpha)x - \sin(\alpha)y \\ v(\mathbf{x}) &= \sin(\alpha)x + \cos(\alpha)y \end{aligned} \tag{5}$$

Here, we have $\partial_y u(\mathbf{x}) = -\sin(\alpha)$, while $\partial_x v(\mathbf{x}) = \sin(\alpha)$. Since these transformation equations do not satisfy the continuity condition, the transformation cannot be implemented with an optical element having a continuous phase function.

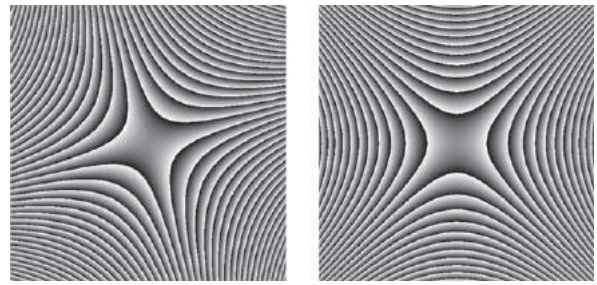
There are a number of different approaches that one can follow to overcome the limitations imposed by the continuity condition in Eq. (4). We'll consider three such approaches: representing the transformation as two separate cascaded transformations, as discussed in Section Cascaded Transformations; implementing the transformation with a multifaceted transmission function, as discussed in Section Multifaceted Transmission Function; or incorporating phase singularities in the phase function of the transmission function, which is discussed in Section Transmission Function With Phase Singularities.

Cascaded Transformations

It was shown (Stuff and Cederquist, 1990) that coordinate transformations, which are bijective[1] geometrical transformations, can be implemented with either one or two transmission functions, depending on whether they obey the continuity condition or not. However, there does not seem to be a general approach to formulate the two separate geometrical transformations that would implement a geometrical transformation not satisfying the continuity condition.

A fairly simple example of such a cascaded implementation is where two mirror transformations are used to implement a rotation transformation. Unlike the rotation transformation, mirror transformations do obey the continuity condition. For a

[1]A bijective transform is one-to-one and covers the whole domain.



(a) (b)

Fig. 2 Two phase functions for the two cascaded mirror transformations, are shown in (a) and (b), respectively. Together they implement a 60 degree rotation. The phase is represented in terms of gray scales, with black and white representing zero and 2π, respectively.

mirror transformation with respect to an arbitrary angle α, the transformation equations are given by

$$u(\mathbf{x}) = \cos(\alpha)x + \sin(\alpha)y v(\mathbf{x}) = \sin(\alpha)x - \cos(\alpha)y \tag{6}$$

Hence, $\partial_y u(\mathbf{x}) = \partial_x v(\mathbf{x}) = \sin(\alpha)$. If this mirror transformation is followed by another mirror transformation with $\alpha = 0$, the result is the rotation transformation, given in Eq. (5).

Using the differential equation for the 2-f system, given in Eq. (2), together with the mirror transformation equations in Eq. (6), we obtain a phase function given by

$$\theta(\mathbf{x}) = \frac{k}{2f}\cos(\alpha)(x^2 - y^2) + \frac{k}{f}\sin(\alpha)xy \tag{7}$$

This phase function defines the phase-only transmission function in the input plane of the first 2-f system. The output plane of the first 2-f system serves as the input plane for another 2-f system having a similar phase-only transmission function in its input plane, but with $\alpha = 0$. The output of the second 2-f system is the required transformed amplitude distribution, which in this case is a rotated version of the amplitude distribution in the input plane of the first 2-f system. Examples of the two phase functions for the two cascaded mirror transformations that together would implement a 60 degree rotation, are shown in **Fig. 2**.

Multifaceted Transmission Function

A general transformation can be implemented optically in a brute force approach, by discretizing the input and output planes. The input plane is divided into a grid of small areas. Each of these areas is then treated as a single point in the input plane, which is to be transformed to a single point in the output plane. A diffraction grating with the appropriate spatial frequency is then placed in each of the small areas. Each of these small areas acts like a small aperture that selects a part of the input field and modulates it with the particular diffraction grating function in that aperture. The effect is that the input field is divided into these input areas and each wavelet then gets diffracted in such a way that the light is placed at the correct location in the output plane.

The phase of the diffraction grating for a particular input area, labelled by the indices m and n for the x- and y-directions, respectively, is given by

$$\theta_{m,n}(\mathbf{x}) = \frac{k}{f}[u(x_m, y_n)x + v(x_m, y_n)y] \tag{8}$$

where (x_m, y_n) are the input coordinates at the center of the small input area and $u(x_m, y_n)$ and $v(x_m, y_n)$ denote the output coordinates that are associated with the input point at (x_m, y_n). The transmission function for each area is then given by

$$t_{m,n}(\mathbf{x}) = \exp[\mathrm{i}\theta_{m,n}(\mathbf{x})] \tag{9}$$

and the total transmission function $t_{\mathrm{tot}}(\mathbf{x})$ is formed as the combination of the transmission functions of all the areas

$$t_{\mathrm{tot}}(\mathbf{x}) = \sum_{m,n} t_{m,n}(\mathbf{x}) \prod\left(\frac{x}{\Delta x} - m, \frac{y}{\Delta y} - n\right) \tag{10}$$

where $\prod(\cdot)$ is a rectangular aperture function defining the regions of the small input areas.

Although this method can implement any coordinate transformation, the output field would have the appearance of a discrete array of spots, rather than a smooth optical field. By making the input regions smaller, the output spots would become bigger and be located closer to each other, so that they may eventually overlap. However, depending on the particular coordinate transformation that is being implemented, the spots may be closer to each other in one part of the output plane than in another. Moreover, the overlapping spots in the output plane may cause undesirable interference fringes.

Another result of the multifaceted transmission function is that the sharp boundaries between adjacent input regions cause additional diffraction effects that may be undesirable. One can reduce this effect by allowing the gratings in the different small regions to merge into those of adjacent regions by providing each with a smooth envelope function that extends into the adjacent region and merges with the envelope function of that region. In that case, one would replace the rectangular aperture function

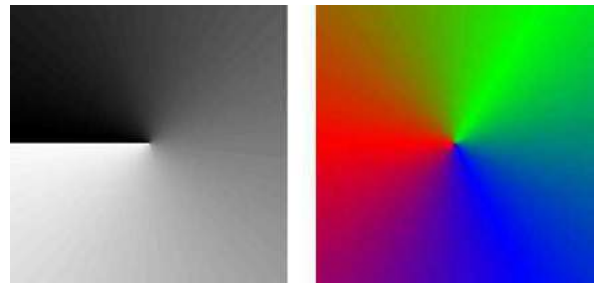


Fig. 3 The phase function of a single isotropic phase singularity. In (a) the phase value is represented in terms of a gray scale, with black and white representing zero and 2π, respectively. Note that the discontinuity along the black-to-white transition is only an artifact of the representation, since 0 and 2π represents the same phase. In (b) a cyclic range of rainbow colors is used to represent the phase values, showing the fact that there is no discontinuity in the phase around the defect in the center.

$\prod(\cdot)$, defined in Eq. (10), by a suitably smooth envelope function. The result would be a varying diffraction grating over the entire input plane. However, forked grating lines would appear at the boundaries where the grating periods don't match up exactly. These forked grating lines represent phase singularities in the phase of the transmission function, which is the topic of the next section.

Phase Singularities

To discuss the method based on the inclusion of phase singularities in the transmission function, we first need to consider phase singularities in general. Their role in those coordinate transformations that do not obey the continuity condition and the process to compute the required phase function are explained in Section Transmission Function With Phase Singularities.

Phase Functions

One can represent phase as a point on a unit circle around the origin on the complex plane. Each point on the circle represents a unique phase value and all possible phase values are represented by points on the circle. There is no 'largest value' or 'smallest value' for phase. A phase value of 2π represents the same phase as zero. The value of the phase always increases (decreases) for anticlockwise (clockwise) motion round the circle. The unit circle on the complex plane is an example of a compact domain, where the values remain finite even though one never reaches a boundary. This also implies that the circle has a nontrivial topological structure; it is not simply connected.

The topological properties of the configuration space of phase give phase functions special characteristics that make them different from ordinary functions.[2] A phase function $\theta(\mathbf{x})$ is a continuous mapping from the two-dimensional (x,y)-plane to the circle of phase values. The continuity of the mapping requires that a connected line of points on the (x,y)-plane maps to a connected line of points on the phase circle. The nontrivial topological structure of the circle now allows for the appearance of topological defects. Consider, for instance, all the possible ways that one can map a closed contour from the (x,y)-plane onto or into the circle of phase values. This could give connected phase values that are all located on one side of the circle, so that if the contour shrinks to a point on the (x,y)-plane, the phase values would also shrink to a point on the phase circle. However, if the mapping of the closed contour produces a wrapping around the phase circle, the phase values on one side of the phase circle would have to 'jump' to the other side when the closed contour shrinks to a point on the (x,y)-plane. The 'jumping' process represents a tearing of the topological space, which is not allowed for any process that conserves the topology of the space. This indicates the presence of a topological defect inside the closed contour on the (x,y)-plane.

For a phase function, such topological defects are called phase singularities, because at such a point the phase is undefined and around such a point the phase takes on all the possible phase values. It is therefore classified as an essential singularity. These phase singularities have topological charges, which may be positive or negative depending on the handedness of the wrapping. The topological charges can also be larger than 1 (or smaller than -1) depending on the number of times that the mapping wraps around the phase circle.

The phase function of a single isotropic phase singularity is shown in **Fig. 3**. Two different ways are used to represent the phase. Henceforth, phase functions are only presented in terms of gray scales.

Mathematically, the presence of a phase singularity inside a close contour can be determined with the aid of an index integral, which integrates the gradient of the phase along the closed contour. The result is an integer multiple of 2π:

$$\oint_C \nabla\theta(\mathbf{x}) \cdot \mathrm{d}\hat{s} = 2\pi n \tag{11}$$

where n is an integer representing the net enclosed topological charge.

[2]In this context, an 'ordinary' function is one with a non-compact range that has no non-trivial homotopy groups.

As an example, consider a phase function that is given by the azimuthal angle $\phi(\mathbf{x})$, expressed as a function of Cartesian coordinates. In this case the index integral can be expressed as

$$\int_0^{2\pi} \partial_\phi \phi \, \mathrm{d}\phi = 2\pi \tag{12}$$

As expected, one finds a phase singularity at the origin with a topological charge of $+1$. This result confirms that phase functions are not ordinary functions.

Derivatives of a Phase Function

Applying the Stokes theorem to Eq. (11), one finds

$$\int_A [\nabla \times \nabla\theta(\mathbf{x})] \cdot \hat{z} \, \mathrm{d}^2x = 2\pi n \tag{13}$$

where A is the area enclosed by the contour. The expression in Eq. (13) implies that $\nabla \times \nabla\theta(\mathbf{x}) \neq 0$. In fact, the curl of the gradient of a phase function produces a Dirac delta function at the location of each phase singularity, multiplied by its topological charge (Roux, 2006)

$$[\nabla \times \nabla\theta(\mathbf{x})] \cdot \hat{z} = 2\pi \sum_p (v)_p \delta(\mathbf{x} - \mathbf{x}_p) \tag{14}$$

where v_p represents the topological charge of each phase singularity and $\mathbf{x}_p$ is the location of the phase singularity. For example, in the case of a single phase singularity, as represented by a shifted function of the azimuthal angle, one gets

$$\nabla \times \nabla\phi(\mathbf{x} - \mathbf{x}_0) = 2\pi\delta(\mathbf{x} - \mathbf{x}_0)\hat{z} \tag{15}$$

Here the azimuthal angle represents the generic phase function for a single isotropic phase singularity, as shown in **Fig. 3**.

In more general terms, one can regard the right-hand side of Eq. (14) as a topological charge distribution $T(\mathbf{x})$. So then

$$\nabla \times \nabla\theta(\mathbf{x}) = 2\pi T(\mathbf{x})\hat{z} \tag{16}$$

By integrating both sides of Eq. (16) over the area A, one recovers Eq. (13), with the understanding that the net topological charge is given by the integral of the topological charge distribution

$$n = \int_A T(\mathbf{x}) \, \mathrm{d}^2x \tag{17}$$

General Phase Function

We see now that a phase function in general consists of a continuous part and a singular part. The continuous part is an ordinary function. The singular part can be represented as a sum of shifted versions of the azimuthal angle, multiplied by their topological charges. Hence, a general phase function can be expressed by

$$\theta(\mathbf{x}) = \theta_{\mathrm{cont}}(\mathbf{x}) + \sum_p (v)_p \phi(\mathbf{x} - \mathbf{x}_p) \tag{18}$$

where the first term on the right-hand side represents the continuous part and the second term represents the singular part.

Transmission Function With Phase Singularities

Required Topological Charge Distribution

With the new understanding of phase functions, we can take another look at the continuity condition. First we note that the requirement for a continuous phase function is the reason why this restriction exists, because it implies that $\nabla \times \nabla\theta_{\mathrm{cont}}(\mathbf{x}) = 0$. Therefore, if we have a set of transformation equations for which $\nabla \times \mathbf{u} \neq 0$, then it simply means that we need to include phase singularities in the phase function.

Let's apply once again the curl on both sides of Eq. (2), this time using Eq. (16). Then we find that

$$\nabla \times \nabla\theta(\mathbf{x}) = 2\pi T(\mathbf{x})\hat{z} = \frac{k}{f} \nabla \times \mathbf{u}(\mathbf{x}) \tag{19}$$

The rotational part of the vector field $\mathbf{u}(\mathbf{x})$ precisely gives the topological charge distribution that is required to implement the transform

$$T(\mathbf{x}) = \frac{1}{\lambda f} [\nabla \times \mathbf{u}(\mathbf{x})] \cdot \hat{z} \tag{20}$$

Note, however, that the topological charge distribution thus obtained is a continuous function and not a summation of discrete topological charges associated with distinct phase singularities. Since phase singularities are by their very nature always discrete

points, the required topological charge distribution can at best be approximated by an arrangement of phase singularities over the (x,y)-plane.

The example of the rotation transformation, defined in Eq. (5), requires a constant topological charge distribution given by

$$T(\mathbf{x}) = \frac{1}{\lambda f}\left[\partial_x v(\mathbf{x}) - \partial_y u(\mathbf{x})\right] = \frac{2\sin(\alpha)}{\lambda f} \tag{21}$$

It should be noted that such a constant topological charge density cannot be implemented over an arbitrary large region. Due to limitations on the density of phase singularities (Roux, 2003), the maximum radius over which one can maintain a constant topological charge density, as given in Eq. (21), is

$$R_{\max} = \frac{f}{\sin(\alpha)} \tag{22}$$

For larger rotation angles α, one would need to increase the focal length.

Continuous Part of the Phase Function

While the required topological charge distribution is readily computed with the aid of Eq. (20), the continuous part of the phase function requires more attention. One needs to obtain a set of modified transformation equations that excludes the contribution of the phase singularities. Such a modified set of transformation equations would then obey the continuity condition (representing an non-rotational vector field) that could be used to compute a continuous phase function.

The total phase function, given in Eq. (18), consist of a continuous part and the singular part. Computing the gradient of the total phase function, one obtains

$$\nabla\theta(\mathbf{x}) = \nabla\theta_{\text{cont}}(\mathbf{x}) + \sum_p (v)_p \nabla\phi(\mathbf{x} - \mathbf{x}_p) \tag{23}$$

The left-hand side is given by the transformation equations according to Eq. (2). The required transformation equations for the continuous part of the phase function is thus given by

$$\nabla\theta_{\text{cont}}(\mathbf{x}) = \frac{k}{f}u(\mathbf{x}) - \sum_p (v)_p \nabla\phi(\mathbf{x} - \mathbf{x}_p) \tag{24}$$

It now remains to determine the last term (the singular term) on the right-hand side of Eq. (24). One can reconstruct the singular term from the topological charge distribution that is required to implement the transform. Although we don't know the locations and topological charges of the individual phase singularities, as required in the last term of Eq. (24), we do know the required topological charge distribution. So one can convert the summation into an integral, where we replace the individual topological charges with the known topological charge distribution

$$\sum_p (v)_p \nabla\phi(\mathbf{x} - \mathbf{x}_p) \to \int T(\mathbf{x}')\nabla\phi(\mathbf{x} - \mathbf{x}')\mathrm{d}^2x' \tag{25}$$

The result is a convolution integral, which convolves the topological charge distribution with the gradient of a single phase singularity function. The resulting differential equation for the continuous part of the phase function now reads

$$\nabla\theta_{\text{cont}}(\mathbf{x}) = \frac{k}{f}\mathbf{u}(\mathbf{x}) - 2\pi\mathbf{F}_s \tag{26}$$

where

$$2\pi\mathbf{F}_s = \int T(\mathbf{x}')\nabla\phi(\mathbf{x} - \mathbf{x}')\mathrm{d}^2x' \tag{27}$$

is the singular phase gradient.

As a test of consistency, we apply a curl operation on both sides of Eq. (26). The result gives

$$\nabla \times \nabla\theta_{\text{cont}}(\mathbf{x}) = \frac{k}{f}\nabla \times \mathbf{u}(\mathbf{x}) - 2\pi\nabla \times \mathbf{F}_s \tag{28}$$

where

$$2\pi\nabla \times \mathbf{F}_s = \int T(\mathbf{x}')\nabla \times \nabla\phi(\mathbf{x} - \mathbf{x}')\mathrm{d}^2x' \tag{29}$$

Since $\theta_{\text{cont}}(\mathbf{x})$ is an ordinary function, the left-hand size of Eq. (28) becomes zero. One can use Eq. (15) to evaluate the integral in Eq. (29). In the end, Eq. (28) becomes

$$0 = \frac{k}{f}\nabla \times \mathbf{u}(\mathbf{x}) - 2\pi T(\mathbf{x}) \tag{30}$$

which is in agreement with Eq. (19). Hence, we have confidence that Eq. (26) represents the required differential equation for the continuous part of the required phase function.

Gradient of the Singular Phase Function

The evaluation of the convolution integral in Eq. (27) requires the expression for the gradient of the generic function for a single phase singularity. One can express the azimuthal angle in terms of the transverse Cartesian coordinate by[3]

$$\phi(\mathbf{x}) = \frac{1}{\mathrm{i}2}\ln\left(\frac{x+\mathrm{i}y}{x-\mathrm{i}y}\right) \tag{31}$$

The gradient of this phase function is

$$\nabla\phi(\mathbf{x}) = \frac{x\hat{y}-y\hat{x}}{x^2+y^2} = \frac{\hat{z}\times\mathbf{x}}{|\mathbf{x}|^2} \tag{32}$$

The convolution integral in Eq. (27) thus becomes

$$2\pi\mathbf{F}_s = \int T(\mathbf{x}')\frac{\hat{z}\times(\mathbf{x}-\mathbf{x}')}{|\mathbf{x}-\mathbf{x}'|^2}\mathrm{d}^2x' \tag{33}$$

or, by redefining the integration variable,

$$2\pi\mathbf{F}_s = \int T(\mathbf{x}-\mathbf{x}')\frac{\hat{z}\times\mathbf{x}'}{|\mathbf{x}'|^2}\mathrm{d}^2x' \tag{34}$$

The expressions in Eqs. (33) and (34) are real-valued. One can simplify the expressions by combining the x- and y-components as the real and imaginary parts of the expression. The results are

$$2\pi F_s = \int \frac{\mathrm{i}T(\mathbf{x}-\mathbf{x}')}{x'-\mathrm{i}y'}\mathrm{d}^2x' \tag{35}$$

or

$$2\pi F_s = \int \frac{\mathrm{i}T(\mathbf{x})}{(x-x')-\mathrm{i}(y-y')}\mathrm{d}^2x' \tag{36}$$

whichever is more convenient to use. The real part of the resulting integral is the x-component and the imaginary part is the y-component.

Application: Rotation Transform

For a constant topological charge density, such as the one required for the rotation transformation, the integral for the singular part of the phase gradient can be readily evaluated with the aid of contour integration. For this purpose we use the expression in Eq. (36) and obtain the result

$$2\pi F_s = \int \frac{\mathrm{i}T_0}{(x-x')-\mathrm{i}(y-y')}\mathrm{d}^2x' = \mathrm{i}\pi(x+\mathrm{i}y)T_0 \tag{37}$$

where T_0 represents a constant topological charge density.

For the rotation transformation, the singular part of the phase gradient is therefore given by

$$2\pi F_s = \frac{k}{f}(\mathrm{i}x-y)\sin(\alpha) \tag{38}$$

which means that

$$2\pi\mathbf{F}_s = \frac{k}{f}(x\hat{y}-y\hat{x})\sin(\alpha) \tag{39}$$

The differential equation for the continuous phase function is obtain by substituting the transformation equations and the singular phase gradient into Eq. (26). For the rotation transformation, the differential equation becomes

$$\nabla\theta_{\mathrm{cont}}(\mathbf{x}) = \frac{k}{f}\cos(\alpha)\mathbf{x} \tag{40}$$

This differential equation can now be solved to give

$$\theta_{\mathrm{cont}}(\mathbf{x}) = \frac{k\cos(\alpha)}{2f}(x^2+y^2) \tag{41}$$

The continuous phase function is a lens function with a focal length that depends on the rotation angle. Together with the required topological charge distribution given in Eq. (21), one can now construct the complete phase function that is required to implement the rotation transformation, using a 2-f system. A part of the phase function is shown in **Fig. 4**. It contains the phase singularities placed on an equilateral triangular grid.

[3]This definition in terms of the ln-function is preferred over that with the arctan-function, because the latter only produces phase values between $-\pi/2$ and $\pi/2$.

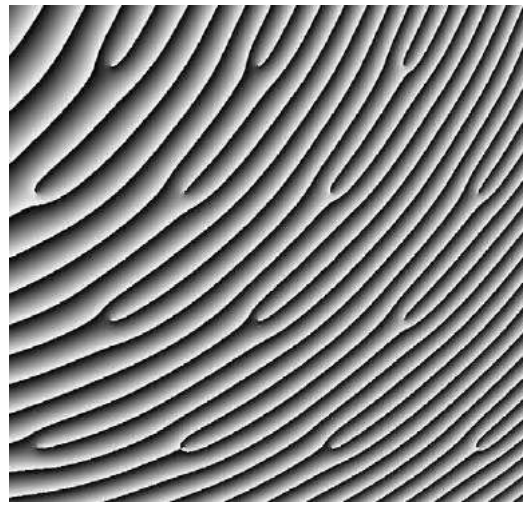


Fig. 4 Part of the phase function with phase singularities of the transmission function for a 45 degree rotation.

Hough Transform

Implicit Transformation

The coordinate transformations, which are discussed above, are defined in terms of a pair of explicit equations that give the output coordinates in terms of the input coordinates and are therefore called explicit transformations. Such transformations are one-to-one, meaning that each output point is associated with only one unique input point.

There are also other transformations that produce one-to-many mappings between input and output points; one input point goes to a set of different output points. Such transformations are defined in terms of implicit equations – a single expression that contains the input coordinates and the output coordinates implicitly – and are therefore referred to as implicit transformations.

Definition of the Hough Transform

An example of an implicit transformation is the conventional Hough transform (Duda and Hart, 1972), which maps lines in the input plane to points in the output plane. The implicit equation for the Hough transform is given by

$$u = x\cos(Kv) + y\sin(Kv) \tag{42}$$

where K is a constant parameter that converts the output coordinate v into an angle. One can define $K = 2\pi/L$, where L is the width of the output region in the v-direction.

For a fixed output point (u,v), the equation in Eq. (42) describes a line in the input plane. Alternatively, for a fixed input point (x,y) the equation in Eq. (42) describes a sinusoidal curve in the output plane.

Optical Implementation

The optical implementation of the Hough transform can be done with the aid of a single transmission function (Roux, 1993b). Such a transmission function needs to be constructed as a superposition of several transmission functions, each implementing an explicit transformations. Here, we'll describe the implementation with the aid of a multifaceted transmission function, as discussed in Section Multifaceted Transmission Function.

The transmission function of a general optical transform (explicit or implicit), implemented with a multifaceted transmission function, can be expressed as

$$t(\mathbf{x}) = \sum_{m,n,p,q} \alpha_{mnpq} A(x - m\Delta x, y - n\Delta y) \times \Phi^*(p\Delta u, q\Delta v; x, y) \tag{43}$$

where α_{mnpq} denotes the binary coefficients (either equal to 1 or 0) that indicate whether or not there is a link between particular input and output points; $A(\mathbf{x})$ is a suitable envelope function; and $\Phi^*(\mathbf{u};\mathbf{x})$ is the complex conjugate of the appropriate kernel function (Fourier or Fresnel), depending on the type of optical system in which the transform is implemented. The grid spacings in the input and output planes along the two orthogonal directions are denoted by Δx, Δy, Δu and Δv, respectively. The envelope function could be $\prod(\cdot)$, which is used in Eq. (10), or it could be a smoothing window function to produce a more continuous transmission function.

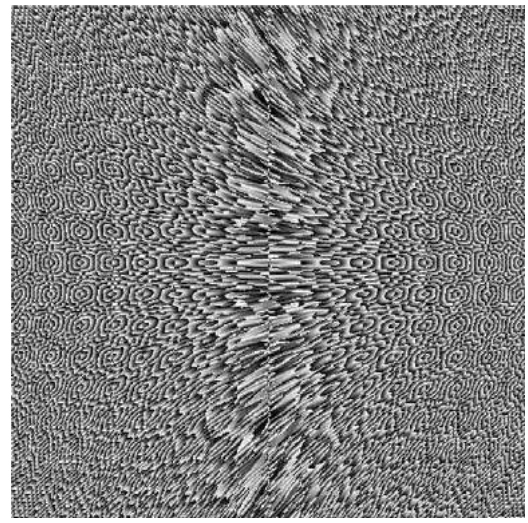


Fig. 5 The phase function of the transmission function for the Hough transform.

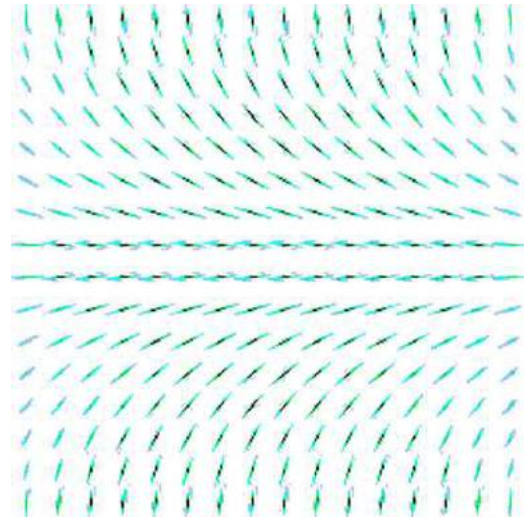


Fig. 6 Output diffraction pattern produced by a fully illuminated transmission function for the Hough transform.

The phase function of the transmission function for an optically implemented Hough transform is shown in **Fig. 5**. It was computed with Eq. (43) for an implementation in a 2-f system. The Hanning window function was used as envelope function. Using the resulting transmission function, we calculated the diffraction pattern in the output plane when the transmission function is fully illuminated. It is shown in **Fig. 6**. The diffraction pattern contains an array of discrete output points as produced due to the discrete nature of the implementation. Each output point is elongated in a direction perpendicular to the orientation of the line in the input plane that is mapped to that output point.

General Comments

There are a number of general considerations that one needs to keep in mind when designing optical systems to implement coordinate transformations or more general implicit transformations, such as the Hough transform. These are:

- The stationary phase approximation that is used to derive the differential equations for the phase of the transmission function ignores the amplitude distribution of the incident optical field. As a result, the effect of the phase of the incident optical field is not taken into account in the design of the optical element. If the incident optical field contains a severely tilted phase front at the location of a particular input point, the location of the associated output point would be noticably shifted away from the

location defined by the transformation. For instance, a helical phase that is unwrapped by the log-polar transform would give a distorted output line in the case of high azimuthal index in the incident optical field.

- The size of an input area, restricted by an aperture, gives rise to a finite area in the output plane with a minimum size. For a smaller (larger) input area, the size of the output area becomes larger (smaller). This relationship is well-known for the 2-f system, where the input and output distributions are related by a Fourier transform,

$$d_{\text{out}} = \frac{\lambda f}{d_{\text{in}}} \tag{44}$$

where d_{in} and d_{out} represent the sizes of the input and output areas, respectively. A similar relationship applies for other systems such as the lens-less system.

The relationship between the sizes of these areas in the input and output can be used to compute a quantity called the space-bandwidth-product (SBP), which is determined as follows. The size of the total usable output area defines a smallest feature size in the input plane via the Fourier relationship in Eq. (44). Anything smaller would scatter light beyond the boundary of the usable output area. At the same time, the input plane also have a finite size for its total usable area that produces a small spot in the output, which thus defines the resolution in the output plane. The SBP is now given by the ratio of the total input area divided by the smallest feature size and it is also equal to the ratio of the total output area divided by the smallest resolution cell in the output plane.

A coordinate transformation, defines a mapping from separate points on the input plane to distinguishable points on the output plane. In practice, these points would be represented by finite size areas both on the input and the output planes. What do the sizes of these areas need to be and how many can there be? One cannot use the number given by the SBP of the system as an indication of the number of points that are mapped between input and output, because if we use the smallest input feature size in the input, then the output points would be as large as the entire output area. Moreover, if one wants the output points to be the size of the resolution cells, the input points would need to cover the entire input area. Hence, the optimal number of points that can be used (in the case of the rotation transform, for instance) would be the square root of the SBP. This gives as many distinguishable output points as separated input points. For this reason, a coordinate transform is not an effective way to implement beam shaping.

- Depending on the type of implementation method that is used, one may have to accept a limited diffraction efficiency. The result is that there could be undiffracted light that also propagates together with the diffracted light, after passing through the devise that implements the transmission function. To separate the diffracted light from the undiffracted light, one can add an overall phase grating (linear phase tilt) to the phase of the transmission function. The magnitude of the phase tilt needs to be larger than the largest spatial frequency in the original phase function to ensure that the desired first diffraction order would be completely separated from the undiffracted zeroth order. The desired first diffraction order then appears next to the undiffracted zeroth order in the output plane and can be separated from it by spatial filtering.
- In the case of an implicit transformation where one input point is mapped to several output points, such as with the Hough transform, the transmission function consists of the superposition of several transmission functions for the individual output points. The superposition causes interference, which modulates the amplitude (magnitude) of the resulting transmission function. As a result the final transmission function for the total implicit transform is not a pure phase function (it is not 'phase-only'), such as the transmission functions for the individual output points.

This gives rise to a number of consequences. To implement the complex amplitude of the total transmission function in a phase-only device, such as a spatial light modulator or diffractive optical elements, special encoding techniques need to be used. Alternatively, one can use older technologies that incorporate absorption. The modulated amplitude of the total transmission function implies that some of the incident optical power would be lost due to absorption or scattering, depending on how the complex amplitude of the transmission function is implemented. The result is a poor power efficiency for the implementation. One can try to use only the phase of the total transmission function to implement a phase-only optical transform, but in general the resulting transformation would then suffer from low fidelity.

Conclusions

Point transforms that are implemented in linear optical systems can be divided into explicit transformations, such as coordinate transformations, and implicit transformations, such as the Hough transform. Explicit transformations include those that only require a continuous phase function in the transmission function and therefore satisfy the continuity condition and those that also require phase singularities in the phase function of the transmission function and therefore do not satisfy the continuity condition.

Those explicit transformations that do satisfy the continuity condition can be implemented by a direct solution of the Bryngdahl differential equations. Explicit transformations that do not satisfy the continuity condition require additional attention. These explicit transformations either need to be separated into two cascaded explicit transformations, each satisfying the continuity condition or they need to be implemented as a multifaceted transmission function or they need the inclusion of phase singularities in the phase function of the transmission function.

The Hough transform is an implicit transformation that maps lines in the input plane to separate points in the output plane. It can be implemented as a superposition of several discrete transmission functions that map particular input points to particular output points.

The optical implementation of geometrical transformations provide a fast effective way to process optical information. However, care must be taken in the design of these systems to ensure that the implementation provide an accurate result.

References

Berkhout, G.C.G., Lavery, M.P.J., Courtial, J., Beijersbergen, M.W., Padgett, M.J., 2010. Efficient sorting of orbital angular momentum states of light. Phys. Rev. Lett. 105, 153601.
Bryngdahl, O., 1974. Geometrical transformations in optics. J. Opt. Soc. Am. 64, 1092–1099.
Case, S.K., Haugen, P.R., Lokberg, O.J., 1981. Multifacet holographic optical elements for wave front transformations. Appl. Opt. 20, 2670–2675.
Cederquist, J.N., Tai, A.M., 1984. Computer-generated holograms for geometrical transformations. Appl. Opt. 23, 3099–3104.
Duda, R.O., Hart, P.E., 1972. Use of the Hough transformation to detect lines and curves in pictures. Commun. ACM 15, 11–15.
Han, C.-Y., Ishii, Y., Murata, K., 1983. Reshaping collimated laser beams with gaussian profile to uniform profiles. Appl. Opt. 22, 3644–3647.
Nye, J.F., Berry, M.V., 1974. Dislocations in wave trains. Proc. R. Soc. Lond. A 336, 165–190.
Roux, F.S., 1993a. Diffractive optical implementation of rotation transform performed by using phase singularities. Appl. Opt. 32, 3715–3719.
Roux, F.S., 1993b. Implementation of general point transforms with diffractive optics. Appl. Opt. 32, 4972–4978.
Roux, F.S., 1994. Branch-point diffractive optics. J. Opt. Soc. Am. A 11, 2236–2243.
Roux, F.S., 1995. Single-element diffractive optical system for realtime processing of synthetic aperture radar data. Appl. Opt. 34, 5045–5052.
Roux, F.S., 2003. Optical vortex density limitation. Opt. Commun. 223, 31–37.
Roux, F.S., 2006. Fluid dynamical enstrophy and the number of optical vortices in a paraxial beam. Opt. Commun. 268, 15–22.
Saito, Y., Komatshu, S., Ohzu, H., 1983. Scale and rotation invariant real time optical correlator using computer generated hologram. Opt. Commun. 47, 8–11.
Stuff, M.A., Cederquist, J.N., 1990. Coordinate transformations realizable with multiple holographic optical elements. J. Opt. Soc. Am. A 7, 977–981.

Single-Pixel Imaging Using the Hadamard Transform

Fernando Soldevila, Pere Clemente, Enrique Tajahuerce, and Jesús Lancis, Jaume I University, Castelló, Spain

Introduction

Sometimes great revolutions occur in science. All the different branches of physics had short periods of time where new ideas flooded their communities and sparkled new technologies. Perhaps, Optics has been the most affected by those kinds of revolutions. We can clearly see at least three periods of time that were crucial to the development of light sciences as we know them today: the improvement in optical lenses fabrication, the generation of laser based light sources, and the development of electronic light sensors. The first one occurred in the 19th century, thanks to the works of Zeiss, Schott, and Abbe. They provided not only technical developments in the fabrication of lenses, but also the theoretical framework of modern optics. In the 1950s, the first laser sources appeared, and proved to exceptional light sources that rapidly replaced white light sources in almost every experiment. With their use, novel techniques emerged, such as lithography, range measuring, and laser cutting. Furthermore, new devices were developed around them, such as optical disk drives. Lastly, at the end of the 1960s, charge-coupled devices (CCD) were invented. Now, optical fields could be digitally stored, and soon digital imaging techniques provided information that would have never been possible by working with conventional photographic films.

Even though this last revolution happened half a century ago, it slowly started a trend that continues today: the symbiosis of Electronics, Computer Science and Optics. Now, new optical techniques not only need to build optical systems considering lenses and light sources, but also the electronics present in their detectors and the computer algorithms they use to recover useful information. Moreover, new electronical devices have been built to control light, such as spatial light modulators, optical filters, etc. This blending of disciplines has proved to be an exceptional tool in a plethora of industries, life sciences, telecommunications, and astronomy.

Although CCD and complementary metal-oxide-semiconductor (CMOS) sensors have stablished as the way-to-go technology in digital imaging, several limitations remain unsolved. Obtaining images in low light level scenarios, in exotic ranges of the electromagnetic spectrum and at extremely high speeds still entails very expensive devices or with great limitations regarding spatial resolution and response time. Due to this, scientists have never stopped their search for new imaging modalities using different detectors. One of the modern solutions to these problems consists of using very specialized sensors without spatial resolution. One can think of those detectors as a digital camera with only one pixel (thus they are usually called single-pixel detectors) (Duarte *et al.*, 2008). Several questions should arise now: why using just one pixel? and, how do I obtain an image (with hundreds or thousands of pixels) with just a single-pixel detector?

The first question has a simple explanation. Building just one sensor with fast response time, sensibility or spectral response is usually easy, and it is always easier than building an array of thousands or millions of sensors with the same characteristics. Then, these cameras using single-pixel detectors present enhanced capabilities in challenging scenarios for traditional digital cameras. Now, the second question needs a little longer explanation, which we will see in the following section.

Single-Pixel Imaging Basics

In order to understand the operation of a single-pixel camera, we can start with a conventional digital camera, as shown in **Fig. 1**. In a conventional setup, the scene to be acquired is imaged with the aid of an optical system (think of the photographic objective of a digital camera, for example) onto the sensor. Due to image formation laws, each region of the scene is imaged onto a different

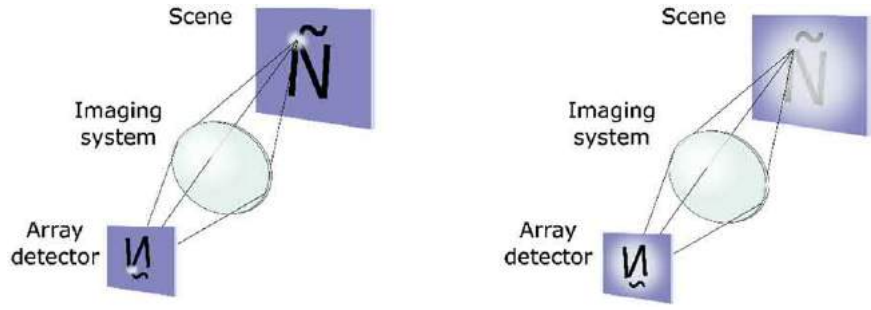


Fig. 1 Conventional digital camera setup. Left: using an imaging system, different regions of the scene are imaged onto different regions of the sensor, that measures the light irradiance. Using this information, a digital image is acquired. Right: as an array detector is used, this process is made in parallel, and each pixel of the detector obtains the information of a different region of the scene at the same time.

region of the sensor, so we have a 1:1 map between regions of the scene and regions of the sensor. Now, each pixel of the sensor measures the light irradiance coming from the scene. In a conventional CCD sensor, every pixel measures at the same time, so the image is acquired in a single shot. However, there are other ways of obtaining the same information.

One can illuminate only a small region of the scene (either by using a galvanometric mirror and a laser beam, or with a spatial light modulator), and measure the light irradiance coming from that region with a detector. This detector does not need to have millions of pixels any more, as light is collected with the optics onto a small spot, and the only relevant quantity to measure is not where the light goes to, but the amount of light coming from the scene. This is usually known in the imaging community as raster scanning, and it's the common operation mode for techniques such as confocal microscopy. The process is depicted in **Fig. 2**. Now, as the detector does not need spatial resolution, instead of a CCD sensor, different sensors can be used. For example, if one needs to work in low light level scenarios, photomultiplier tubes of photon counting detectors can be used. Even beam spectrometers and polarimeters have been used, providing polarimetric and multispectral images. This measurement process can be depicted by using simple mathematical operations. We consider a bidimensional object, $O(x,y)$. In order to obtain an image, we make the superposition of the object and a function that belongs to an orthogonal basis. In the raster scanning example, those functions are Dirac deltas, each one centered at a different spatial position. Then, each measurement, I_p, can be expressed as

$$I_p = O(x,y) \cdot \delta_p(x - x_p, y - y_p) \tag{1}$$

For each delta function, the single-pixel detector measures the amount of light coming from the object. After all the measurements are done, we can recover the object with the equation

$$O = \sum_p I_p \cdot \delta_p \tag{2}$$

It must be stated that, in order to recover an image with a number N of pixels, N delta functions need to be generated and sequentially measured with the detector, whereas in a conventional camera all the measurements are done in one single shot.

After the introduction of the raster scanning approach, it is easy to understand that there is no need to limit oneself to the use of delta functions. The same mathematical problem can be solved in any base of orthogonal functions, and this knowledge can be very useful in practical scenarios. As shown in **Fig. 3**, now instead of scanning the scene, different light patterns (functions of the basis) are sequentially generated onto the object, performing what is usually known as a basis scan. The main advantage of using different functions is that, instead of just illuminating a small region of the scene, several parts of the object are lighted at the same time. This increases the signal level at the detector, which ultimately improves the signal-to-noise ratio of the measurements. This is usually known as the Fellgett or multiplex advantage. Here, a plethora of functions have been discussed in the references, such as Fourier functions, DCT, and wavelets. Even non-orthogonal functions, such as random functions, can be used to obtain images, as

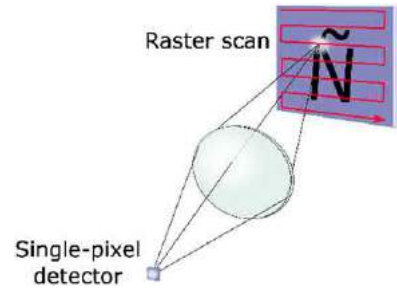


Fig. 2 Raster scanning imaging. The scene under study is scanned with a light beam, and the reflected irradiance is collected by an optical system and measured with a single-pixel detector. By scanning the full scene, the image can be recovered.

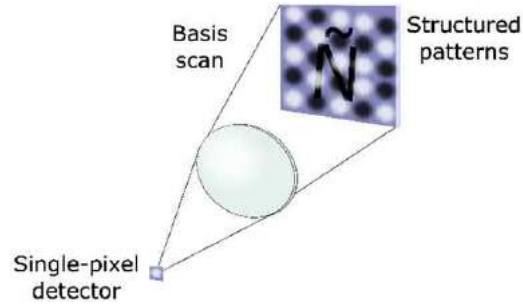


Fig. 3 Basis scanning imaging. The scene under study is lighted with a complete set of functions from an orthogonal basis, and the reflected irradiance is collected by an optical system and measured with a single-pixel detector. After all the functions have been projected, the set of measurements can be used to recover the image of the scene.

in the case of computational ghost imaging techniques. Each one of the basis presents its benefits and characteristic experimental implementations given its mathematical properties, and usually the choice between one on another can be made depending on the specifications of the device that one wants to build. Here, we are going to discuss the use of Walsh-Hadamard functions, its properties, and the requirements needed in order to build a single-pixel camera using them.

Walsh-Hadamard Single-Pixel Imaging

Walsh functions can be derived from Hadamard matrices, which are square matrices with dimensions of power of 2, entries either $+1$ or -1, and the property that the product of two different rows or columns is zero (Pratt *et al.*, 1969) (see **Fig. 4**). Then, each row (column) can be understood as a square binary function which is orthogonal to all the other rows (columns). Hadamard matrices can be calculated for arbitrary big dimensions, and are easily stored and transmitted. Furthermore, due to being binary matrices, its technical implementation using spatial light modulators tends to be straightforward. Given those reasons, they are an excellent candidate to operate in a single-pixel architecture. Using vector form, now the measurement process can be described by

$$\boldsymbol{y} = H \cdot \boldsymbol{x} \tag{3}$$

where the object under study is denoted by a vector, $\boldsymbol{x}$, containing the N pixels of the scene arranged in column form. The measurement process is represented by a Hadamard matrix, H, with dimension $N \times N$. Each row of this matrix contains a Walsh function. The measurements are arranged in the column vector $\boldsymbol{y}$. In each element of the measurement vector the superposition of a Walsh function and the whole object is stored. This process of projecting the Walsh functions and measuring the resulting irradiance can be understood as an optical way of calculating the Walsh-Hadamard transform of the object (see **Fig. 5**). This is similar to the measurement of the Fourier Transform of an object by using a digital camera in the focal plane of a lens. Finally, after the measurements are done, one can recover the object by inverting the problem in Eq. (3). This can also be understood as performing the inverse Walsh-Hadamard transform of the measurement vector, $\boldsymbol{y}$.

However, there can be a problem regarding the implementation of Walsh functions in spatial light modulators. As their entries consist of either $+1$ or -1, amplitude modulators cannot directly implement Walsh functions (there is no available -1 state). A usual workaround entails the use of two different functions to codify a function. Instead of working with the true functions, a Hadamard matrix can be expressed as $H = H^{+} - H^{-}$. The first matrix is built by substituting the -1 entries by 0, and the second one by replacing the $+1$ entries by zero, and the -1 entries by 1. Then, the measurement process can be expressed by

$$\boldsymbol{y} = H \cdot \boldsymbol{x} = H^{+} \cdot \boldsymbol{x} - H^{-} \cdot \boldsymbol{x} \tag{4}$$

which matrices can be easily codified on amplitude spatial light modulators. It has to be noted that, if one uses phase spatial light modulators, this problem does not occur. However, due to the extremely fast operation modes of amplitude spatial light modulators and the sequential nature of single-pixel cameras, those are the ones usually used.

Once the fundamentals of Walsh-Hadamard imaging have been stablished, here we show an example of a single-pixel camera using a digital micromirror device (DMD) as an amplitude spatial light modulator. A scheme of the camera setup is shown in **Fig. 6**. Single-pixel imaging can be divided into two main stages: illumination (or codification) and detection. In the illumination

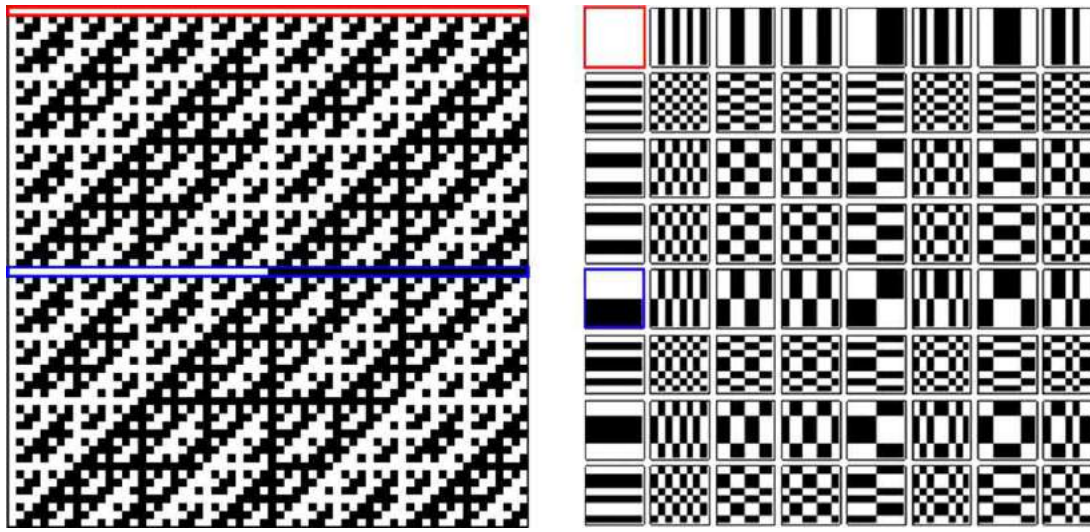


Fig. 4 Hadamard matrix and patterns example. Left: Hadamard matrix of dimension 64. White regions represent $+1$ entries and black regions represent -1 entries. Right: Hadamard patterns derived from the Hadamard matrix. In order to build each one of the 64 patterns, one row of the matrix is selected and arranged into a 8×8 pattern. Two example patterns, highlighted red and blue, are marked with their correspondent rows of the Hadamard matrix.

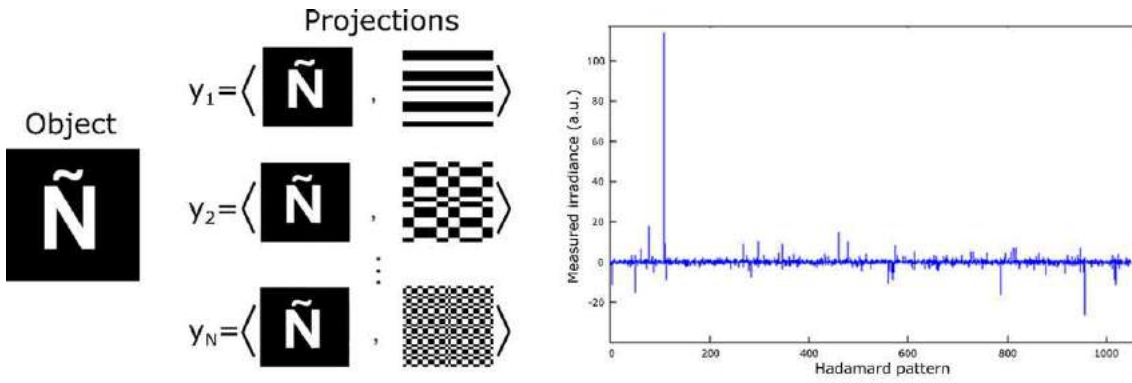


Fig. 5 Single-pixel Hadamard measuring fundamentals. The object is measured by making the projections onto the basis of Hadamard patterns. Those projections consist on overlapping both the object and each Hadamard pattern, and then measuring the total irradiance coming to the detector. By doing this, one gets the decomposition of the object in the Hadamard basis (this graph is the 1D Walsh-Hadamard transform of the object expressed in vector form). Once this decomposition is measured, the recovery can be obtained by performing the inverse Walsh-Hadamard transform of the coefficient vector.

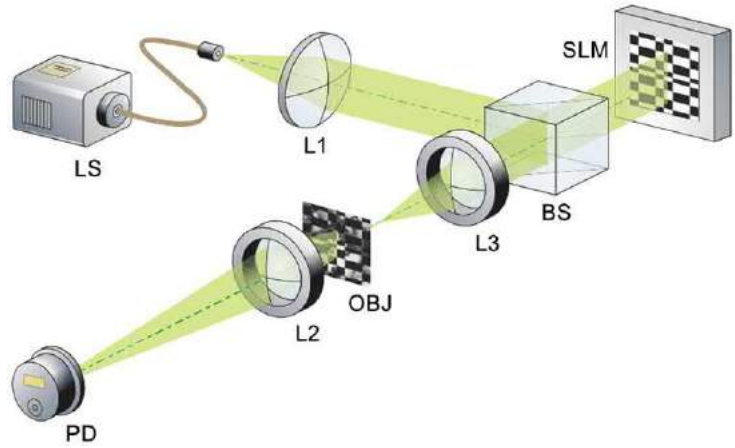


Fig. 6 Single-pixel camera setup. Light coming from a light source (LS) is structured with a spatial light modulator (SLM). An image of the SLM is projected onto the object (OBJ) with several lenses and a beam-splitter (BS). After the superposition of both the Walsh-Hadamard functions and the object, the resulting irradiance is measured with a single-pixel detector (a photodiode in this case, PD).

stage, a light source illuminates a spatial light modulator, where the Walsh functions are generated. In this case, each one of the rows of the Hadamard matrix is rearranged in a bidimensional mask, usually called Hadamard pattern. Those patterns are projected into the scene that one wants to measure with the aid of an optical projection system. Then, the object and the Hadamard patterns overlap in one plane. In the detection stage, the resulting light distribution is collected with the aid of another optical system. A single-pixel detector measures this irradiance value, which is the projection of the object into one of the functions of the Walsh basis. Then, the spatial light modulator codifies another Walsh function and the detector measures again. After the full set of Hadamard patterns have been projected, the image of the object can be recovered. Two examples of images obtained by using this approach are shown in Fig. 7.

It must be stated that the spatial codification of information is realized when both the Hadamard patterns and the object overlap. Here we have shown one way of doing this, by projecting the patterns generated by the spatial light modulator onto the object. However, this projection can be done in different ways. For example, one can image the object onto the spatial light modulator with an optical system, and then make the overlapping in the spatial light modulator plane. After that, light is collected and measured with a single-pixel detector in the same way as in the first approach. This flexibility can be exploited in different scenarios, given that sometimes it is easy to work with an active illumination (sending light patterns to an object) and sometimes it is better to work in a passive configuration (imaging the scene onto the spatial light modulator).

Walsh-Hadamard Single-Pixel Applications

Given the benefits mentioned before, examples of single-pixel imaging techniques using Walsh-Hadamard functions can be extensively found in the references. They can be divided in four main groups. First, given the ease to build single-pixel detectors

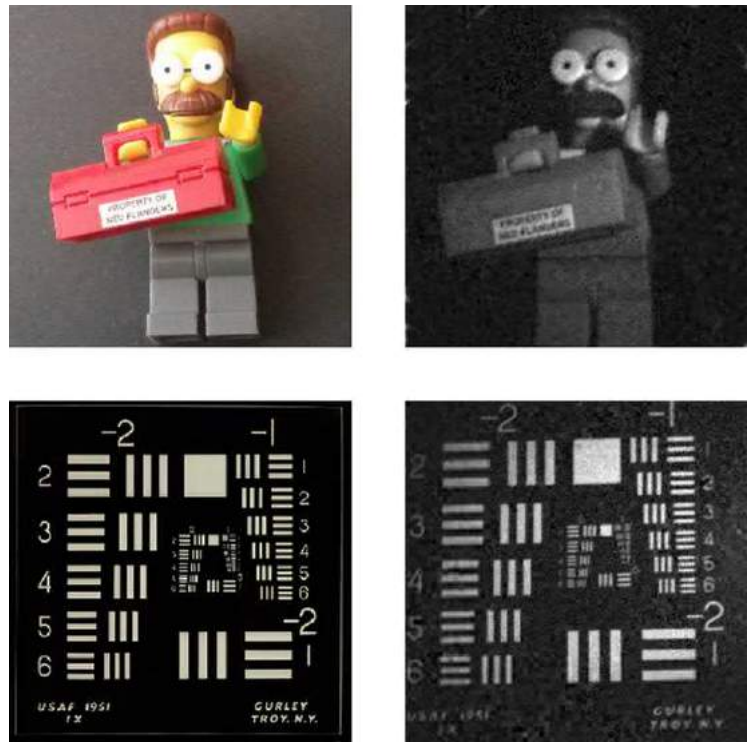


Fig. 7 Single-pixel imaging using Hadamard patterns. In the top row, we show a 256×256 pixels LEGO® Ned Flanders picture (left) and its single-pixel reconstruction. In the bottom row, we show a 512×512 pixels USAF1951 test image (left) and its single-pixel reconstruction (right). Images extracted from Soldevila, F., Salvador-Balaguer, E., Clemente, P., Tajahuerce, E., Lancis, J., 2015. High-resolution adaptive imaging with a single photodiode. Scientific Reports (Nature Publishing Group) 5, 7 (Article No. 14300). Licensed under CC BY 4.0.

that work in exotic ranges of the spectral domain, multiple single-pixel cameras have been built in the IR and THz spectral regions (Radwell *et al.*, 2014; Watts *et al.*, 2014). Second, using simple detectors in combination with other optical and mechanical elements has allowed to perform multidimensional imaging. Here we can name multi and hyper spectral imaging (Soldevila *et al.*, 2013; Li *et al.*, 2017; Studer *et al.*, 2012), polarimetric imaging (Durán *et al.*, 2012; Welsh *et al.*, 2015), phase imaging (Clemente *et al.*, 2013), and 3D imaging (Zhang *et al.*, 2016), among others. Third, given the simplicity of the detection scheme, dedicated sensors can also be used to work in challenging scenarios, such as low-light level regimes, by using photomultiplier tubes or photon counting detectors. This has enabled single-pixel cameras to provide images, for example, in presence of highly scattering media, where distinguishing between non-scattered photons and diffuse light is of paramount interest (Tajahuerce *et al.*, 2014; Durán *et al.*, 2015). Lastly, single-pixel ideas have also been applied to well stablished optical techniques, improving their performance in demanding situations. Here we can name several applications, such as improving the SNR in two-photon microscopy (Ducros *et al.*, 2013) and photoacoustic imaging (Wang *et al.*, 2012), and enhancing resolution in traditional microscopy setups (Rodríguez *et al.*, 2014).

Challenges, Future Work

The main drawback of single-pixel cameras is the need to perform a sequential codification of information: whereas a traditional digital camera images the scene in a single shot, single-pixel cameras need to project a set of patterns onto the scene to recover an image. Furthermore, the higher the spatial resolution one wants to obtain, the higher the number of projections that need to be measured. Even though fast spatial light modulators are used, such as digital micromirror devices, that have repetition rates above 20 kHz, images higher than 64×64 pixels are rarely seen in applications where real-time operation is needed. For several applications, this resolution is high enough, but there are situations where high spatial resolution is needed, such as medical imaging.

In order to tackle this problem, multiple approaches have been proposed. Using Compressive Sensing techniques enable single-pixel cameras to reduce the number of required pattern projections (Duarte *et al.*, 2008; Candès, 2006; Romberg, 2008). Even though they usually entail high post-processing times, novel approaches have also been used to provide real-time video (Sankaranarayanan *et al.*, 2012). Other techniques, based on adaptive algorithms, have also been proposed as a way to reduce the measurement time with no need of complex post-processing stages (Soldevila *et al.*, 2015; Phillips *et al.*, 2017). However, none of those solutions have been able to provide results comparable to traditional digital cameras in terms of spatial resolution and frame

rate as of today. Further development in both computational algorithms and technical improvements in the fabrication of spatial light modulators (i.e., the development of faster SLMs) will be key to solve the present limitations of single-pixel cameras.

References

Candès, E., 2006. Compressive sampling. In: Proceedings of the International Congress of Mathematicians Madrid, August 22–30, 2006. Zurich, Switzerland: European Mathematical Society Publishing House, pp. 1433–1452.
Clemente, P., Durán, V., Tajahuerce, E., *et al.*, 2013. Compressive holography with a single-pixel detector. Opt. Lett. 38 (14), 2524–2527.
Duarte, M.F., Davenport, M.A., Takhar, D., *et al.*, 2008. Single-pixel imaging via compressive sampling. IEEE Signal Process. Mag. 25 (2), 83–91.
Ducros, M., Goulam Houssen, Y., Bradley, J., de Sars, V., Charpak, S., 2013. Encoded multisite two-photon microscopy. Proc. Natl. Acad. Sci. USA 110 (32), 13138–13143.
Durán, V., Clemente, P., Fernández-Alonso, M., Tajahuerce, E., Lancis, J., 2012. Single-pixel polarimetric imaging. Opt. Lett. 37 (5), 824–826.
Durán, V., Soldevila, F., Irles, E., *et al.*, 2015. Compressive imaging in scattering media. Opt. Express 23 (11), 14424.
Li, Z., Suo, J., Hu, X., *et al.*, 2017. Efficient single-pixel multispectral imaging via non-mechanical spatio-spectral modulation. Sci. Rep. 7, 41435.
Phillips, D.B., Sun, M.-J., Taylor, J.M., *et al.*, 2017. Adaptive foveated single-pixel imaging with dynamic supersampling. Sci. Adv. 3 (4), e1601782.
Pratt, W., Kane, J., Andrews, H., 1969. Hadamard transform image coding. Proc. IEEE 57 (1), 58–68.
Radwell, N., Mitchell, K.J., GIbson, G., *et al.*, 2014. Single-pixel infrared and visible microscope. Optica 1 (5), 285–289.
Rodríguez, A.D., Clemente, P., Irles, E., Tajahuerce, E., Lancis, J., 2014. Resolution analysis in computational imaging with patterned illumination and bucket detection. Opt. Lett. 39 (13), 3888–3891.
Romberg, Justin, 2008. Imaging via compressive sampling. IEEE Signal Process. Mag. 25 (2), 14–20.
Sankaranarayanan, A.C., Studer, C., Baraniuk, R.G., 2012. CS-MUVI: Video compressive sensing for spatial-multiplexing cameras. In: 2012 IEEE International Conference on Computational Photography (ICCP), pp. 1–10.
Soldevila, F., Irles, E., Durán, V., *et al.*, 2013. Single-pixel polarimetric imaging spectrometer by compressive sensing. Appl. Phys. B Lasers Opt. 113 (4), 551–558.
Soldevila, F., Salvador-Balaguer, E., Clemente, P., Tajahuerce, E., Lancis, J., 2015. High-resolution adaptive imaging with a single photodiode. Sci. Rep. 5.14300.
Studer, V., Bobin, J., Chahid, M., *et al.*, 2012. Compressive fluorescence microscopy for biological and hyperspectral imaging. Proc. Natl. Acad. Sci. USA 109 (26), E1679–E1687.
Tajahuerce, E., Durán, V., Clemente, P., *et al.*, 2014. Image transmission through dynamic scattering media by single-pixel photodetection. Opt. Express 22 (14), 16945–16955.
Wang, Y., Maslov, K., Wang, L.V., 2012. Spectrally encoded photoacoustic microscopy using a digital mirror device. J. Biomed. Opt. 17 (6), 66020.
Watts, C.M., Shrekenhamer, D., Montoya, J., *et al.*, 2014. Terahertz compressive imaging with metamaterial spatial light modulators. Nat. Photonics 8, 605–609.
Welsh, S.S., Edgar, M.P., Bowman, R., Sun, B., Padgett, M.J., 2015. Near video-rate linear Stokes imaging with single-pixel detectors. J. Opt. 17 (2), 25705.
Zhang, Y., Edgar, M.P., Sun, B., *et al.*, 2016. 3D single-pixel video. J. Opt. 18 (3), 035203.

Linear Canonical Transforms

Kurt B Wolf, National Autonomous University of Mexico, Cuernavaca, Morelos, Mexico

Introduction

Canonical transformations preserve the *structure* of Hamiltonian systems, be they *classical* as geometric optics or classical mechanics, or *wave* as monochromatic optics or quantum mechanics. When these transformations are *linear*, they can be represented by matrices, a most useful technique to keep account of their composition, which facilitates finding the end product of optical elements in a setup.

One can broadly define a D-dimensional system to be *Hamiltonian* when it contains variables or observables of *position*, $\mathbf{Q}=\{Q_i\}_{i=1}^{D}$, and an equal number of *momentum* variables or observables $\mathbf{P}=\{P_i\}_{i=1}^{D}$, which constitute its *phase space*; and a Hamiltonian function $H(\mathbf{Q},\mathbf{P})$ that determines its *canonical evolution* along an optical axis z, or time t as in mechanics. If the transformation is to be *linear*, one must assume that phase space is the full real $2D$-dimensional space $\mathbb{R}^{2D}$. This proviso allows only *paraxial* optical models to be addressed directly, as well as non-relativistic classical and quantum mechanics, since their treatment is mathematically similar to the corresponding classical and wave models in paraxial optics.

Conservation of Hamiltonian Structure

An important property of Hamiltonian systems is their *structure*; this boils down to a set of relations that must be conserved between the phase space coordinates under any transformation: *Poisson brackets* in the classical models and *commutators* in the wave models, namely

$$\{q_i,p_j\}=\delta_{i,j},\quad \{q_i,q_j\}=0,\quad \{p_i,p_j\}=0,\quad q_i\equiv Q_i\in\mathbb{R} \tag{1}$$

$$[\hat{Q}_i,\hat{P}_j]=\mathrm{i}\tilde{\lambda}\delta_{i,j},\quad [\hat{Q}_i,\hat{Q}_j]=0,\quad [\hat{P}_i,\hat{P}_j]=0,\quad \hat{Q}_i\equiv Q_i \text{ on } \mathcal{L}^2(\mathbb{R}^D) \tag{2}$$

These operations are bilinear: $\{[c_1A_1+c_2A_2,B]\}=c_1\{[A_1,B]\}+c_2\{[A_2,B]\}$, skew-symmetric: $\{[A,B]\}=-\{[B,A]\}$, satisfy the Jacobi identity and the Leibniz rule. They generate thus the Lie algebra with derivation referred by Heisenberg and Weyl. The wave models (of reduced wavelength $\tilde{\lambda}=\lambda/2\pi$, or $\hbar=h/2\pi$ in quantum mechanics) further require the Hilbert space $\mathcal{L}^2(\mathbb{R}^D)$ of Lebesgue square-integrable 'wave'-functions, where the operators $\hat{Q}_i$ and $\hat{P}_j$ are essentially self-adjoint.

The 4×4 matrix of Lie brackets (1)–(2) with the $2D$ column vector $\mathbf{W}:=\binom{\mathbf{Q}}{\mathbf{P}}$ and its transpose $\mathbf{W}^{\mathrm{T}}$, written as

$$\{[\mathbf{W}^{\mathrm{T}},\mathbf{W}]\}:=\left\{\left[(\mathbf{Q},\mathbf{P}),\begin{pmatrix}\mathbf{Q}\\ \mathbf{P}\end{pmatrix}\right]\right\}:=\begin{pmatrix}\{[Q_i,Q_j]\} & \{[P_i,Q_j]\}\\ \{[Q_i,P_j]\} & \{[P_i,P_j]\}\end{pmatrix}=\begin{pmatrix}0 & -\delta_{i,j}\\ \delta_{i,j} & 0\end{pmatrix} \tag{3}$$

requires that, under linear transformation by a $2D\times2D$ matrix $\mathbf{M}=\begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix}$ with four $D\times D$ blocks, which maps $\mathbf{W}\mapsto\mathbf{M}\mathbf{W}$, should continue to satisfy

$$\{[\mathbf{W}^{\mathrm{T}},\mathbf{W}]\}=\{[(\mathbf{M}\mathbf{W})^{\mathrm{T}},\mathbf{M}\mathbf{W}]\}=\mathbf{M}\boldsymbol{\Omega}\mathbf{M}^{\mathrm{T}}=\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}=:\boldsymbol{\Omega} \tag{4}$$

In 2×2 block form,

$$\begin{pmatrix}\mathbf{0} & -\mathbf{1}\\ \mathbf{1} & \mathbf{0}\end{pmatrix}=\begin{pmatrix}-\mathbf{A}\mathbf{B}^{\mathrm{T}}+\mathbf{B}\mathbf{A}^{\mathrm{T}}-\mathbf{A}\mathbf{D}^{\mathrm{T}}+\mathbf{B}\mathbf{C}^{\mathrm{T}}\\ -\mathbf{C}\mathbf{B}^{\mathrm{T}}+\mathbf{D}\mathbf{A}^{\mathrm{T}}-\mathbf{C}\mathbf{D}^{\mathrm{T}}+\mathbf{D}\mathbf{C}^{\mathrm{T}}\end{pmatrix} \tag{5}$$

this entails the $D(2D-1)$ *symplectic conditions*:

$$\begin{gathered}\mathbf{A}\mathbf{B}^{\mathrm{T}},\ \mathbf{C}\mathbf{D}^{\mathrm{T}} \text{ are symmetric (and also } \mathbf{A}^{\mathrm{T}}\mathbf{C},\ \mathbf{B}^{\mathrm{T}}\mathbf{D}),\\ \mathbf{A}\mathbf{D}^{\mathrm{T}}-\mathbf{B}\mathbf{C}^{\mathrm{T}}=\mathbf{1}\end{gathered} \tag{6}$$

$$\Rightarrow\quad \mathbf{M}^{-1}=\boldsymbol{\Omega}\mathbf{M}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}=\begin{pmatrix}\mathbf{D}^{\mathrm{T}} & -\mathbf{B}^{\mathrm{T}}\\ -\mathbf{C}^{\mathrm{T}} & \mathbf{A}^{\mathrm{T}}\end{pmatrix} \tag{7}$$

Matrices that satisfy (6) are called *symplectic* and provide a minimal matrix representation of linear canonical transformations. The product of two such matrices is also symplectic, and so is the unit $\mathbf{1}$ and the inverse (7). They are thus a $D(2D+1)$-dimensional manifold that forms a *group*, denoted $\mathrm{SP}(2D,\mathbb{R})$, where $\boldsymbol{\Omega}$ is called the symplectic metric matrix.

Geometric Canonical Transforms

In plane paraxial optics (with line $D=1$ screens), canonical transformations are represented by the 3-parameter manifold of 2×2 matrices of unit determinant $\mathsf{Sp}(2,\mathbb{R})$ acting on $\binom{q}{p}$. In the common setups with plane $D=2$ screens, the group of canonical transformations is the 10-parameter symplectic group $\mathsf{Sp}(4,\mathbb{R})$ represented by 4×4 matrices acting on $(\mathbf{q},\,\mathbf{p})^{\mathsf{T}}=(q_x,\,q_y,\,p_x,\,p_y)^{\mathsf{T}}$.

The correspondence between quadratic functions of the phase space coordinates, matrices and optical elements is established when the former are expressed in terms of the *Poisson operators* of differentiable functions $f(\mathbf{q},\,\mathbf{p})$, given by

$$\{f,\circ\}:=\sum_{i=1}^{D}\left(\frac{\partial f(\mathbf{q},\mathbf{p})}{\partial q_i}\frac{\partial}{\partial p_i}-\frac{\partial f(\mathbf{q},\mathbf{p})}{\partial p_i}\frac{\partial}{\partial q_i}\right) \tag{8}$$

acting on phase space functions with the Poisson *bracket*: $\{f,\circ\}g(\mathbf{q},\,\mathbf{p}):=\{f,g\}(\mathbf{q},\,\mathbf{p})$. When exponentiated with Taylor series in powers $\{f,\circ\}^n:=\{\{f,\circ\}^{n-1},\circ\}$, one finds in the $D=2$ group $\mathsf{Sp}(4,\mathbb{R})$

3 anamorphic lens parameters:

$$\exp\left\{-\sum_{i\le j}c_{i,j}q_iq_j,\circ\right\}\begin{pmatrix}\mathbf{q}\\ \mathbf{p}\end{pmatrix}=\begin{pmatrix}\mathbf{1} & \mathbf{0}\\ -\mathrm{c} & \mathbf{1}\end{pmatrix}\begin{pmatrix}\mathbf{q}\\ \mathbf{p}\end{pmatrix},\quad \mathrm{c}=\begin{pmatrix}c_{x,x} & c_{x,y}\\ c_{x,y} & c_{y,y}\end{pmatrix} \tag{9}$$

3 anisotropic free space displacements:

$$\exp\left\{-\sum_{i\le j}b_{i,j}p_ip_j,\circ\right\}\begin{pmatrix}\mathbf{q}\\ \mathbf{p}\end{pmatrix}=\begin{pmatrix}\mathbf{1} & \mathbf{b}\\ \mathbf{0} & \mathbf{1}\end{pmatrix}\begin{pmatrix}\mathbf{q}\\ \mathbf{p}\end{pmatrix},\quad \mathrm{b}=\begin{pmatrix}b_{x,x} & b_{x,y}\\ b_{x,y} & b_{y,y}\end{pmatrix} \tag{10}$$

4 anisotropic magnifiers:

$$\exp\left\{-\sum_{i,j}a_{i,j}q_ip_j,\circ\right\}\begin{pmatrix}\mathbf{q}\\ \mathbf{p}\end{pmatrix}=\begin{pmatrix}e^{\mathrm{a}} & \mathbf{0}\\ \mathbf{0} & e^{-\mathrm{a}^{\mathsf{T}}}\end{pmatrix}\begin{pmatrix}\mathbf{q}\\ \mathbf{p}\end{pmatrix},\quad \mathrm{a}=\begin{pmatrix}a_{x,x} & a_{x,y}\\ a_{y,x} & a_{y,y}\end{pmatrix} \tag{11}$$

When applying matrices on phase space vectors to represent canonical transformations in an optical setup, one should remember that (if light goes from left to right) the leftmost optical element will correspond with the rightmost matrix factor (to act first), and successively in inverse order.

The group of $D=2$ canonical transformations $\mathsf{Sp}(4,\mathbb{R})$ contains 4 *rotations* of phase space which close into the *Fourier group*, which is isomorphic to the unitary group of 2×2 (complex) matrices $\mathsf{U}(2)_{\mathrm{F}}$. This contains x- and y- fractional Fourier transforms (FrFT), gyrations and screen rotations, which are also *generated* by exponentials of Poisson operators of quadratic functions of phase space given below. Writing for brevity $c:=\cos\theta$ and $s:=\sin\theta$, and identifying their generator functions, they are

$$\begin{array}{cccc}\text{isotropic FrF T} & \text{anisotropic FrF T} & \text{gyrations} & \text{rotations}\\ \begin{pmatrix}c & 0 & -s & 0\\ 0 & c & 0 & -s\\ s & 0 & c & 0\\ 0 & -s & 0 & c\end{pmatrix}; & \begin{pmatrix}c & 0 & -s & 0\\ 0 & c & 0 & s\\ s & 0 & c & 0\\ 0 & -s & 0 & c\end{pmatrix}, & \begin{pmatrix}c & 0 & 0 & -s\\ 0 & c & -s & 0\\ 0 & s & c & 0\\ s & 0 & 0 & c\end{pmatrix}, & \begin{pmatrix}c & -s & 0 & 0\\ s & c & 0 & 0\\ 0 & 0 & c & -s\\ 0 & 0 & s & c\end{pmatrix}\\ \frac{1}{2}\left(p_x^2+p_y^2+q_x^2+q_y^2\right) & \frac{1}{2}\left(p_x^2-p_y^2+q_x^2-q_y^2\right) & p_xp_y+q_xq_y & q_xp_y-q_yp_x\end{array} \tag{12}$$

These matrices are both symplectic and *orthogonal* (called ortho-symplectic); the isotropic FrFT commutes with the other three and is generated by the isotropic harmonic oscillator Hamiltonian. In general D dimensions, $\mathsf{Sp}(2D,\mathbb{R})$ contains a $\mathsf{U}(D)_{\mathrm{F}}$ subgroup of such phase space rotations represented by $2D\times 2D$ ortho-symplectic matrices.

Canonical Integral Transforms

The transformation between input and output wavefields due to the passage through an ideal paraxial optical setup are canonical transforms. Their kernel was found by Stuart A. Collins in optics and, independently, by Marcos Moshinsky and Christiane Quesne as a fundamental question in quantum mechanics, in 1970 and 1971.

Wave optics and quantum mechanics provide a Hilbert space of square-integrable wavefunctions $f(\mathbf{q})$ of $\mathbf{q}\in\mathbb{R}^D$ on which the phase space operators, satisfying (2) act as $\hat{Q}_if(\mathbf{q})=q_if(\mathbf{q})$ and $\hat{P}_if(\mathbf{q})=-\mathrm{i}\partial f(\mathbf{q})/\partial q_i$ (for $\tilde{\lambda}=1$ or $\hbar=1$). Canonical transformations $\mathcal{C}_{\mathrm{M}}$ characterized by generic symplectic matrices $\mathbf{M}$ will act on these operators through,

$$\mathcal{C}_{\mathrm{M}}\begin{pmatrix}\hat{\mathbf{Q}}\\ \hat{\mathbf{P}}\end{pmatrix}\mathcal{C}_{\mathrm{M}}^{-1}=\mathbf{M}^{-1}\begin{pmatrix}\hat{\mathbf{Q}}\\ \hat{\mathbf{P}}\end{pmatrix}=\begin{pmatrix}\mathbf{D}^{\mathsf{T}}\,\hat{\mathbf{Q}} & - & \mathbf{B}^{\mathsf{T}}\,\hat{\mathbf{P}}\\ -\mathbf{C}^{\mathsf{T}}\,\hat{\mathbf{Q}} & + & \mathbf{A}^{\mathsf{T}}\,\mathbf{P}\end{pmatrix} \tag{13}$$

The *inverse* matrix (7) is required, as can be easily ascertained verifying that $\mathcal{C}_{\mathrm{M}_1}\mathcal{C}_{\mathrm{M}_2}=\sigma\mathcal{C}_{\mathrm{M}_1\mathrm{M}_2}$, but for a sign σ that is not detectable in (13).

Regarding the action of $\mathcal{C}_{\mathrm{M}}$ on wavefunctions, since the Fresnel transform of free flight is an integral transform, and so is the fractional Fourier transform, canonical transformations $\mathcal{C}_{\mathrm{M}}$ will be *integral transform* operators that act as

$$f_{\mathrm{M}}(\mathbf{q}) \equiv (\mathcal{C}_{\mathrm{M}} f)(\mathbf{q}) := \int_{\mathbb{R}^D} \mathrm{d}^D\mathbf{q}' C_{\mathrm{M}}(\mathbf{q},\mathbf{q}') f(\mathbf{q}') \tag{14}$$

where the integral kernel $C_{\mathrm{M}}(\mathbf{q},\mathbf{q}')$ is found applying $\mathcal{C}_{\mathrm{M}}$ to $\hat{Q}_i f$ and $\hat{P}_i f$,

$$\mathcal{C}_{\mathrm{M}}(\hat{Q}_i f) = \left(\mathcal{C}_{\mathrm{M}}\hat{Q}_i\mathcal{C}_{\mathrm{M}}^{-1}\right)\mathcal{C}_{\mathrm{M}} f = \sum_{j=1}^{D}(D_{j,i}\hat{Q}_j - B_{j,i}\hat{P}_j) f_{\mathrm{M}} \tag{15}$$

$$\mathcal{C}_{\mathrm{M}}(\hat{P}_i f) = \left(\mathcal{C}_{\mathrm{M}}\hat{P}_i\mathcal{C}_{\mathrm{M}}^{-1}\right)\mathcal{C}_{\mathrm{M}} f = \sum_{j=1}^{D}\left(-\mathcal{C}_{j,i}\hat{Q}_j + A_{j,i}\hat{P}_j\right) f_{\mathrm{M}} \tag{16}$$

Replacing the integral (14) on the left, $\hat{Q}_i$ and $\hat{P}_i$ will act on the $\mathbf{q}'$ argument of f, while on the right they act on the exterior argument $\mathbf{q}$ of the kernel $C_{\mathrm{M}}(\mathbf{q},\mathbf{q}')$. The operator $-\mathrm{i}\partial/\partial q_i'$ in the left can be integrated by parts and applied on the $\mathbf{q}'$ argument of the kernel; this leads to $2D$ simultaneous differential equations:

$$q_i' C_{\mathrm{M}}(\mathbf{q},\mathbf{q}') = \sum_{j=1}^{D}\left(D_{j,i}q_i + \mathrm{i}B_{j,i}\frac{\partial}{\partial q_j}\right) C_{\mathrm{M}}(\mathbf{q},\mathbf{q}') \tag{17}$$

$$-\mathrm{i}\frac{\partial}{\partial q_i'} C_{\mathrm{M}}(\mathbf{q},\mathbf{q}') = \sum_{j=1}^{D}\left(C_{j,i}q_i + \mathrm{i}A_{j,i}\frac{\partial}{\partial q_j}\right) C_{\mathrm{M}}(\mathbf{q},\mathbf{q}') \tag{18}$$

Multiplying these vector equations by $\mathbf{B}^{-1}$ and proposing a quadratic exponential solution, with appropriate normalization one finds

$$C_{\mathrm{M}}(\mathbf{q},\mathbf{q}') = K_{\mathrm{M}} \exp\,\mathrm{i}\left(\frac{1}{2}\mathbf{q}^{\mathrm{T}}\mathbf{B}^{-1}\mathbf{D}\mathbf{q} - \mathbf{q}^{\mathrm{T}}\mathbf{B}^{-1}\mathbf{q}' + \frac{1}{2}\mathbf{q}'^{\mathrm{T}}\mathbf{A}\mathbf{B}^{-1}\mathbf{q}'\right) \tag{19}$$

$$K_{\mathrm{M}} := \frac{1}{\sqrt{(2\pi\mathrm{i})^D \det\,\mathbf{B}}} \equiv \frac{e^{-\mathrm{i}\pi D/4}\exp\,\mathrm{i}\left(-\frac{1}{2}\arg\,\det\,\mathbf{B}\right)}{\sqrt{(2\pi)^D|\det\,\mathbf{B}|}} \tag{20}$$

where $\arg\det\mathbf{B} \in \{0, \pm\pi\}$. When $\mathrm{M}_0 = \begin{pmatrix}\mathbf{A} & 0\\ \mathbf{C} & \mathbf{A}^{-1}\end{pmatrix}$, the kernel collapses to a Dirac δ^D and the transformation is no longer integral, but *point-to-point*:

$$(\mathcal{C}_{\mathrm{M}_0} f)(\mathbf{q}) = \frac{\exp\,\mathrm{i}\left(\frac{1}{2}\mathbf{q}^{\mathrm{T}}\mathbf{C}\mathbf{A}^{-1}\mathbf{q}\right)}{\sqrt{\det\,\mathbf{A}}} f\left(\mathbf{A}^{-1}\mathbf{q}\right) \tag{21}$$

The normalization and phase are determined by the requirement that, at the unit transformation, $\lim_{\mathrm{M}\to 1} C_{\mathrm{M}}(\mathbf{q},\,\mathbf{q}') = \delta^D(\mathbf{q}-\mathbf{q}')$. The canonical integral transforms are *unitary* in $\mathcal{L}^2(\mathbb{R}^D)$:

$$(f, \mathcal{C}_{\mathrm{M}} g) = \left(\mathcal{C}_{\mathrm{M}}^{\dagger} f, g\right) = \left(\mathcal{C}_{\mathrm{M}}^{-1} f, g\right) = \left(\mathcal{C}_{\mathrm{M}^{-1}} f, g\right) \tag{22}$$

because

$$C_{\mathrm{M}}(\mathbf{q}',\mathbf{q})^* = C_{\mathrm{M}^{-1}}(\mathbf{q},\mathbf{q}') \tag{23}$$

As in the geometric case, integral canonical transforms are fully invertible and no information is lost between input and output images.

Because it is emblematic, we should write the $D=1$ case of (19)–(21) explicitly; it is

$$C_{\mathrm{M}}(q,q) = \frac{1}{\sqrt{2\pi\mathrm{i}B}} \exp\frac{\mathrm{i}}{2B}\left(Dq^2 - 2qq' + Aq'^2\right) \tag{24}$$

where $1/\sqrt{2\pi\mathrm{i}B} \equiv e^{-\mathrm{i}\pi/4}\exp\left(-\frac{\mathrm{i}}{2}\arg B\right)/\sqrt{2\pi|B|}$, and the lower-triangular case (21) follows directly.

Issues Pertaining Integral Transforms

Canonical integral transforms include the Fresnel and fractional Fourier transforms as particular one-parameter subgroups of translation and a phase space rotation (of a spin-like nature – see below). They also include the time evolution generated by all quadratic wave/quantum operators and allow for complex extensions of some transformation parameters, thereby adding coherent states to the functions subject to group theoretical treatment.

The Metaplectic Sign

The square root in (20) and (24) indicates a double valuation: indeed, when for symplectic $\mathbf{M}_1\mathbf{M}_2=\mathbf{M}_3$ we check carefully the composition of the integral kernels we find that

$$\int_{\mathbb{R}^D} \mathrm{d}^D\mathbf{q}'\, C_{\mathrm{M}_1}(\mathbf{q},\mathbf{q}')\, C_{\mathrm{M}_2}(\mathbf{q}',\mathbf{q}'') = \sigma_{1,2;3}\, C_{\mathrm{M}_3}(\mathbf{q},\mathbf{q}'') \tag{25}$$

with the *metaplectic* sign $\sigma_{1,2;3} = \mathrm{sign}(\det \mathbf{B}_3/\det \mathbf{B}_1\mathbf{B}_2)$.

Although overall signs are not commonly detected in wavefields, this sign 'ambiguity' is real and mathematically unavoidable. It is due to the fact that the set of canonical *integral* transforms covers the geometric $\mathsf{Sp}(2D,\mathbb{R})$ group *twice*. The issue reminds us of the double cover that the spin group affords over the group of space rotations. A handle on this matter is provided by the kernel of the $1D$ Fourier transform $\mathscr{F}$, namely $F(q,q') = e^{\mathrm{i}qq'}/\sqrt{2\pi}$, versus the *canonical* Fourier transform kernel for $\mathbf{F} = \left(\begin{smallmatrix}0 & -1\\ 1 & 0\end{smallmatrix}\right) = \mathbf{\Omega}$, which is $C_{\mathbf{F}}(q,q') = e^{\mathrm{i}\pi/4}e^{\mathrm{i}qq'}/\sqrt{2\pi}$. Thus $\mathcal{C}_{\mathrm{F}} = e^{\mathrm{i}\pi/4}\,\mathcal{F}$ and, while $\mathscr{F}^4 = 1$, one has $\mathcal{C}_{\mathrm{F}}^4 = -1$, and only $\mathcal{C}_{\mathrm{F}}^8 = 1$. The manifold of the symplectic group has the topology of a one-sheeted hyperboloid, whose *waist* is the 'circle' subgroup of isotropic fractional Fourier transforms, which may be infintely covered. The double cover is called the *metaplectic* group $\mathsf{Mp}(2D,\mathbb{R})$, of which linear canonical transforms are a faithful representation.

Exponentials of Second-Order Differential Operators

There is an important connection between canonical integral transforms and exponentials of *second*-order differential operators which follows the generation of the geometric $\mathsf{Sp}(2D,\mathbb{R})$ presentation in (9)–(11) and in (12), replacing q_i, p_j with $\hat{Q}_i, \hat{P}_j$, Poisson brackets with commutators, and $\exp(\circ)$ with $\exp(\mathrm{i}\circ)$.

There follow properties of self-reproduction of certain functions under a subgroup of canonical transforms: for example, writing $\mathcal{C}_{\mathrm{M}} \equiv \mathcal{C}(\mathbf{M})$, the Hermite-Gauss wavefunctions $\Psi_n(q) = e^{-q^2}/2H_n(q)/\sqrt{2^n n!\sqrt{\pi}}$ in a planar waveguide satisfy

$$\exp\left(\mathrm{i}t\frac{1}{2}\left(\hat{P}^2+\hat{Q}^2\right)\right)\Psi_n(q) = \mathcal{C}\left(\exp t\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}\right)\Psi_n(q) = \left(\mathcal{C}\begin{pmatrix}\cos t & -\sin t\\ \sin t & \cos t\end{pmatrix}\Psi_n\right)(q) = e^{\mathrm{i}\left(n+\frac{1}{2}\right)}\Psi_n(q) \tag{26}$$

Upon decomposing $\left(\begin{smallmatrix}a & b\\ c & d\end{smallmatrix}\right) = \left(\begin{smallmatrix}\alpha & 0\\ \gamma & \alpha^{-1}\end{smallmatrix}\right)\left(\begin{smallmatrix}\cos t & -\sin t\\ \sin t & \cos t\end{smallmatrix}\right)$, and using (26) and (21), one can then find the general linear canonical transform of these Hermite-Gauss functions as

$$\left(\mathcal{C}\begin{pmatrix}a & b\\ c & d\end{pmatrix}\Psi_n\right)(q) = \frac{\exp\left(-\frac{d-\mathrm{i}c}{a+\mathrm{i}b}\frac{q^2}{2}\right)}{\sqrt{2^n n!\sqrt{\pi}\left(\frac{a+\mathrm{i}b}{a-\mathrm{i}b}\right)^n}}\,\frac{H_n\left(q/\sqrt{a^2+b^2}\right)}{\sqrt{a+\mathrm{i}b}} \tag{27}$$

Similar manipulations can be performed with the eigenfunctions of all quadratic operators: the repulsive oscillator $\frac{1}{2}\left(\hat{P}^2-\hat{Q}^2\right)$, free waves $\exp(\mathrm{i}kq)/\sqrt{2\pi}$ of $\frac{1}{2}\hat{P}^2$, eigenfunctions $\sim q^{\mathrm{i}\lambda-\frac{1}{2}}$ of the dilatation generator $\frac{1}{2}(\hat{Q}\hat{P}+\hat{P}\hat{Q})$, or $\delta(q-\lambda)$ of $\hat{Q}^2$. Finally, Airy eigenfunctions of the linear potential $\frac{1}{2}\hat{P}^2+\hat{Q}$ can be brought into the picture with the addition of linear terms of $\hat{Q}$, $\hat{P}$, and $\hat{1}$ to generate the six-parameter *Weyl-symplectic* group $\mathsf{WSp}(2,\mathbb{R})$.

Complex Extension of Canonical Transforms

The parameters of the symplectic groups can be extended into a *complex* domain, characterizing thus $\mathsf{Sp}(2D,\mathbb{C})$. The integral kernel (19) remains valid as long as it is an oscillatory and/or *decreasing* Gaussian in $\mathrm{q}'\in\mathbb{R}^D$. This condition requires that the term $\mathrm{i}\,\mathbf{q}'^{\mathsf{T}}\mathbf{A}\mathbf{B}^{-1}\mathbf{q}'$ must have a zero or negative real part. In the $1D$ case of (24) this means that $\mathrm{Im}\,(b/a)\le 0$.

Out of a Dirac delta at c, $\delta_c(q)\equiv\delta(q-c)$, one can produce a Gaussian of width $\omega>0$ through the complex canonical transform

$$G_\omega(q-c) := \left(\mathcal{C}\begin{pmatrix}1 & -\mathrm{i}\omega\\ 0 & 1\end{pmatrix}\delta_c\right)(q) = \frac{1}{\sqrt{2\pi\omega}}\exp\left(-\frac{1}{2\omega}(q-c)^2\right) \tag{28}$$

Thus $\mathcal{C}\left(\begin{smallmatrix}1 & -\mathrm{i}\tau\\ 0 & 1\end{smallmatrix}\right)$ is the complex canonical transform for *diffusion* in time $\tau\ge 0$. It is no longer unitary in $\mathcal{L}^2(\mathbb{R})$, but within certain conditions can be made unitary in Bargmann-type Hilbert spaces over the complex domain.

The evolution of Gaussian wavefunctions in a harmonic waveguide, under fractional Fourier transforms, or under a quantum harmonic oscillator potential in time τ, can be computed using matrix algebra,

$$G_\omega(q-c;\tau) = \left(\mathcal{C}\begin{pmatrix}\cos\tau & -\sin\tau\\ \sin\tau & \cos\tau\end{pmatrix}\mathcal{C}\begin{pmatrix}1 & -\mathrm{i}\omega\\ 0 & 1\end{pmatrix}\delta_c\right)(q) = \left(\mathcal{C}\begin{pmatrix}\cos\tau & -\sin\tau & -\mathrm{i}\omega\cos\tau\\ \sin\tau & \cos\tau & -\mathrm{i}\omega\sin\tau\end{pmatrix}\delta_c\right)(q) \tag{29}$$

Coherent states are also complex canonical transforms of a Dirac δ_c. These are $\Upsilon_c(q) := \exp(c\hat{A}^\dagger)\Psi_0(q)$, with the raising operator $\hat{A}^\dagger := \frac{1}{\sqrt{2}}(\hat{Q}-\mathrm{i}\hat{P})$, the ground oscillator state $\Psi_0(q) = \pi^{-1/4}e^{-q^2/2}$, and of displaced center $c\sqrt{2}$,

$$\Upsilon_c(q) = \pi^{-1/4}e^{\frac{1}{2}c^2}\exp\left(-\frac{1}{2}\left(q-c\sqrt{2}\right)^2\right) = (2\pi)^{1/4}\left(\mathcal{C}\begin{pmatrix}1/\sqrt{2} & -\mathrm{i}/\sqrt{2}\\ -\mathrm{i}/\sqrt{2} & 1/\sqrt{2}\end{pmatrix}\delta_c\right)(q) \tag{30}$$

This the *Bargmann* transform that also maps the Hermite-Gauss wavefunctions $\Psi_n(q)$ on power functions $\sim z^n$ of a variable z on the complex plane. The time evolution of $\Upsilon_c(q)$ under a harmonic waveguide can be found as in (29), and results in the center $c(\tau)=ce^{i\tau}$ oscillating with position $\sqrt{2}\mathrm{Re}\ c(\tau)$ and momentum $\sqrt{2}\ \mathrm{Im}\ c(\tau)$, drawing out circles on phase space.

Radial Canonical Transforms

When an $\mathrm{Sp}(2D,\mathbb{R})$ matrix has diagonal blocks $\begin{pmatrix} a\mathbf{1} & b\mathbf{1} \\ c\mathbf{1} & d\mathbf{1} \end{pmatrix}$, canonical transforms of functions with definite angular momentum are reduced to $1D$ *radial* canonical transforms that conserve the angular momenta. The $2D$ kernel can be separated in polar coordinates as

$$q_1 = r\cos\theta, \quad q_2 = r\sin\theta, \quad r\in\mathbb{R}^+=[0,\infty), \quad \theta\in\mathbb{R} \bmod 2\pi, \tag{31}$$

resulting in the measure $\mathrm{d}^2\mathbf{q}=r\,\mathrm{d}r\,\mathrm{d}\theta$. The generator of rotations, $\hat{L}=\hat{Q}_1\hat{P}_2-\hat{Q}_2\hat{P}_1=-\mathrm{i}\partial_\theta$ is invariant under all $\mathcal{C}\begin{pmatrix} a\mathbf{1} & b\mathbf{1} \\ c\mathbf{1} & d\mathbf{1} \end{pmatrix}$'s due to (6) and so are its eigenspaces of functions $f(r)e^{\mathrm{i}m\theta}$, for all integers $m\in\mathbb{Z}$. When ϕ is the angle between $\mathbf{q}\cdot\mathbf{q}'=rr'\cos\phi$, in each of these spaces we can perform the integral over ϕ to obtain

$$e^{\mathrm{i}m\theta}C_{\mathrm{M}}^{(m)}(r,r') := \int_{-\pi}^{\pi}\mathrm{d}\phi C_{\mathrm{M}}(\mathbf{q},\mathbf{q}')e^{\mathrm{i}m(\theta-\phi)} \tag{32}$$

This provides kernel for the m-radial canonical transforms,

$$f_{\mathrm{M}}^{(m)}(r) \equiv \left(\mathcal{C}_{\mathrm{M}}^{(m)}f\right)(r) = \int_{R^+} r\,\mathrm{d}r' C_{\mathrm{M}}^{(m)}(r,r')f(r') \tag{33}$$

$$C_{\mathrm{M}}^{(m)}(r,r') = e^{\mathrm{i}\pi(m-1)/2}\frac{1}{b}\exp\left(\frac{\mathrm{i}}{2b}\left(dr^2+ar'^2\right)\right)J_m\left(\frac{rr'}{b}\right) \tag{34}$$

where $J_m(z)$ is the Bessel function of the first kind. These transforms are unitary in the Hilbert space $\mathcal{L}^2(\mathbb{R}^+)$, with the inner product

$$(f,g)_{\mathcal{L}^2(\mathbb{R}^+)} := \int_0^\infty r\,\mathrm{d}r\, f(r)^* g(r) \tag{35}$$

Among the $1D$ quadratic generators, squared momentum now displays a 'centrifugal' term:

$$\hat{P}_{(m)}^2 = -\left(\frac{\mathrm{d}^2}{\mathrm{d}r^2}+\frac{1}{r}\frac{\mathrm{d}}{\mathrm{d}r}-\frac{m^2}{r^2}\right) \tag{36}$$

while $\frac{1}{2}\left(\hat{Q}_{(m)}\hat{P}_{(m)}+\hat{P}_{(m)}\hat{Q}_{(m)}\right)=r\mathrm{d}/\mathrm{d}r$ and $\hat{Q}_{(m)}^2=r^2$. The eigenfunctions will thus be of the corresponding quantum potentials, such as Laguerre-Gauss functions for the harmonic waveguide Hamiltonian $\frac{1}{2}\left(\hat{P}_{(m)}^2+\hat{Q}_{(m)}^2\right)$.

Of course, the set $\mathcal{C}_{\mathrm{M}}^{(m)}$ with $\mathrm{M}\in\mathrm{Sp}(2,\mathbb{R})$ also represents that group, and belongs to *Bargmann's discrete* representation series D_+^k for $k=\frac{1}{2}(m+1)$. The well-known *oscillator* representation of $\mathrm{Sp}(2,\mathbb{R})$ that occupied the previous two sections belongs to $D_+^{1/4}\oplus D_+^{3/4}$ for even and odd states on $\mathbb{R}$. There, $m=\mp\frac{1}{2}$ corresponds to kernels with $J\left\{\begin{matrix}\cos z\\ \sin z\end{matrix}\right.$ on $\mathbb{R}^+$, which on the full real line reconstitute the oscillator representation that is double valued because it corresponds to half-integer m. For integer m the representations are single-valued so the composition sign σ in (25) is unity.

Acknowledgements

Support for this research has been provided by the *Óptica Matemática* projects of UNAM, presently PAPIIT IN101115.

See also: The Fractional Order Fourier Transform and Fresnel Diffraction

Further Reading

Bargmann, V., 1947. Irreducible unitary representations of the Lorentz group. Ann. Math. 48, 568–642.
Bargmann, V., 1961. On a Hilbert space of analytic functions and an associated integral transform, Part I. Commun. Pure Appl. Math. 20, 187–214.
Bargmann, V., 1970. Group representation in Hilbert spaces of analytic functions. In: Gilbert, P., Newton, R.G. (Eds.), Analytical Methods in Mathematical Physics. New York, NY: Gordon & Breach, pp. 27–63.
Collins Jr, S.A., 1970. Lens-system diffraction integral written in terms of matrix optics. J. Opt. Soc. Am. 60, 1168–1177.
Condon, E.U., 1937. Immersion of the Fourier transform in a continuous group of functional transformations. Proc. Natl. Acad. Sci. 23, 158–163.
Forbes, G.W., Man'ko, V.I., Ozaktas, H.M., Simon, R., Wolf, K.B. (Eds.), 2000. Feature issue on wigner distributions and phase space in optics. J. Opt. Soc. Am. A 17 (12), 2440–2463
Healy, J.J., Alper Kutay, M., Ozaktas, H.M., Sheridan, J.T. (Eds.), 2016. Linear Canonical Transforms. Theory and Applications. New York: Springer.
Kauderer, M., 1994. Symplectic Matrices. First Order Systems and Special Relativity. Singapore: World Scientific.

Liberman, S., Wolf, K.B., 2015. Independent simultaneous discoveries visualized through network analysis: The case of linear canonical transforms. Scientometrics 104, 715–735.
Moshinsky, M., Quesne, C., 1971. Linear canonical transformations and their unitary representation. J. Math. Phys. 12, 1772–1780.
Namias, V., 1980. The fractional order Fourier transform and its applications in quantum mechanics. IMA J. Appl. Math. 25, 241–265.
Ozaktas, H.M., Zalevsky, Z., Alper Kutay, M., 2001. The Fractional Fourier Transform with Applications in Optics and Signal Processing. Chichester: Wiley.
Quesne, C., Moshinsky, M., 1971. Linear canonical transformations and matrix elements. J. Math. Phys. 12, 1780–1783.
Rodrigo, J.A., Alieva, T., Bastiaans, T.J., 2011. Phase space rotators and their applications in optics. In: Cristóbal, G., Schelkens, P., Thienpont, H. (Eds.), Optical and Digital Image Processing: Fundamentals and Applications. Weinheim: Wiley-VCH Verlag, pp. 251–271.
Sánchez-Mondragón, J., Wolf, K.B. (Eds.), 1986. *Lie Methods in Optics*. Lecture Notes in Physics 250. Heidelberg: Springer.
Simon, R., Wolf, K.B., 2000. Fractional Fourier transforms in two dimensions. J. Opt. Soc. Am. A 17, 2368–2381.
Wolf, K.B., 1974. Canonical transforms. I. Complex linear transforms. J. Math. Phys. 15, 1295–1301.
Wolf, K.B., 1979. Integral Transforms in Science and Engineering. New York: Plenum.
Wolf, K.B. (Ed.), 1989. *Lie Methods in Optics. II*. Lecture Notes in Physics 352. Heidelberg: Springer.
Wolf, K.B., 2004. Geometric Optics on Phase Space. Heidelberg: Springer.

Phase-Space Representations of Freeform Optical Systems

Alois M Herkommer, University of Stuttgart, Stuttgart, Germany

Introduction

Since the early days of optics the classical shape of an optical surface used to be spherical or at least rotational symmetric, mainly because for this shape accurate manufacturing methods have been available. Accordingly the majority of optical systems was, and still is, rotational symmetric. Based on that symmetry assumption the well-known aberration theories, such as Seidel aberration theory (Gross, 2005), or Aldis calculus (Cox, 1964; Brewer, 1976) have mathematically been developed and intensely used for system analysis. However modern manufacturing methods, such as high precision diamond turning or 3D-printing, today allow for the accurate generation of optical surfaces which are free of any symmetry. Such surfaces are attractive to the optical designer, since they offer additional degrees of freedom to correct for astigmatic or other aberrations in complex folded systems (Thompson and Rolland, 2012). However, as freeform systems per definition do not allow any assumptions on symmetry, the well-developed aberration theories cannot be applied and more sophisticated analysis, such as nodal aberration theory (Fuerschbach *et al.*, 2014), is required. However these approaches are quite complex and require in-depth mathematical analysis of the system. Therefore it is desirable to find alternate simple methods to calculate and visualize aberration contributions within freeform systems, in order to allow the designer to analyze the optical system in a similar simple way as for rotational symmetric systems.

In this article we will apply the method of phase space in optics (Testorf *et al.*, 2010; Torre, 2005) for this purpose. The method allows an alternate access to the analysis of optical systems by employing an illustration of positions and angles of limiting rays. This has proven to be especially helpful for paraxial system analysis, but as we will see also aberrations can be very intuitively discussed in this context. Since symmetry assumptions are not required, this method proves to be a general tool to analyze and visualize freeform systems.

Within this article we will first introduce the general principles of phase space in geometrical optics. In a next step we will employ the method to visualize aberrations in symmetric and non-symmetric systems for the simplified two-dimensional case. Extending the analysis to all dimensions is however possible and finally allows an exact mathematical treatment and extraction of individual aberration contributions.

The Concept of Phase Space in Geometrical Optics

The concept of phase space in geometrical optics is based on an illustration of ray angles versus ray positions. This is most easily explained if we for the moment only consider meridional rays in the xz-plane of an optical system, as illustrated in **Fig. 1**. Any meridional ray at a given z-position is completely defined by two quantities, namely the distance to the optical axis x and the angle θ to the optical axis. If we furthermore weight the ray direction with the refractive index n of the propagation medium, we can define

$$u = n \cdot \tan\theta \tag{1}$$

Thus the vector $\mathbf{r} = (x,u)^T$ completely defines the ray-tracing behavior of any ray in an optical system. The vector components (x,u) of the ray can be illustrated as one point in a xu-diagram, which is called the phase space diagram. Propagation of the ray in an optical system in consequence corresponds to trajectories, respectively transformations in phase space. In **Fig. 1** the relationship between standard ray propagation versus the corresponding trajectory in phase space is illustrated.

In the paraxial regime the ray-propagation and the corresponding trajectory in phase space is closely related to the so-called ABCD matrix-optics, respectively linear optical systems theory (Kloos, 2007). In this approximation the real optical system can mathematically be replaced by a sequence of linear transformations. This is illustrated in **Fig. 2** where a sequence of reference surfaces D_i and D'_i, located at the vertex of the refractive surface, is introduced. The paraxial ray propagation then corresponds to

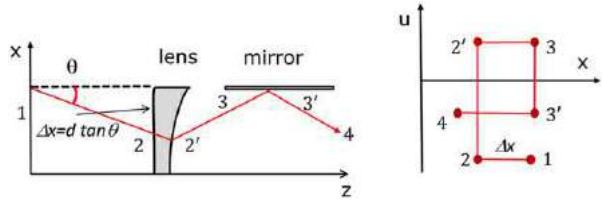


Fig. 1 Relationship between ray-propagation (left) and phase space trajectory (right) in an optical system. Reproduced from Rausch, D., Herkommer, A. 2012. Phase space approach to the use of integrator rods and optical arrays in illumination systems. Advanced Optical Technologies. Available at: https://doi.org/10.1515/aot-2011-0002.

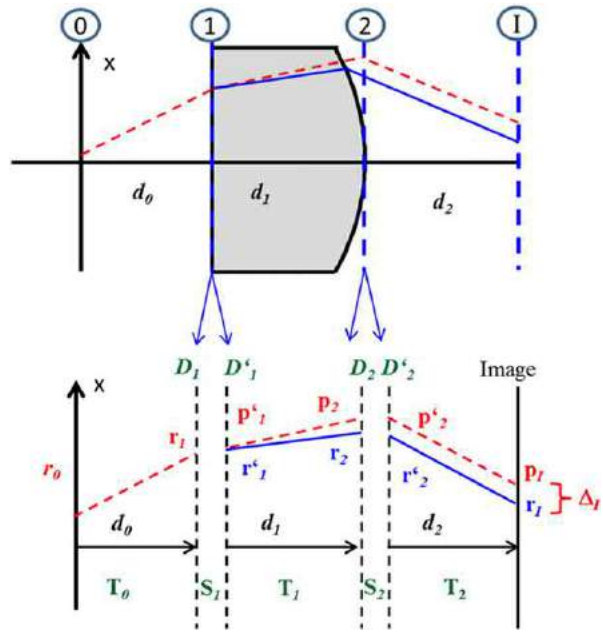


Fig. 2 Real geometry (top) and paraxial representation (bottom) of a refractive lens system. The dashed line represents the paraxial ray-tracing, the solid line real ray-tracing.

the application of propagation steps $\mathbf{T}_i$ in between the surfaces D_i and D_{i+1}, and paraxial refraction $\mathbf{S}_i$ from D_i to D'_i at the spherical surfaces. If we denote the paraxial ray vector by $\mathbf{p}$, the matrix-optics rules define the following set of linear transformations as the ray propagates:

$$\mathbf{p}_{i+1} = \mathbf{T}_i \cdot \mathbf{p}_i \text{ where } \mathbf{T}_i = \begin{pmatrix} 1 & d/n \\ 0 & 1 \end{pmatrix} \tag{2}$$

$$\mathbf{p}'_i = \mathbf{S}_i \cdot \mathbf{p}_i \text{ where } \mathbf{S}_i = \begin{pmatrix} 1 & 0 \\ -\phi & 1 \end{pmatrix} \tag{3}$$

Here d denotes the distance between surfaces along the optic axis and ϕ is the optical power of the surface. Thus any paraxial ray, corresponding to an input ray $\mathbf{r}_0$ can paraxial be traced to surface D_i by application of a sequence of linear operations. Mathematically this is:

$$\mathbf{p}_i = \underbrace{\mathbf{T}_{i,i-1} * \mathbf{S}_{i-1} \cdots \mathbf{S}_1 * \mathbf{T}_{1,0}}_{\mathbf{M}_{i,0}} * \mathbf{r}_0 \tag{4}$$

We can now translate the paraxial ray-tracing behavior into transformations within phase space. A visualization of appropriate reference rays in phase space, for example, the marginal ray and the chief ray, reveals the basic transformation properties and parameters of an optical system. This is illustrated in **Fig. 3** for a 2f-lens system, representing the propagation from the front focal plane to the back focal plane of a single lens. According to Eq. (2) the linear matrix $\mathbf{T}$ for free propagation corresponds to a shear into the x-direction, whereas a paraxial lens $\mathbf{S}$, Eq. (2), corresponds to a shear in the angular u-direction. Thus the total propagation from the front focal plane to the back focal plane corresponds to a 90° rotation in phase space, which is expected since the 2f lens system translates angles into position and vice versa. It is also worth to note that area in phase space has an important interpretation. If the chief ray $(\bar{x}, \bar{u})$ and the marginal ray (x,u) of an optical system are employed as reference points, then the product $L = \bar{u} \cdot x - \bar{x} \cdot u$ is equivalent to the Lagrange invariant (Gross, 2005), which is a conserved quantity in conventional optical systems. In the phase space diagram this product corresponds to the shaded area, as illustrated in **Fig. 3**. In this sense we can generally state that area in phase space is conserved and paraxial propagation corresponds to a sequence of affine transformations to this area.

Visualization of Aberrations in Phase Space

Already analyzing the linear behavior in phase space reveals a lot of information about the system, especially the primary quantities like magnification or focal length (Kloos, 2007). However the paraxial equations do not contain any aberration effects, since each paraxial ray will behave in the same linear way and lead to the same transformation behavior. For aberration analysis it is therefore necessary to record and visualize the exact ray-tracing behavior and compare it to the above linear behavior. To do so,

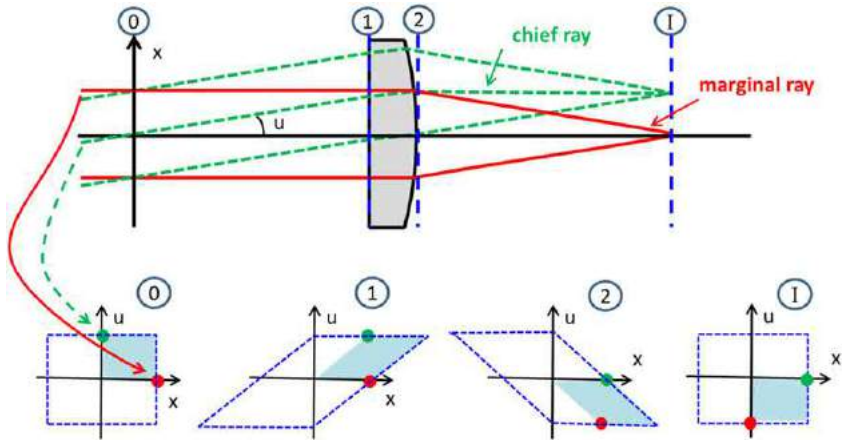


Fig. 3 Illustration of chief and marginal reference rays in phase space and the paraxial transformation representing the system. The enclosed area corresponds to the Lagrange invariant.

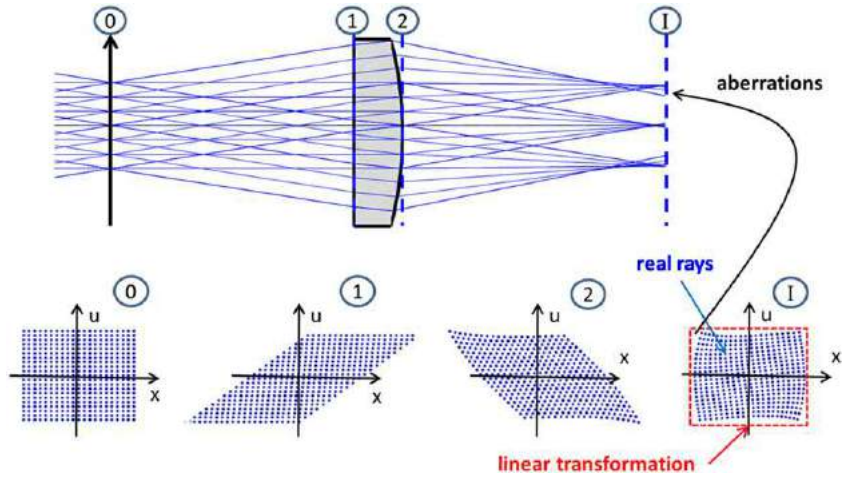


Fig. 4 Illustration of real ray behavior in phase space, revealing nonlinear distortions corresponding to aberrations.

we can introduce a series of planar dummy surfaces D_i and D'_i at the vertex of each refractive surface, as illustrated in Fig. 2. In a sequential ray-tracer the surface D_i is placed directly before the real surface, and records the input ray, whereas the surface D'_i is arranged at the same position but sequentially arranged after the real surface und thus monitors the output ray. By performing raytracing the real ray intersection data $\mathbf{r}_i$ at each "input"-dummy surface D_i, and $\mathbf{r}'_i$ on the corresponding "output"-dummy surfaces D'_i can be recorded. Since the dummy surfaces are at the same position as the paraxial power of the linear system model, we can directly compare the paraxial behavior and real raytracing behavior. To illustrate this, a grid of rays, sampling the object (x) and aperture (u) space, is traced through the system and each ray is plotted in the phase space diagram. The result is presented in Fig. 4 and reveals the phase space transformation of the real system.

If we compare this result to the paraxial transformation as was shown in the previous Fig. 3 we notice, that non-linear deformations of the ray-pattern, respectively the phase space area, become apparent. At the image plane these deviations have a clear meaning and are illustrated in detail in Fig. 5. Since one ray fan at the image plane, corresponding to one column of the ray-grid, represents a set of rays that arrive at the same image point but under different angles, the lateral shift relative to the central ray simply corresponds to the transverse aberrations in x-direction. This is depicted in Fig. 5(b). In other words at the image plane we can define a general aberration vector, describing the difference between the linear ray $\mathbf{p}_I$ and the corresponding real ray $\mathbf{r}_I$

$$\mathbf{\Delta}_I = \mathbf{r}_I - \mathbf{p}_I = \begin{pmatrix} \Delta x_I \\ \Delta u_I \end{pmatrix} \tag{5}$$

Obviously the spatial component Δx_I of this vector then represents the transverse aberration of this ray. The angular component can also be interpreted as the pupil aberration of the system. Following along these lines, we can define an aberration vector at any surface i within the system in the same way.

$$\mathbf{\Delta}_i = \mathbf{r}_i - \mathbf{p}_i \tag{6}$$

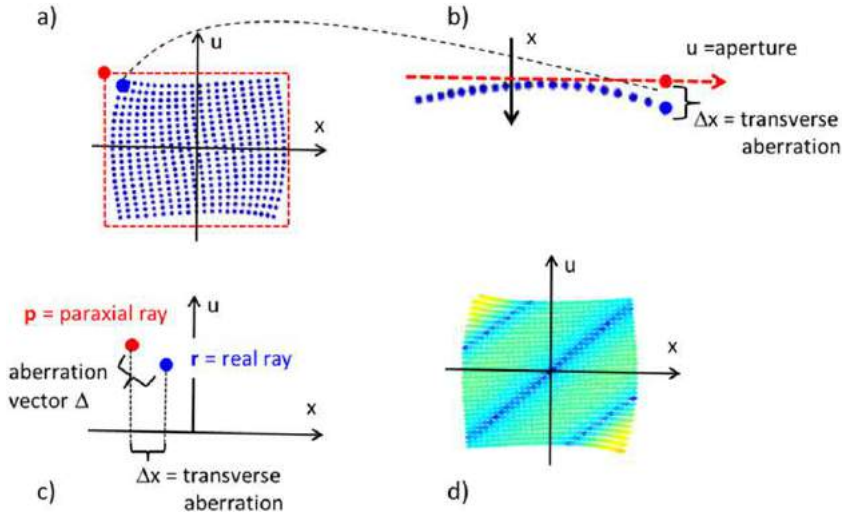


Fig. 5 Difference in ray positions at the image corresponding to aberrations: (a) real ray grid at the image plane and linear behavior, (b) one ray fan corresponding to transverse aberrations (c) aberration vector in phase space, (d) color coded ray grid, where the color indicates the length of the aberration vector.

This vector represents the local deviation of the real ray from the linear ray in phase space and is shown in **Fig. 5(c)**. The length of this aberration vector may thus be interpreted as a general measure of the current magnitude of the aberration of this ray. In consequence we can visualize the aberration of any ray in phase space by illustrating the length of the corresponding aberration vector $|\boldsymbol{\Delta}|$, for example, as a color code. In **Fig. 5(d)** we show the color coded aberration field at the image.

It is worth to note, that since we have defined $u = n \tan\theta$ via Eq. (1), the free propagation between parallel reference surfaces D_i and D_{i+1} is of exactly the same linear form as the paraxial matrix in Eq. (2). Thus real raytracing in between parallel dummy surfaces exactly corresponds to the paraxial raytracing transformation.

$$\mathbf{r}_{i+1} = \mathbf{p}_{i+1} = \mathbf{T}_i \cdot \mathbf{r}_i \tag{7}$$

In other words free propagation does not introduce differences between paraxial and real ray propagation, i.e., aberrations. Therefore as expected the aberrations within the system are only due to the spherical surfaces and correspond to the difference between the real raytracing via the spherical interface according to Snell's law versus the linear behavior according to the paraxial refraction (Herkommer, 2013).

Visualizing Aberrations in Freeform Systems

The above described method is quite general, since all that is required to calculate the aberration vector of Eq. (6) are real ray-tracing data and linear ray-propagation data. In consequence this analysis can be transferred to more general optical systems, such as freeform systems (Chen and Herkommer, 2016). To illustrate this we use the following freeform prism from literature (Takahashi, 1997), which is often employed as a test and reference system for freeform designs. The design is illustrated in **Fig. 6(a)** and the detailed design data are listed in **Table 1**. The illustrated prism geometry is, for example, employed in head-mounted displays to image a display into the pupil of the eye. Note that the design is usually obtained backwards, so in **Fig. 6(a)** the eye-pupil corresponds to the entrance pupil of the prism and the display corresponds to the image.

To illustrate aberrations in this freeform system the paraxial and real ray propagation in phase space can be compared, just as before for the rotational symmetric case. Again we denote the real rays by $\mathbf{r}$ and the paraxial rays by $\mathbf{p}$. However in a freeform system moreover a reference ray, serving as the reference axis and origin of the coordinate system, has to be defined, as we propagate through the system. In the rotational symmetric case this was not required, since the rotation axis naturally and uniquely defines the optical axis, however in freeform systems the reference has to be fixed. Here we choose the center ray on the pupil starting with an angle of 0° as a reference axis, as illustrated in **Fig. 6(b)**. This reference ray is traced through the freeform system, defining surface intersection points at each surface S1 to S4. At each intersection we again insert a pair of dummy surfaces D_i and D'_i similar to the rotational symmetric system. The position of both dummy surfaces is identical to the reference ray intersection point, however again the first dummy is mathematically before, whereas the second dummy is after the surface in a sequential ray-tracer. Moreover all dummy surfaces have to be tilted to be exactly orthogonal to the reference ray in order to properly define the ray angels relative to the reference ray. This is illustrated in **Fig. 6(b)** for the example of surface 2 only. The positions of all dummy surfaces are listed in **Table 1**. Note that positioning of these dummy surfaces can be performed in any ray-tracer after recording the global ray-tracing data and intersections of the reference ray.

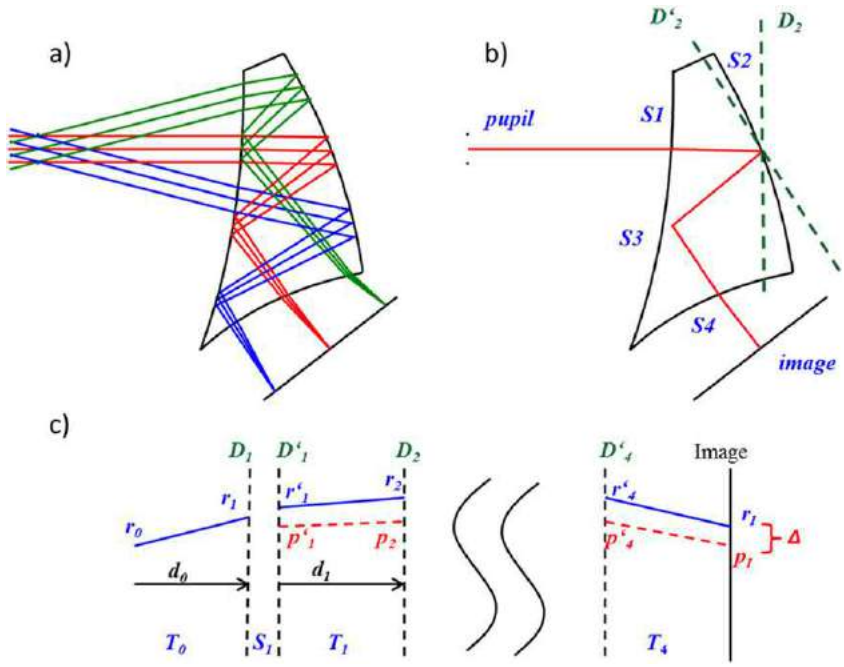


Fig. 6 Illustration of a patented freeform prism (a) and corresponding location of auxiliary reference planes (b), leading to an unfolded auxiliary system of a sequence of dummy surfaces (c).

What should be mentioned is that the tilt and decenter of the surfaces in **Table 1** are all in global coordinates and referred to the entrance pupil (surface 0). The horizontal and the vertical field angle of this lens as we analyze here is $\pm 15^\circ$ in each dimension. Besides, the pupil diameter is 4 mm. The surfaces 1, 2 and 3 are anamorphic freeform surfaces, where K_x and K_y represent the conic constant, AR and BR describe the 4th and 6th order symmetric coefficients, AP and BP illustrate the 4th and 6th order asymmetric coefficients, Y and Z correspond to the decenter of the surface in Y and Z direction, and α shows the surface tilt.

After the above design preparations we can perform raytracing and record the real raytracing data $\mathbf{r}_i$ on each "input"-dummy surface D_i, and $\mathbf{r'}_i$ on the corresponding "output"-dummy surfaces D'_i with the help of any ray-tracer. Just in the same way as we did in the rotational symmetric case, we can thus define an unfolded auxiliary system of sequential dummy surfaces to record the ray patterns in phase space, as illustrated in **Fig. 6(c)**. In order to compare the real rays to the paraxial rays we additionally have to determine the linear transformation matrices $\mathbf{S}_i$ and $\mathbf{T}_i$. For the propagation matrices $\mathbf{T}_i$ all that is required are the distances d_i between consecutive parallel dummy surfaces D'_i and D_{i+1}. These data are available from any ray-tracer. Determination of the linear transformation $\mathbf{S}_i$ of the tilted and decentered real surfaces is more challenging and several methods can be implemented: The real ray grid data at the dummy surfaces D_i and D'_i can be used to determine the best fit linear transformation $\langle \mathbf{S}_i \rangle$ from a sufficiently large set of rays and the corresponding set of equations (Chen and Herkommer, 2016). Alternatively we can use the predefined ABCD-function for tilted and decentered systems in Synopsis CodeV10.8 to calculate the paraxial transformation matrix $\mathbf{S}_i$. Both methods are appropriate to visualize aberrations fields, since the average linear transformation can be subtracted, however the later method will be used later in this paper for exact surface aberration calculations. In **Fig. 7** we illustrate the phase space behavior of the freeform prism. The color scale corresponds to difference between the real ray and the corresponding best fit linear ray and represents the length of the aberration vector $|\Delta|$, as defined in Eq. (6). **Fig. 7** thus illustrates the aberration field, which is built up as the rays propagate through the freeform system. The figure nicely illustrates the nodes of the aberration field, i.e., the areas where the aberration vector is small, moreover the missing symmetry leads to non-symmetric distributions.

Extension to 4d Phase Space

To explain the principles and allow visualization the above description was limited to the 2d phase space, since we only considered rays in the *xz*-plane. However this allows describing meridional rays within an optical system only. Already for rotational symmetric systems with finite field we are missing the saggital ray fans. Moreover in freeform systems rotational symmetry cannot be assumed. Therefore for the general case we have to consider a 4d phase space, which means that the position and direction of each ray in a certain reference plane has to be described by four quantities, namely $\mathbf{r}=(x,y,u,v)^T$. Here x,y are the

Table 1 Lens data of the freeform prism of reference including additional dummy surfaces

Surface	*Surface shape/radius (mm)*		*Surface position*	*Inclination angle(α)*	*Glass(n,v)*
0	Infinity (pupil)		Origin		
D1	Infinity		Z 30.002		
1	R_y −108.187	AR 5.542E-7		−14.700°	1.492, 57.471
	R_x −73.105	BR 8.176E-11	Y −24.028		
	K_y 0	AP −0.080	Z 26.360		
	K_x 0	BP −1.379			
D′1	Infinity		Z 30.002	−1.066°	1.492, 57.471
D2	Infinity		Y −0.251	−1.066°	1.492, 57.471
			Z 43.485		
2	R_y −69.871	AR-7.233E-11		36.660°	1.492, 57.471
	R_x −60.374	BR-4.529E-12	Y 19.109		
	K_y −0.1368	AP29.075	Z 33.339		
	K_x −0.123	BP −2.085			
D′2	Infinity		Y −0.251	38.376°	1.492, 57.471
			Z 43.485		
D3	Infinity		Y −11.858	38.376°	1.492, 57.471
			Z28.827		
3	R_y −108.187	AR5.542E-7		−14.700°	1.492, 57.471
	R_x −73.105	BR8.176E-11	Y −24.028		
	K_y 0	AP −0.080	Z 26.360		
	K_x 0	BP −1.379			
D′3	Infinity		Y −11.858	−55.019°	1.492, 57.471
			Y −11.858		
D4	Infinity		Y −23.067	−55.019°	1.492, 57.471
			Z 36.667		
4	77.772		Y −35.215	−47.770°	
			Z 18.817		
D′4	Infinity		Y −23.067	−50.668°	
			Z 36.667		
5 (Image)	Infinity		Y −30.892	−50.668°	
			Z 43.083		

Source: Reproduced from Chen, B., Herkommer, A.M., 2016. Generalized Aldis theorem for calculating aberration contributions in freeform systems. Optics Express 24 (23), 26999–27008.

intersection points of the ray with the reference surfaces and u,v are the index-weighted direction tangents, which can easily be computed from the normalized direction vector (L, M, N) and the refractive index n by using

$$u = n \cdot \frac{L}{N}, \quad v = n \cdot \frac{M}{N} \tag{8}$$

In consequence the above employed matrix calculus needs to be extended to 4d, which however is straightforward. If we for example consider free ray-propagation between two consecutive parallel reference planes D'_i and D_{i+1}, as illustrated in **Fig. 8**, the corresponding exact ray-transfer matrix is quite similar to the 2d case.

$$\mathbf{r}_{i+1} = \mathbf{T}_i \cdot \mathbf{r}_i' \text{ where } \mathbf{T}_i = \begin{pmatrix} \mathbf{I} & d \cdot \mathbf{I} \\ 0 & \mathbf{I} \end{pmatrix} \text{ and } \mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{9}$$

Thus only the introduction of the two-dimensional unit matrix $\mathbf{I}$ is required to extend the matrix $\mathbf{T}$ to the general case. Here it is again important to note, that since we employ the exact ray direction tangents in Eq. (8) and not any "paraxial" angles the above propagation matrix is exact, linear and identical to the paraxial transformation matrix also for larger angles. In other words again free propagation does not introduce any deviation between paraxial and exact ray-tracing.

Thus aberrations are only induced by the action of the real surface in between each dummy surface pair D_i and D'_i. The linear/paraxial transformation $\mathbf{S}_i$ of any 4d-ray input ray $\mathbf{r}_i$ into a paraxial output ray $\mathbf{p}'_i$ can be described by a linear matrix, which however now is 4-dimensional:

$$\mathbf{p}'_i = \mathbf{S}_i \cdot \mathbf{r}_i \text{ where } \mathbf{S}_i = \begin{pmatrix} A_{xx} & A_{xy} & B_{xx} & B_{xy} \\ A_{yx} & A_{yy} & B_{yx} & B_{yy} \\ C_{xx} & C_{xy} & D_{xx} & D_{xy} \\ C_{yx} & C_{yy} & D_{yx} & D_{yy} \end{pmatrix} \tag{10}$$

For the case of rotational symmetric lenses the matrix simplifies to a similar form as Eq. (9) and can be calculated form the surface power, in the same way as discussed in Eq. (2). In the general case we can determine the matrix either from a set of rays and

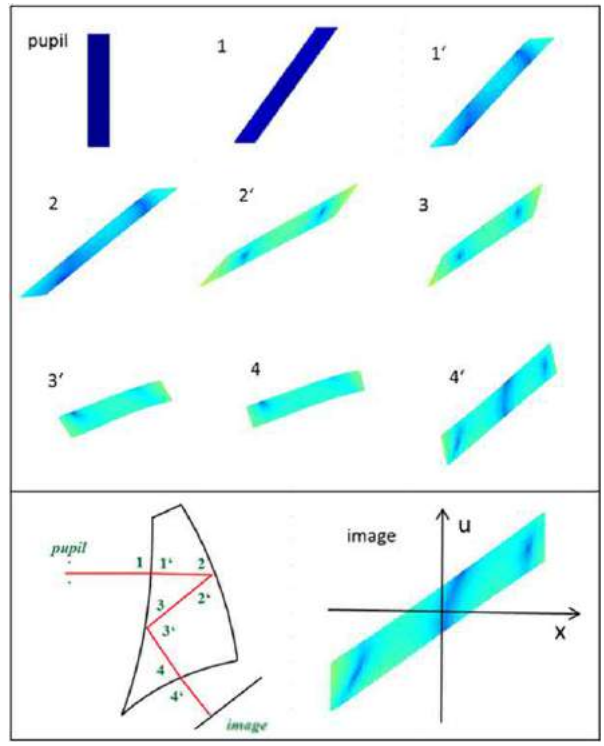


Fig. 7 Illustration of the phase space transformations corresponding to propagation through a freeform prism. The color code represents the normalized length of the aberration vector.

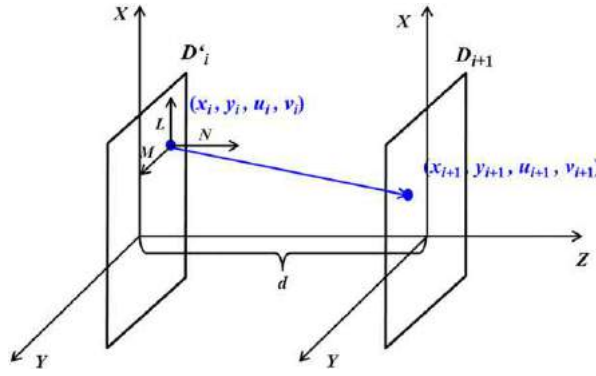


Fig. 8 Ray definition and propagation in between parallel reference planes. Reproduced from Chen, B., Herkommer, A.M., 2016. Generalized Aldis theorem for calculating aberration contributions in freeform systems. Optics Express 24 (23), 26999–27008.

then solving the equations to find the best-fit linear four-dimensional matrix. Another option is to employ the ABCD-function in Synopsis CodeV (Synopsis CodeV10.8 News), which is valid for decentered and tilted surfaces and defined for the general 4d-case. Application of this function in between D_i and D'_i will return the above 16-element linear transformation matrix. Note that in the general freeform case the sub-matrixes A, B, C and D will not be symmetric, since especially a freeform surface will contain different curvatures in x and y direction and also prismatic structures.

If we, in comparison to the above linear analysis, now employ any ray-tracer to find the real surface intersection coordinates $\mathbf{r}_i$ at all of the reference surfaces we again can define, according to Eq. (6), an aberration vector $\mathbf{\Delta}$, which however is 4-dimensional. Therefore in summary the complete method as described in 2d is mathematically straightforward to be extended to the general case. However illustration of a 4d-vector and aberration field is challenging and in consequence visualization of the full information is difficult. Thus typically a reduction (projection) of the data will be required for visualization. For example the length of the aberration vector $|\mathbf{\Delta}|$ can again be understood of as an aberration measure and illustrated as a color code, leading to similar pictures as shown in **Fig. 7** in any 2d-visualization of the 4d-parameter field.

Even though visualization of the general situation is difficult, calculation is not. We will show below that the 4d-calculus is able to quantify aberrations and aberration contributions within freeform systems.

Calculation of Aberration Contributions at the Image

For rotational symmetric systems a long established theory of aberrations exists, and different classes of aberrations are defined, such as spherical, coma and astigmatism. The corresponding third order aberrations can be derived from paraxial raytracing only, and individual surface contributions can be calculated (Gross, 2005). Also an exact calculation of higher order surface contributions to lateral aberrations at the image exists: Within Aldis theory (Cox, 1964; Brewer, 1976) the aberration calculation is based on real ray tracing and paraxial (linear) ray-tracing of a given input ray. Application of this theorem allows exact calculation of the surface induced contribution to lateral aberrations at the image plane including all orders. Unfortunately for freeform systems this theorem can, due to the lack of symmetry, not be applied.

The phase space method described here does not require symmetry and is, same as Aldis theorem, based only on paraxial and exact rays. We will now mathematical prove that it can be applied to exactly quantify surface aberration contributions in freeform systems. To do this, the aberration generation and propagation in phase space has to be understood. Per definition the general aberration vector at a reference surface is a 4d-vector describing the deviation between the real ray coordinates and the paraxial ray coordinates. To better explain the relationships and propagation laws for aberrations let us switch back to the 2d-case for visualization. However the mathematical treatment is applicable to both, 2d and 4d cases.

In Fig. 9 the aberration generation and propagation is illustrated for the simple system of Fig. 2, where only two surfaces are involved. We will now mathematically and graphically follow the ray and aberration vectors through phase space.

During the analysis we will consider the aberration vector in front and after of each surface until we reach the image plane. We start with a general input ray $\mathbf{r}_0$ at the object plane, as for example illustrated in Fig. 2. The free propagation to the first system surface D_1 is exact, via Eq. (7), and thus the real ray and the corresponding paraxial ray are equivalent:

$$\mathbf{r}_1 = \mathbf{p}_1 = \mathbf{T}_0 \cdot \mathbf{r}_0$$

Therefore the aberration vector in front of the real surface S1 is zero, as

$$\boldsymbol{\Delta}_1 = \mathbf{r}_1 - \mathbf{p}_1 = \mathbf{0}$$

As the ray passes the first surface (propagation from D_1 to D'_1) the ray picks ups some aberration, since the input ray splits into a paraxial ray $\mathbf{p}'_1$ (resulting from the linear surface matrix $\mathbf{S}_1$) and the real ray $\mathbf{r}'_1$ (refracted at the real surface), as illustrated in Fig. 9(a). Therefore we find the local aberration vector after surface S1 to be:

$$\boldsymbol{\Delta}'_1 = \mathbf{r}'_1 - \mathbf{p}'_1 = \underbrace{\mathbf{r}'_1 - \mathbf{S}_1\mathbf{r}_1}_{\delta_1}$$

Here and for later use we define the surface contribution $\boldsymbol{\delta}_i$ to be the difference between the real traced ray $\mathbf{r}'_i$ and the paraxial traced ray $\mathbf{S}_i\mathbf{r}_i$ at surface i as:

$$\boldsymbol{\delta}_i = \mathbf{r}'_i - \mathbf{S}_i\mathbf{r}_i \tag{11}$$

After free propagation from surface S1 to the front of surface S2 we find the transformed, respectively sheared aberration vector, as illustrated in Fig. 9(b). In mathematical terms this corresponds to the application of the propagation matrix:

$$\Delta_2 = \mathbf{r}_2 - \mathbf{p}_2 = \mathbf{T}_1\boldsymbol{\delta}_1$$

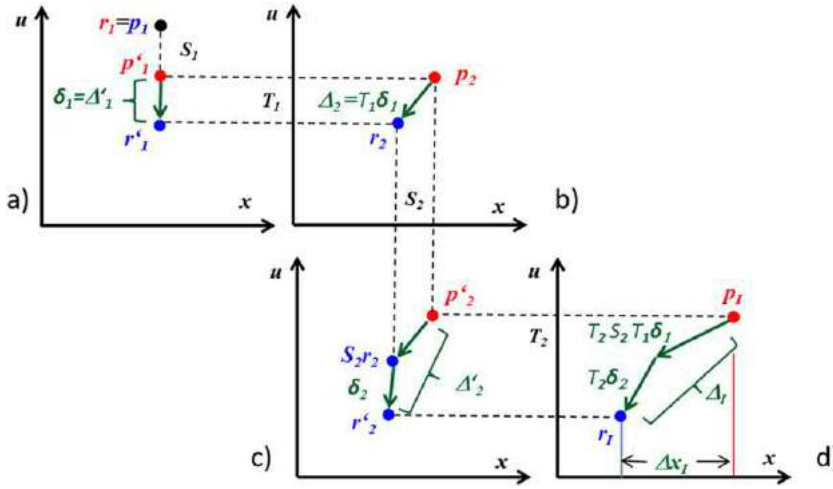


Fig. 9 Illustration of aberration propagation in phase space, where (a) to (b) represents the free propagation to surface S2, (b) to (c) illustrates the action of surface S2, and (c) to (d) resembles propagation to the image.

Similar as before the action of the subsequent surface S2 introduces some additional aberration contribution, which due to the above relations between the rays can be calculated as:

$$\boldsymbol{\Delta}'_2 = \mathbf{r}'_2 - \mathbf{p}'_2 = \mathbf{r}'_2 - \mathbf{S}_2\mathbf{p}_2 = \mathbf{r}'_2 - \mathbf{S}_2\mathbf{T}_1\mathbf{p}'_1 = \mathbf{r}'_2 - \mathbf{S}_2\mathbf{T}_1(\mathbf{r}'_1 - \boldsymbol{\delta}_1) = \mathbf{S}_2\mathbf{T}_1\boldsymbol{\delta}_1 + \underbrace{(\mathbf{r}'_2 - \mathbf{S}_2\mathbf{r}_2)}_{\delta_2}$$

The result can be interpreted in such a way that the surface S2 introduces an additional aberration term $\boldsymbol{\delta}_2$ which is added to the existing transformed (linear propagated) aberration vector of surface S1. In Fig. 9(c) we have again illustrated this graphically. Further propagation of the complete aberration vector to the front of surface S3 yields:

$$\Delta_3 = \mathbf{T}_2\mathbf{S}_2\mathbf{T}_1\boldsymbol{\delta}_1 + \mathbf{T}_2\boldsymbol{\delta}_2$$

In our graphical illustration of Fig. 9(d) this already corresponds to the image plane. At this point it has to be remembered that at the image plane the spatial component directly represents the lateral aberration of the particular ray. The equation above already suggests, that a recurrence relation can be established, since the aberration vector consists of terms corresponding to a sum of linear transformed aberrations $\boldsymbol{\delta}_i$ from all proceeding surfaces. Indeed if we have n surfaces in the system we can continue this calculation to the image plane to find the following general result.

$$\Delta_I = \mathbf{r}_I - \mathbf{p}_I = \sum_{i-1}^{n} \mathbf{M}_{I,i}\delta_i = \sum_{i-1}^{n} \Delta_I^i \tag{12}$$

Here $\mathbf{M}_{I,i}$ represents the linear propagation matrix from the exiting dummy surface D'_i to the image plane I and is given by:

$$\mathbf{M}_{I,i} = \mathbf{T}_n\mathbf{S}_n \ldots \mathbf{S}_{i+1}\mathbf{T}_i \tag{13}$$

Here the index n represents the last surface in the system, and $\mathbf{T}_n$ is the propagation from this last surface n to the image plane.

To summarize the above mathematical treatment: With Eq. (12) we have found a mathematical exact prove, that the total aberration of any ray at the image plane of an optical system consisting of n surfaces, can be expressed as the sum of individual surface contributions. Those individual surface contributions result from the differential rays at each surface i (i.e., the difference between linear and real ray tracing across surface i), which are linearly propagated to the image. In other words we have found a mathematical relation between surface aberration contributions and the exact ray aberrations at the image plane, including all

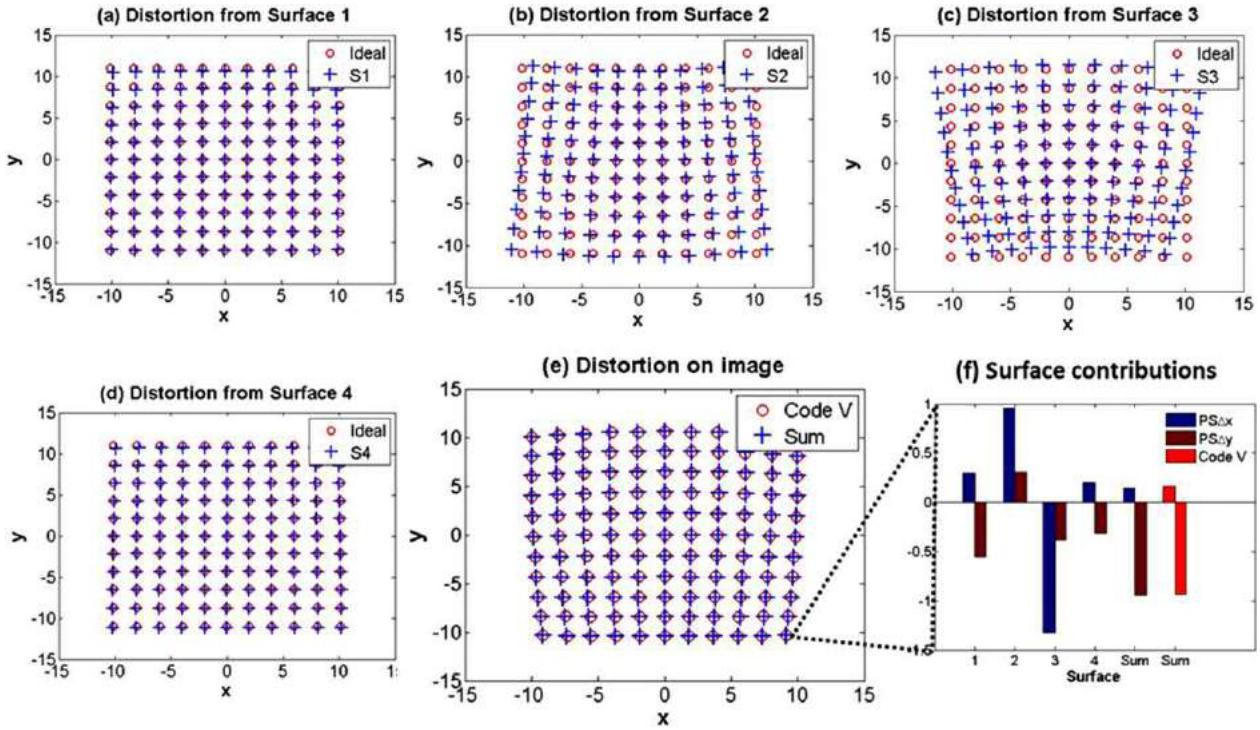


Fig. 10 Distortion analysis: In (a)–(d) the individual surface contributions are shown, where the red circles are the paraxial positions of the rays on image, and the blue cross represents the lateral distortion contribution of the surface, (e) illustrates the sum of the distortion contributions and is compared to the exact ray-tracing result at the image, (f) reveals the individual surface contributions for a single chief ray. Reproduced from Chen, B., Herkommer, A.M., 2016. Generalized Aldis theorem for calculating aberration contributions in freeform systems. Optics Express 24 (23), 26999–27008.

orders, quite similar to Aldis theory, however valid for freeform systems. The individual surface aberration contributions at the image plane, given in Eq. (12), are:

$$\Delta_I^i = \mathbf{M}_{I,i}\delta_i = \mathbf{M}_{I,i}(\mathbf{r}'_i - \mathbf{S}_i\mathbf{r}_i) \tag{14}$$

This expression contains only real ray-tracing data before and after each surface and the linear System matrix $\mathbf{M}_{I,i}$. Therefore this contribution vector can be obtained from the paraxial system representation and ray-tracing data only. No assumptions are made about symmetry or surface shape.

Aberration Contributions in a Freeform Prism

In order to demonstrate the application of the above mathematical treatment we apply it to the already introduced freeform prism. Following the above preconditioning, we have inserted dummy surfaces, as given in **Table 1**, and moreover determined all necessary transformation matrices $\mathbf{S}_i$, $\mathbf{T}_i$ and $\mathbf{M}_{I,i}$ by using CodeVs ABCD-linear system analysis (Synopsis CodeV10.8 News). Therefore calculation of the aberration contribution $\mathbf{\Delta}^i_I$ of every ray at the image is now possible via Eq. (14), if in a prior real ray-tracing the ray-coordinates $\mathbf{r}_i$ and $\mathbf{r}'_i$ are recorded.

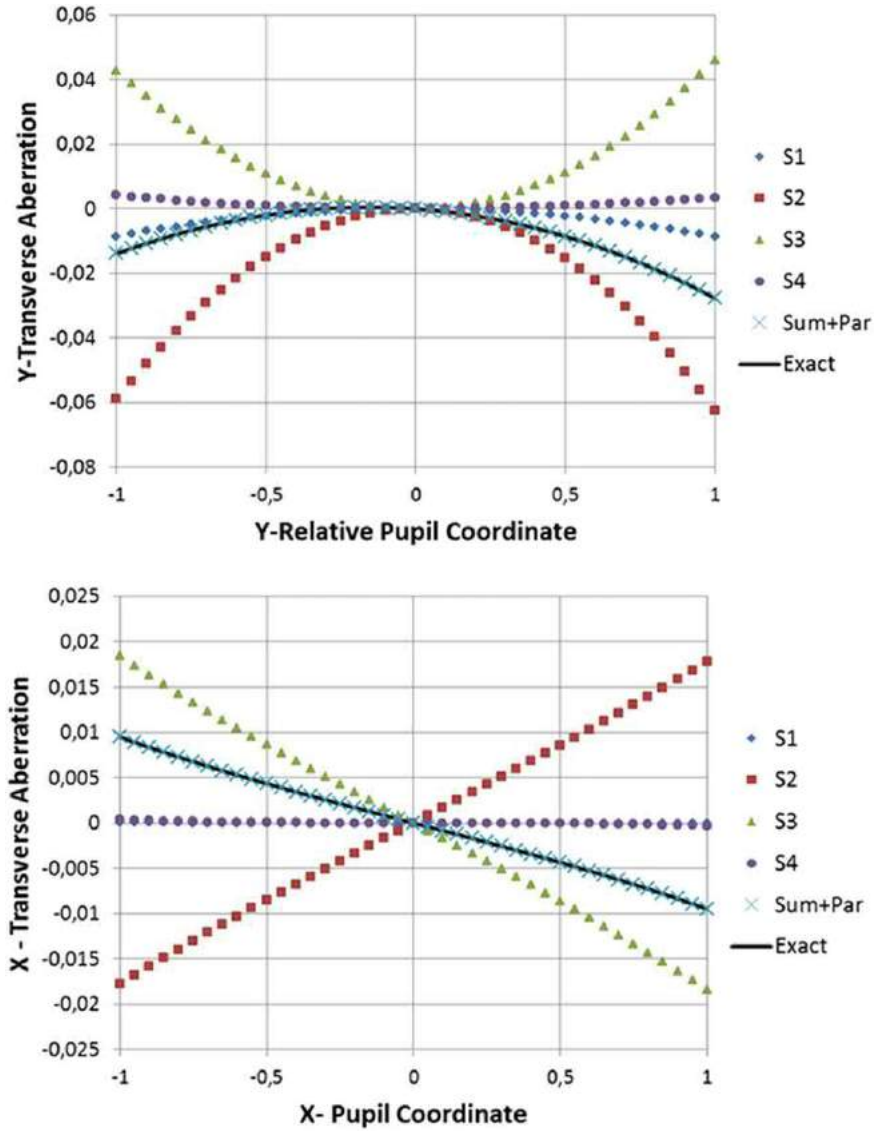


Fig. 11 Transverse aberrations for the (0°,0°) field, resolved for individual surface contributions. Reproduced from Chen, B., Herkommer, A.M., 2016. Generalized Aldis theorem for calculating aberration contributions in freeform systems. Optics Express 24 (23), 26999–27008.

As a first simple application the aberration contribution of a set of single rays can be studied. The chief-rays, representing different field angles, allow for a first useful application of the method. Since lateral aberrations of the chief-rays at the image plane correspond to the distortion of the system, we can simply apply our method to study distortion contributions in the freeform prism. To do so, we trace a set of chief-rays (i.e., rays starting at the center of the pupil) through the system. We use a set of 11×11 rays, scanning the field of view of $\pm 15^\circ$ in the x and y-direction. For every ray the ray-coordinates $\mathbf{r}_i$ and $\mathbf{r}_i'$ are recorded at all dummy surfaces. By applying Eq. (14) we can calculate the aberration contribution vector $\mathbf{\Delta}_I^i$ on the image plane. The spatial (x,y) components of this aberration vector correspond to lateral distortion, i.e., the distortion, which is induced by each of the surfaces. This is illustrated in **Fig. 10**.

In **Fig. 10(a)–(d)**, red circles represent the ideal positions of the rays on the image, calculated by the linear system matrixes, whereas the blue cross represents the ray aberration contributions of surfaces 1–4 on the image. Therefore the deviations of the blue cross from the red dots are the distortion contributions of each surface. The summation of all distortion contributions corresponds to the overall distortion at the image plane and are drawn as blue crosses in **Fig. 10(e)** and compared to the red circles directly obtained from real ray-tracing of the full system to the image in CodeV. As expected from Eq. (12) they exactly agree. The benefit of the phase space method is illustrated in **Fig. 10(f)**, where the individual surface contributions are shown for a single ray of this grid. Obviously surface S2 and S3 are strongly contributing to distortion.

As a further example of the application of this method we can calculate and illustrate transverse ray-aberrations for a particular spot in the image plane. In **Fig. 11** the meridional and saggital ray aberration contributions for each surface are illustrated for the central spot (0°,0°) of the freeform system. The figure reveals that here surface S2 and surface S3 are strongly compensating each other. The sum of all surfaces, plus the paraxial contribution, perfectly coincides with the results achieved by exact tracing of these rays to the image surface, which again proves the validity of Eq. (12).

Conclusions

This paper presents a matrix based method to visualize aberrations via an analysis of phase space transformations. The difference between real and paraxial rays define a general measure which can be employed to illustrate the aberration generation and propagation in freeform systems.

Moreover mathematical exact aberration contributions in freeform systems could be derived. It was mathematical proven, that the exact ray aberrations at the image plane is identical to the sum of individual surface contributions. Those individual surface contributions can be calculated form a ray-tracing analysis of the system and paraxial propagation of the real and paraxial differential vector induced from an individual surface. All calculations can be performed with standard ray-tracing, provided additional dummy surfaces are installed in the system before. No assumptions about symmetry were required. In other words the method resembles a kind of Aldis theory for freeform systems, allowing the exact calculation of surface aberration contributions.

We have illustrated the method at the example of a freeform prism and shown that distortion and transverse ray-fans can be calculated and agree with the result of a standard ray-tracer.

This general work thus allows an "easy to implement" visualization and analysis of the aberration contributions within any freeform system. This will be beneficial in comparing freeform designs, performing tolerance analysis and improved optimization strategies.

References

Brewer, S.H., 1976. Surface contribution algorithms for analysis and optimization. Journal of the Optical Society of America 66 (1), 8–13.
Chen, B., Herkommer, A.M., 2016. High order surface aberration contributions from phase space analysis of differential rays. Optics Express 24 (6), 5934–5945.
Cox, A., 1964. A System of Optical Design. Focal Press.
Fuerschbach, K., Jannick, P.R., Kevin, P.T., 2014. Theory of aberration fields for general optical systems with freeform surfaces. Optics Express 22, 26585–26606.
Gross, H. (Ed.), 2005. Handbook of Optical Systems, vols. 2 and 3. Wiley VCH.
Herkommer, A.M., 2013. Phase space optics: An alternate approach to freeform optical systems. Optical Engineering 53 (3), 031304.
Kloos, G., 2007. Matrix Methods for Optical Layout, vol. 77. SPIE Press.
Synopsis CodeV10.8 News. Available at: https://optics.synopsys.com/codev/codev-whatsnew.html
Takahashi, K., 1997. Head or face mounted image display apparatus. U.S. Patent No. 5,701,202.
Testorf, M., Hennelly, B., Ojeda-Castaneda, J., 2010. Phase-Space Optics. McGraw-Hill.
Thompson, K.P., Rolland, J.P., 2012. Freeform optical surfaces: A revolution in imaging optical design. Optics and Photonics News 23 (6), 30–35.
Torre, A., 2005. Linear Ray and Wave Optics in Phase Space. Amsterdam: Elsevier.

Silicon Photonics; Ring Modulator Transmitters

M. Ashkan Seyedi and Marco Fiorentino, Hewlett Packard Labs, Palo Alto, CA, United States

Introduction

A silicon-based photonics integrated circuit (PIC) is a device that integrates multiple photonics components in a single planar structure. Silicon PICs lag behind compound semiconductor PICs in maturity. However the prospect of building large numbers of complex PICs for data communication applications at low cost by leveraging the CMOS industrial base has attracted much attention in academia and industry. One of the key components of a data communication PIC is a modulator. An integrated optical modulator is used to vary the properties of a light beam propagating in a PIC. Most commonly, modulators change the beam intensity but phase modulators are also used. An electro-optical modulator is used to translate an electrical signal into an optical signal that can be then transmitted and received through a photonic circuit. A ring optical modulator consist of a ring-shaped optical cavity that is optically coupled with a bus waveguide. Schematic top views of ring resonators are shown in **Fig. 1**. In a silicon PIC, the waveguides have a rectangular cross section with dimensions ranging between hundreds of nanometers to a few microns and the ring diameters for modulators are of the order of ten microns, whereas larger rings with diameters up to hundreds of microns can be used as filters.

Fig. 1(a) shows the waveguide arrangement for an "all pass" resonator while **Fig. 1(b)** shows the arrangement for an "add-drop filter" resonator. In the case of the "all pass" arrangement, all light incident in the device is passed through except for the resonant wavelengths that are absorbed or scattered in the resonator. In the "add-drop filter" configuration, resonant light is dropped through the "drop" port while non resonant light is transmitted to the "pass" port. The transfer function of both configurations from the input to the pass port is characterized by a Lorentzian notch centered at the cavity resonant frequency. Changing the effective index of the ring cavity shifts the resonant frequency and this mechanism can be used to modulate the amplitude of a laser in a wavelength-selective way. This is the main attraction of ring modulators: rings of different resonant wavelengths can be cascaded on a bus waveguide to modulate a number of different laser wavelengths. Similarly resonant "add-drop" filters can be used to de-multiplex a multi-wavelength signal at the receiver. In a silicon-based modulator, the change in the effective index can be achieved in a number of ways. Most commonly one uses plasma dispersion effects that link changes in the carrier concentration to changes in the real and imaginary parts of the index of refraction of silicon. Changes in the refractive index Δn as well as in the absorption coefficient $\Delta\alpha$ at a wavelength of 1.55 μm are given by (Soref and Bennet, 1987).

$$\Delta n = -\left[8.8 \times 10^{-22} \times \Delta n_e + 8.5 \times 10^{-18} \times (\Delta n_h)^{0.8}\right] \tag{1}$$

$$\Delta\alpha = 8.5 \times 10^{-18} \times \Delta n_e + 6.0 \times 10^{-18} \times \Delta n_h \tag{2}$$

where Δn_e and Δn_h indicate changes in the free-electron and free-hole carrier concentrations, respectively. The number of free electrons and free holes in the silicon can be changed using charge depletion, charge injection, or charge accumulation.

Various modulator schemes are shown in **Fig. 2**. A horizontal depletion modulator is shown in **Fig. 2(a)**. The modulator consists of a light waveguide with a PN junction in the middle. By reverse-biasing the junction, one can expand the depletion region and therefore change the charge concentration. This modulation scheme is very fast and energy efficient but, because of the intrinsic doping of the light-carrying waveguide, it adds considerable optical losses to the device. It should also be noted that the maximum amount of index change is determined by the initial doping of the waveguide. A vertical depletion modulator (**Fig. 2(b)**) uses a similar modulation mechanism but the PN junction runs parallel to the plane of the circuit. This reduces the requirements for overlay control between the various doping steps. A current injection modulator consists of a PIN junction where the intrinsic region is the light carrying waveguide (**Fig. 2(c)**). This modulation scheme is relatively slow as it is limited by the majority carrier lifetime and has a relatively high power consumption due to the need to continuously

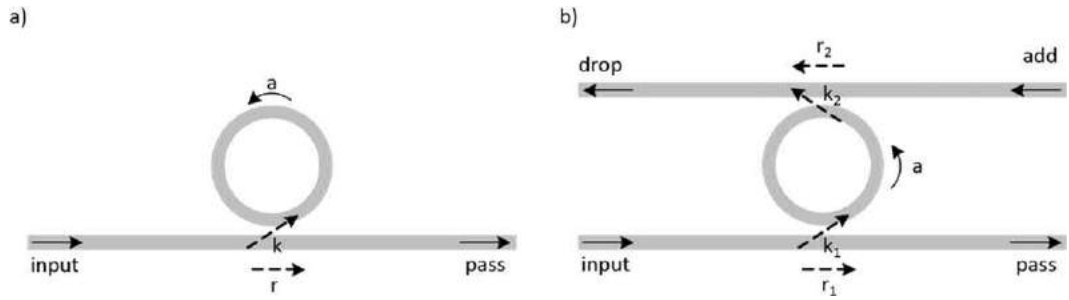


Fig. 1 Schematic top view of ring resonators in two configurations: (a) "all Pass" and (b) "add-drop filter".

 doi:10.1016/B978-0-12-803581-8.09762-9

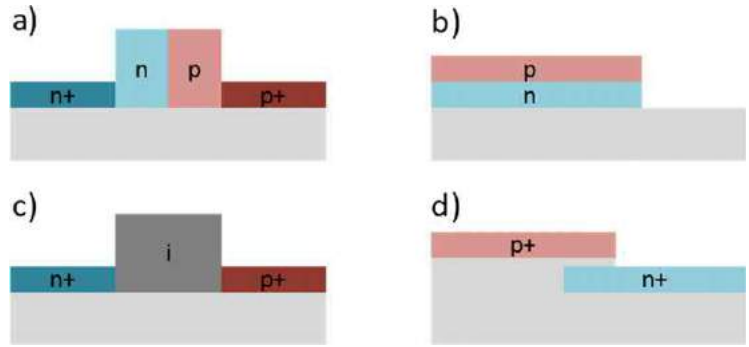


Fig. 2 Waveguide cross sections for various types of modulators: (a) horizontal depletion, (b) vertical depletion, (c) injection, and (d) accumulation.

replace carriers in the waveguide by injecting current in the junction. Balancing this disadvantages is the fact that large changes of the index are possible and the fact that the modulator can be built with low-loss intrinsic silicon. An accumulation modulator is shown in **Fig. 2(d)**. The modulator consist of a waveguide with a MOS capacitor in the middle (running parallel to the plane of the PIC in this case). The capacitor can be charged and discharged through the p and n electrodes and the charge accumulating around the oxide barrier creates the index change. This scheme has the advantage of being fast and can in principle be built with relatively low intrinsic losses. It should be noted, however, that while it is possible to build a ring or a disk modulator based on this modulation mechanism, none has been demonstrated so far. There are other possible modulation schemes. One can change the index of the waveguide by changing the temperature of the silicon. This method is slow and it is not used for data encoding but given that it allows a large tuning range, it is usually used to tune ring resonance wavelengths to a desired value. Other methods use an electro-optical effect in the material that clads the silicon waveguide, for example there have been several demonstration of polymer-based ring modulators.

Fabrication

Photonic devices and circuits require an optical isolation layer and thus are based on a silicon-on-insulator (SOI) wafer. In order to create such a wafer, a layer of silicon with a thickness of a few hundred nanometers is placed atop a layer of silicon dioxide that is a few microns thick. Below this buried oxide is a bulk silicon substrate. The index contrast provided by the oxide layer below and air or a deposited oxide above enables optical confinement in the upper, thin, silicon layer in the vertical direction (Bruel, 1995).

The fabrication of ring-based modulators uses basic CMOS foundry process steps such as implantation of ions to form junctions, vias for ohmic contact, and a multi-layer back end of line. However, to enable lateral optical confinement, waveguides are formed by a partial etch of the silicon layer using plasma gas dry etch processes, a key step that enabled adoption of photonics into CMOS processes. In the most advanced foundries, 193 nm immersion lithography is used to define the optical structures in a resist, but lower resolution lithography has also been used successfully. For ring-based fabrication the critical dimension is the gap between the ring and the bus waveguide that can be as small as 100nm and requires tight control to avoid large dispersion in the optical properties. The lithographic pattern is then transferred into an oxide-based hard mask and finally into the SOI layer using a plasma etch. The optical propagation loss is a critical metric of performance for PICs and is highly dependent on the quality of the lithography and ensuing etch steps to ensure smooth, straight sidewalls. The depth of the partial etch (or alternatively, the height of the waveguide side wall) determines this loss value as it exposes more roughness to the optical signal. Therefore, this depth and its uniformity across a wafer is a tightly-controlled process parameter.

Often, there are two partial etches of the SOI wafer: one to define optical grating structures used for input/output of the optical signal onto the chip and the second for waveguides, often referred to as a shallow waveguide etch. The advantage of this shallow etch, which is usually 50% of the top silicon layer of the SOI wafer, is that it has low propagation loss on the order a few dB per centimeter for a straight waveguide. However, as mentioned, rings are formed by tight bends of the waveguide where the performance of shallow etched waveguides is very poor. This is primarily due to low optical confinement that causes the optical field to leak out of the waveguide through radiative modes. Bend losses can be alleviated by a deeper partial etch, resulting in a higher waveguide sidewall. Of course, this enhanced optical confinement comes at the expense of increased optical propagation loss. Therefore, these competing mechanisms present an optimization problem between process controls, etch depth and ring resonator bend radius in order to achieve optimized optical performance. A cross section of a partially etched waveguide is shown in **Fig. 3** where important dimensions are denoted. A further complication of ring modulators that use a lateral junction is the resistance of the remaining portion of the un-etched SOI layer, referred to as the rib. A deeper etch results in better confinement which enables a tighter bend radius. However, increased optical loss and higher resistance of the rib layer present a design trade-off that must be properly accounted for to optimize device performance.

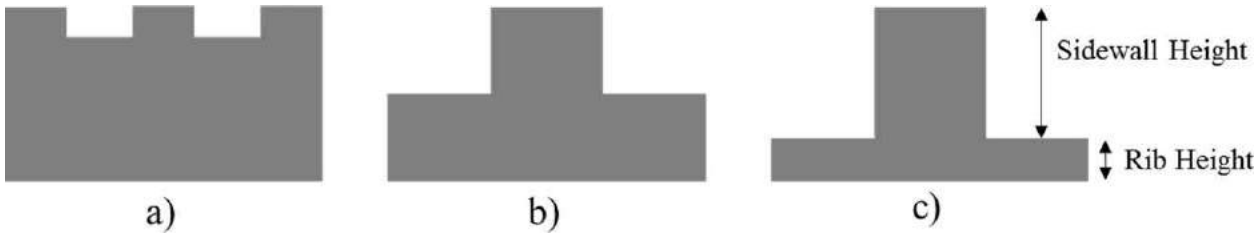


Fig. 3 Cross section of various etch depths for (a) TE-polarized gratings, (b) shallow rib waveguide, and (c) deep rib waveguide.

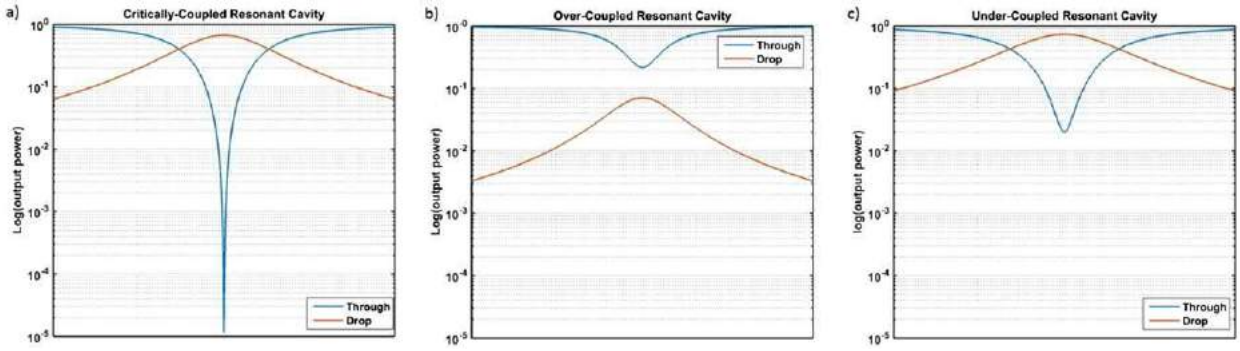


Fig. 4 Through and Drop port spectra for ring resonator in (a) critically coupled, (b) over-coupled, and (c) under coupled conditions.

Optoelectronic Properties

As mentioned before, the transfer function of a ring resonator structure is a Lorentzian, characterized by its extinction ratio (ER) and full-width half-maximum (FWHM), a parameter which is often used to calculate the cavity quality factor (Q) by an inverse relation. A resonant cavity can be under, critically, or over-coupled which is easily determined by the trends of the ER and cavity Q. Following the treatment in Bogaerts *et al.*, (2012), the transmitted optical spectrum for an all-pass configuration, as shown in **Fig. 1 (a)**) is given by the following equation:

$$T_n = \frac{I_{pass}}{I_{input}} = \frac{a^2 - 2ra\cos\phi + r^2}{1 - 2ar\cos\phi + (ra)^2} \tag{3}$$

where $a^2 = \exp(-\alpha L)$ is the internal loss mechanisms and comprises radiative losses due to surface roughness, bend loss, and free-carrier absorption from junction doping. These factors are largely process-dependent and can be considered static for a ring resonator design. However, careful consideration must be given to the coupler region design to ensure a proper balance between these three factors to achieve the desired coupling state. The coefficient r quantifies the coupling between the ring and the bus waveguide. Similarly, the transmitted optical power for the through and drop port for a ring resonator in the add/drop configuration shown in **Fig. 1(b)**) is given by

$$T_p = \frac{I_{pass}}{I_{input}} = \frac{r_2^2 a^2 - 2r_1 r_2 a\cos\phi + r_1^2}{1 - 2r_1 r_2 a\cos\phi + (r_1 r_2 a)^2} \tag{4}$$

$$T_d = \frac{I_{drop}}{I_{input}} = \frac{\left(1 - r_1^2\right)\left(1 - r_2^2\right)a}{1 - 2r_1 r_2 a\cos\phi + (r_1 r_2 a)^2} \tag{5}$$

where r_1 and r_2 are the coupling coefficients between the ring and the add and drop waveguide, respectively. The state of coupling to the ring resonator is a relation between the input coupling, internal loss, and output coupling. Referring to the add/drop configuration in **Fig. 1(b)**, the relationship

$$r_1 = r_2 a \tag{6}$$

defines the critical coupling condition. When $r_1 > r_2 a$ the cavity is under-coupled whereas when $r_1 < r_2 a$ the cavity is over-coupled.

In the critically-coupled regime the amount of light coupled to the cavity is equal to the amount coupled out and radiated due to losses. In this condition the extinction ratio (ER) and quality factor (Q), as can be seen from the spectra of the through and drop port plotted in **Fig. 4(a)**, are very high. For SOI-based ring resonator cavities, a cavity Q in excess of 15,000 is routinely possible with greater than 15 dB ER. As the resonator cavity is moved into the under-coupled regime, the ER drops drastically as Q increases still, as can be seen from the plotted spectra in **Fig. 4(b)**). An over-coupled resonator has both low ER and low Q. the spectra of which are plotted in **Fig. 4(c)**). Therefore, ring resonator amplitude modulators that are critically coupled are desired as this

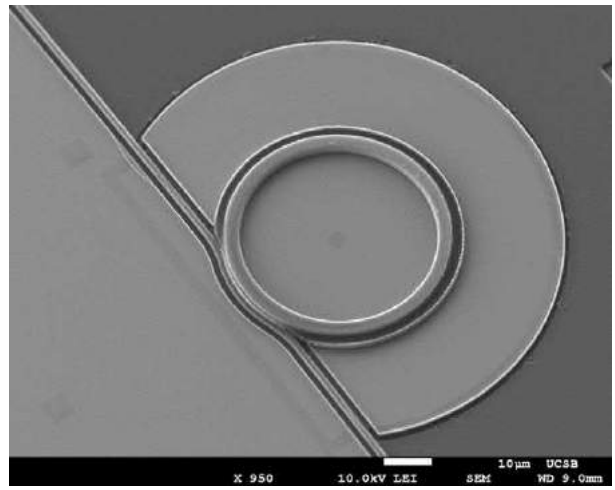


Fig. 5 Curved bus/ring resonator cavity coupler region.

maximizes the ER of the optical signal while keeping the cavity Q within a reasonable range. The cavity Q presents yet another design optimization as higher Q allows for denser channel spacing due to higher spectral isolation. However, higher Q means long photon lifetime within the optical cavity, increasing inter-symbol interference (ISI) as the modulation data rate increases.

There are many design types for coupler regions. In one example two parallel waveguides of a desired length are used to transfer power between the bus waveguide and ring cavity. This design allows for an analytical solutions for coupling strength using coupled mode theory (Marcuse, 2013). However, as is often desired for critical coupling, low input/output coupling ($\sim$5% or less) are difficult to achieve due to the required short interaction length. Another design type of the coupler region is a co-curved bus/ring interaction region, shown in **Fig. 5** where the bus waveguide follows the curve of the ring cavity with a constant gap separation.

This design offers a great degree of flexibility and allows for low coupling percentages between the bus and ring waveguides and is often preferred for processes where the minimum gap possible between two waveguides is large (on the order of 250 nm or greater). This design allows for shorter effective interaction lengths between the two waveguides as compared to the parallel waveguide design, it allows the resonant cavity to maintain a fully circular shape (which is often required for maintaining a constant loss and mode profile inside the cavity). However, it is more prone to fabrication variation in the gap between the ring/bus waveguide. The third design technique for the coupler region uses a straight bus waveguide and circular ring cavity as shown in **Fig. 1**, but with a narrower waveguide width for the coupler (W_{cp}) versus the ring cavity. In using a narrower waveguide width for the bus, the effective indices for the fundamental states of the two waveguides are increased, thereby enhancing the coupling between the two fundamental waveguide modes for a fixed effective interaction length. Another design factor for the add/drop configuration of the ring resonator cavity are the gaps between the ring cavity and the through and drop gap, referred to as G_{thru} and G_{drop}, respectively. For a fixed W_{cp} and ring width, one can control precisely the coupling percentage between the two waveguides by narrow variations in the gap between the two waveguides. This allows for accurate control of the input/output coupling to achieve critical coupling, as defined above.

A ring resonant cavity can be used as an amplitude modulator when its wavelength of resonance is modulated by electro-optic interactions. However, it is difficult, if not impossible, to control the absolute value of the resonance wavelength by fabrication techniques alone. The resonance wavelength can be controlled by changing the effective index of the material by using either the electro-optic effect or localized heating using integrated heaters. As mentioned, injecting carriers into the ring cavity causes a blue shift in the resonance wavelength and this is the main scheme for modulation. However, by maintaining a small DC bias on the diode, the resonance wavelength can be controlled with the same diode that is performing the modulation. This scheme is challenging to implement along with high-speed modulation in the driver design. By using a localized heater, the effective index of silicon can be finely tuned causing a red shift in the resonance wavelength.

The two main designs for an integrated heater are using the lowest metal layer directly above the ring cavity or with a resistive heater in the silicon rib material. To maximize the thermal profile overlap, the metal layer traces over the ring rib waveguide for most of the circumference of the ring cavity. This design is more robust to fabrication errors and burn-out from high current. However, due to the low resistivity of the typical metal layers available in CMOS-compatible processes, this design requires higher voltages as compared to other designs. Furthermore, due to CMOS back-end-of-line (BEOL) processes, the metal layer is several hundred microns away, further decreasing its thermal impact on the underlying silicon waveguide. Decreasing this length is not directly an advantage as interaction of the optical field with metal can lead to very high losses. These challenges can be somewhat mitigated by using highly resistive compounds like TiN instead of copper or aluminum, however, this material is not compatible with CMOS processes.

An integrated resistive heater in the silicon rib is created by forming a conductive channel in the otherwise intrinsic silicon material. Ohmic contacts are formed to this region by forming regions of high doping and conventional BEOL steps. By using an integrated resistor in the silicon rib, the local temperature within a few microns of the ring rib can be controlled, which in turn controls the material effective index, allowing for accurate resonance wavelength control. This design offers very efficient wavelength control and consumes power on the order of tens of microwatts per gigahertz in the wavelength range of interest. Careful consideration must be given to the resistor design in doping type and concentration of the current channel, its length and width to avoid leakage/fringing fields, and to optimize the resistance to achieve efficient wavelength control. A drawback of this design is the minimum proximity of the current channel to the waveguide due to foundry design rules. By using local temperature change, often near the coupling regions of the cavity and bus waveguides, a relatively small footprint of a few microns squared is required to efficiently control the wavelength of resonance, furthering the advantage of this heater design.

As mentioned previously, injection-type ring modulators use a P-i-N junction inside the waveguide to modulate the optical signal. The intrinsic bandwidth of this diode, however, is typically ~1 GHz. This is due to long lifetime of carriers in the intrinsic region. To overcome this bandwidth limit, equalization, or pre-emphasis, of the driving signal is required. The general principle of this approach is to decrease the amplitude of the lower frequency content of the driving signal, thereby boosting the amplitude of the high frequency content of the driving signal, relatively. The convolution of this driving signal frequency content and the intrinsic electrical frequency response of the diode results in increased bandwidth. This technique has achieved open eye diagrams up to 25 Gb/s using carrier-injection ring modulators with in intrinsic bandwidth of 1 GHz (Wu, *et al.*, 2016). As mentioned previously, there exists also an intrinsic bandwidth of the modulator from the photon lifetime due to cavity Q. So while pre-emphasis may boost the electrical bandwidth, the photon lifetime is a consequence of optical cavity design and is not affected by a pre-emphasized electrical signal.

In order to first understand the intrinsic electrical bandwidth of a ring modulator, an accurate circuit model is needed. The BEOL metal contribute small but known parasitic effects in the form of resistance and capacitance due to metal vias, substrate and metal layers, and are independent from the type of ring modulator, namely carrier depletion or injection. To develop a large signal model, the P-i-N junction may be modeled as a current source and a capacitor in parallel which are placed in series with a resistor, as shown in **Fig. 6(a)** fromWu *et al.* (2016). Similarly, a carrier-depletion ring modulator can be analyzed by the effective circuit shown in **Fig. 6(a)** fromRhim *et al.* (2015).

By performing S_{11} electrical measurements, the parasitics from these circuits can be derived through fitting. Subsequent DC current-voltage tests will further define the diode's current behavior. Note that the optical bandwidth due to the photon lifetime effect is also captured in the carrier-injection model. It is critical to note that the capacitor in this model, represented by C_{diff} is not a representation of the junction capacitance but that it is a mathematical consequence of the diode's small-signal behavior. Namely, it captures the transit time of the carriers across the junction and represents the change in current as a function of a bias change in time, which is the definition of capacitance. Therefore, C_{diff} must be carefully calculated using physical parameters and fitting models to accurately predict the time-dependent behavior of the modulator. Once the frequency response of the diode current is known, it can be mapped directly to a change in material index and thus modulation of the optical signal. This allows for an opto-electrical co-simulation to understand the high-speed modulation response of the modulator and allows for optimization of the pre-emphasized driving scheme often required to overcome the bandwidth limitations of a carrier-injection modulator.

The pre-emphasized signal can be generated by a few approaches and two will be briefly discussed here. By using a differential PRBS pattern, one can include a tunable delay and DC bias on the complementary output of the differential pair and combine the two outputs. This forms then the over/under-shoot components of the driving signal. The amount of delay relative to the data rate UI determines the frequency boost seen by the higher frequency components of the driving signal. Analogously, using a multi-tap FIR filter can produce a similar scheme. Often 2 or 3 taps are sufficient to produce the correct frequency response necessary for this driving mechanism. By using external hardware, one can accurately understand the required frequency response, making the design of a CMOS driver ASIC more direct. Once this frequency response is known, a tunable FIR filter integrated into the CMOS driver design can be easily implemented. Both approaches have been demonstrated (Li *et al.*, 2014) successfully to drive carrier-injection ring modulators.

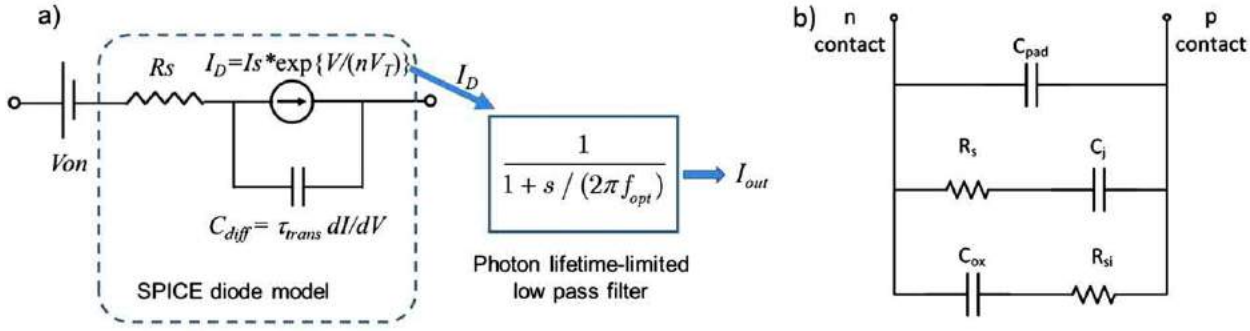


Fig. 6 Equivalent circuit for a (a) carrier-injection and (b) carrier-depletion ring modulator.

Nonlinear Effects

Light that is tightly confined in a silicon photonic waveguide will experience optical nonlinearities. Whereas silicon is a centrosymmetric material and therefore does not exhibit any nonlinear effect of the lowest order ($\chi^{(2)}$) third-order ($\chi^{(3)}$) effects are going to be present. The resonant nature of the ring resonators further reduces the threshold power necessary to observe nonlinear effects. Some researchers have strived to demonstrate and exploit these effects for a range of applications including Raman amplification and lasing, self-phase modulation, all-optical processing, and even photon-pair generation. In the context of modulation, however, nonlinear effects are detrimental. In particular the most commonly observed nonlinear effect in ring resonators derives from thermal effects. Light absorption in the waveguide (either from two-photon absorption or from absorption by silicon defects) creates carriers that thermalize thus increasing the waveguide temperature. Because the index of refraction depends on temperature this effect will change the resonator resonant wavelength and therefore its transfer functions. The effect of this nonlinear behavior is shown in **Fig. 7**. As the power increases, the transfer function of the drop port of an "add-drop filter" changes (**Fig. 7(a)**). The system also shows a bi-stable behavior as evidenced by the difference in the transfer function measured in the same resonator by scanning the wavelength in the two directions. Initial modeling of this effects for high-speed depletion modulators has been published (Shin *et al.*, 2016).

Applications: Ring-Based DWDM Transceivers

The most important application of silicon microring modulators is in dense wavelength multiplexing (DWDM) transceivers. Because of their wavelength-selective nature microrings naturally lend themselves to DWDM applications. A schematic view of a DWDM transceiver is shown in **Fig. 8**. In this example a multi-wavelength laser source is injected in the transmitter circuit comprising a number of modulator rings each resonant with one wavelength of the incoming source. The modulators are used to encode a data stream on the individual laser lines. At the receiver side add-drop resonators are used de-multiplex the incoming data streams and direct them on individual detectors. Waveguides and possibly optical fibers connect the transmitter and the receiver. Transceivers like the one described here have been studied but no commercial implementation of this scheme is available at the time of writing. A key issue to warrant widespread deployment of ring-based DWDM transceivers is the implementation of

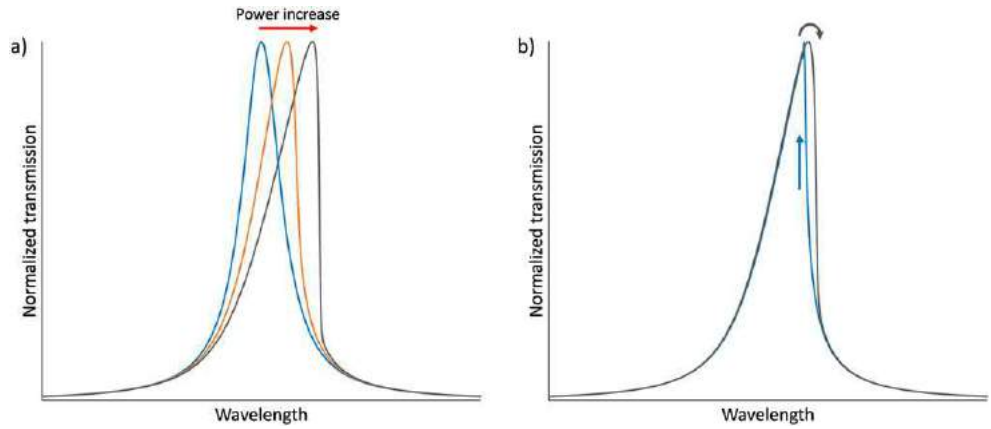


Fig. 7 Nonlinear bistability in ring resonators. (a) Change in the normalized transfer function of a drop port as a function of optical power. (b) Bistability, the transfer function changes depending on the direction of the wavelength sweep.

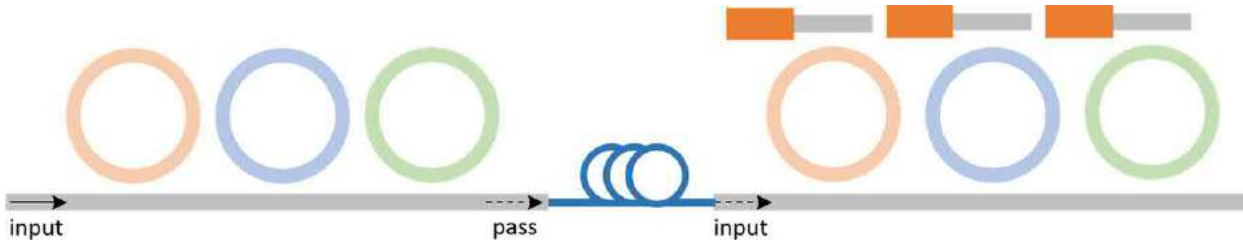


Fig. 8 Dense wavelength division multiplexing link. A multi-frequency source is injected in the transmitter and the various components are separately modulated. After transmission involving on-chip waveguides and off-chip fibers the modulated signals are coupled into a receiver chip where the various components are demultiplexed using drop filers and detected with on-chip photodiodes.

appropriate electronic circuits that drive the modulators, amplify and discriminate the received signal, and tune the rings to be in resonance with the appropriate laser line. Much work has gone into the design of special purpose circuits to achieve these goals. Most of these circuits have been designed in standard CMOS processes to take advantage of the large industrial base that supports these processes and given the relatively low analog bandwidth (10–20 GHz) required for these circuits.

Transmitter Circuits

Transmitter circuits differ considerably depending on the nature of the ring modulator. Depletion rings behave as capacitive loads and have bandwidths in excess of 20 GHz. For this reason relatively simple drivers comprising a chain of inverters and a final driver can be used. With such drivers bit rates in excess of 50 Gbps have been demonstrated. The bandwidths of injection rings on the other hand is severely limited by the carrier recombination time and typically are of the order of 1 GHz or less. For this reason injection-ring drivers require a significant amount of pre-emphasis and relatively large driving voltages. To achieve this pre-emphasis or finite impulse response filter stages need to be added to the driver circuitry. These additional circuits increase the complexity and power consumption of the drivers. With these caveats injection-ring modulator drives have been demonstrated for bit rates in excess of 20 Gbps.

Receiver Circuits

The receiver circuits for microring links do not present any peculiarity compared to standard photonic receivers. The main blocks of a receiver circuit include a trans-impedance amplifier, optional additional amplification stages, discrimination stages, and clock data recovery. The fact that the Silicon Photonic circuit can be closely integrated with the receiver circuit either through 3D integration or monolithic integration simplifies the design of the receiver first stage. The receiver sometimes is integrated with the control circuitry necessary to keep the drop filter in resonance with the incoming modulated laser.

Controllers

Controller circuitry is used to keep the transmitter modulator and drop-off filter in the receiver on resonance with the laser line. These circuits require a sensor and an actuator. For the sensing element a photodiode is often used to verify that the peak level of the modulated light is optimized. This can be accomplished using a peak detector in the receiver circuit. On the transmitter side a similar approach can be used if an add-drop configuration is used for the modulator ring. Two actuator strategies are commonly used with microrings. Increasing a ring temperature allows one to red-shift the resonant frequency. Charge injection allows one to blue-shift the resonance. Charge injection is more energy efficient than thermal tuning but the tuning range is limited. Ideally a combination of the two methods should be used in order to optimize power consumption.

Laser Sources

A multi-wavelength laser source is a key element of the DWDM link described earlier. Such source can be realized by combining multiple lasers through an AWG (array waveguide grating) or similar multiplexing component. Intrinsically multi-wavelength sources can also be used. Comb lasers based on quantum dot gain materials provide multiple independent low-noise lines that can be used for modulation. This solution is much simpler and less expensive than using multiplexed individual lasers.

Link-Level Considerations

Not much can be said about link-level tradeoffs here. In general a system-level analysis of the link is necessary to achieve optimal performance whereas optimizing a single component or figure of merit (such as the line toggle rate) will most likely result in sub-optimal link performance. It is also important to consider the compatibility and availability of different technologies in a single platform.

References

Bogaerts, W., *et al.*, 2012. Silicon microring resonators. Laser & Photonics Reviews 6 (1), 47–73.
Bruel, M., 1995. Silicon on insulator material technology. Electronics Letters 31 (14), 1201–1202.
Li, C., *et al.*, 2014. Silicon photonic transceiver circuits with microring resonator bias-based wavelength stabilization in 65 nm CMOS. IEEE Journal of Solid-State Circuits 49 (6), 1419–1436.

Marcuse, D., 2013. Theory of dielectric optical waveguides, 2nd ed London: Elsevier.
Rhim, J., *et al*., 2015. Verilog-A behavioral model for resonance-modulated silicon micro-ring modulator. Opt. Express 23, 8762–8772.
Shin, M., *et al*., 2016. Parametric characterization of self-heating in depletion-type Si micro-ring modulators. IEEE Journal of Selected Topics in Quantum Electronics 22 (6), 116–122.
Soref, R., Bennet, B., 1987. Electroopticaleffects in silicon. IEEE J. Quant. Electron. 123–129.
Wu, R., *et al*., 2016. Large-signal model for small-size high-speed carrier-injection silicon microring modulator. New Orleans, LA: OSA Integrated Photonics Research, p. IW1B.4.

Optical Switches

Dritan Celo, Dominic J Goodwill, and Eric Bernier, Huawei Technologies Canada, Kanata, ON, Canada

Introduction

An optical switch is a networking element that directs light from optical input fiber(s) to optical output fiber(s) without converting the light to an electronic signal. An optical switch is controlled by an electronic controller. Applications for optical switches include DWDM network reconfiguration, test automation and packet networking.

Evolution From Static Links to Optical Switching

An optical fiber link carries information from a transmitter to a receiver. Early optical fiber network deployments comprised static optical links, of three general types. Wavelength add-drop networks are used for metro and long-haul deployments, and comprise wavelength add-drop rings and multi-wavelength point-to-point links. Passive optical networks (PON) have a root and leaf topology with time-division multiplexing and, in some cases, wavelength filtering. Intra-datacenter, LAN and super-computer deployments comprise single-point to single-point optical links. In all three environments, all of the optical paths were fixed when the equipment was installed. Switching of circuits, flows or packets happens in electronic elements positioned between an optical receiver and an optical transmitter, in an architecture known as optical-to-electrical-to-optical (OEO), Fig. 1. Even today this remains the conventional way to switch information. However, such OEO conversion uses transceivers at the switch, which are expensive, bulky, energy-intensive, and whose design usually supports a small set of bit-rates and protocols.

It is also possible to reconfigure or switch the optical links. This capability is known generically as optical switching. In optical switching, an electronic controller actuates a photonic element to change the connection of light between the optical transmitters and optical receivers. The architecture is known as optical-optical-optical (OOO), where the center "O" represents the optical switch itself. As will be described below, in some cases the photonic element is a switch that is positioned part-way along the optical path, and in other cases some aspect of the transmitter and/or receiver is actuated (e.g., the wavelength is tuned) which results in a change in the optical path.

An optical switch is transparent to signaling format and protocol (within limitations of spectral bandwidth, loss budget and other physical optical effects) and therefore simplifies the hardware requirements in the data plane. Because there is no electronic conversion of the signals, it is believed that this method of switching will consume less energy, an important requirement to maintain growth of data centers and super computers. Therefore, upgrades of bit-rate or protocol can be accommodated without the need to replace the switch.

Terminology of Optical Switching

The term used in this article is "optical switching". The term "photonic switching" is also used for this capability, as it is the destination of the photons which is changed.

The term "all-optical switching" is sometimes used, but this term is ambiguous. For many years, in research labs, there have been elements where a control optical beam directly affects the destination of an optical signal, without the intervention of any transistors – this is truly an "all-optical" switch. It is out of scope for this article, as its practical use in communications systems is rare to date.

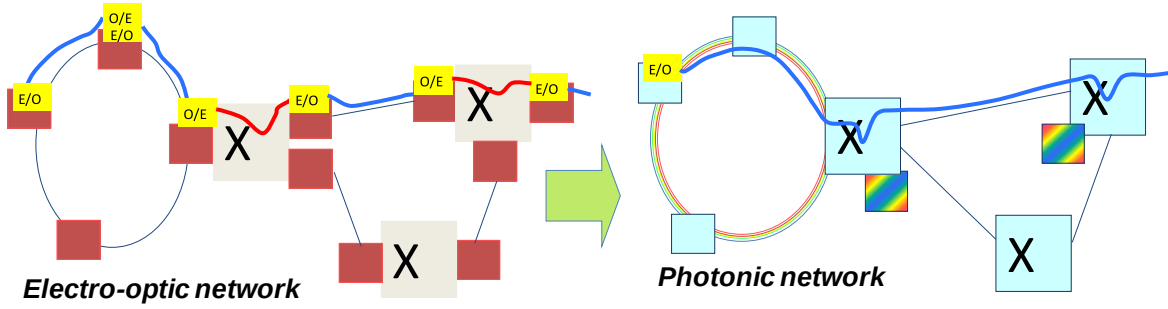


Fig. 1 A conceptual electrically switched optical-to-electrical-to-optical (OEO) network, compared to an optically switched optical-optical-optical (OOO) network.

 doi:10.1016/B978-0-12-803581-8.09769-1

Existing and Proposed Deployments of Optical Switches

Protection and provisioning

Optical switching is already widely deployed in metro and long-haul networks. The switching granularity is either a fiber, typically using 1×2 and 2×1 switches for protection switching when a fiber fails, or an individual wavelength, typically for adding new customers or adding additional bandwidth to a network, using reconfigurable optical add/drop (ROADM) and wavelength selective switch (WSS) technology. In current deployments of wavelength-granularity systems, the actual switching happens very rarely, although dynamic reconfiguration is frequently proposed as an important future opportunity. Switching time of milliseconds is typical within the switch itself, but the overall optical link takes 10's of seconds to several minutes to achieve stability in existing production networks.

Patch panels

Optical switching with fiber granularity is used in lawful intercept (fiber taps with selector switch), video production and test automation applications. Here, switches act as automated patch panels. Example technologies are: free-space MEMS to redirect beams from an input fiber collimator to a focusing lens of an output fiber, piezoelectric mechanical actuators to move a fiber relative to collimating and focusing lenses, and robot arms that move fiber patch cords. Switching time of milliseconds to seconds is typical.

Reconfiguration within data center

Optical switching with fiber granularity is currently being promoted for data center use where it can provide reconfiguration as workloads change, and can mitigate fiber wiring errors by craft people. Switching time of milliseconds is typical. However, there is a lack of published examples of real deployments.

Packet switch capacity growth

Optical switching with very fine granularity is a subject of much ongoing research, with the intent of capping or displacing OEO flow switches and packet switches, although there are currently no published commercial deployments. Switching time of nanoseconds to microseconds is typical. In this application, the energy-per-bit is a key metric, and photonic switching can in principle be more energy efficient than electrical switching. For clarity, optical switching applications are summarized in **Table 1**.

Switch Characteristics

Granularity

The granularity of a switch is the smallest change that it can make. Example granularities of optical switches are fiber, wavelength, flow, and packet (ranked from coarsest granularity to finest granularity). As a rule of thumb, optical switches with coarser granularity are simpler to implement, easier to integrate with network operations, have better optical performance, are cheaper, and are more widely deployed than optical switches with finer granularity. Nonetheless, optics is penetrating further and further into electronic switches, and the opportunity for fine granularity optical switches is growing.

Scalability

The scalability of a switch is a conceptual term that indicates whether the technology and architecture of a switch can implement a large number of ports, without a catastrophic reduction in the quantitative parameters, compared to a switch with a small number of ports.

Quantitative Performance

Optical switches are commonly characterized by the following quantitative performance parameters.

Blocking probability is the fraction of legal connection maps that cannot be met because there is no available path inside the switch. Illegal connection maps are not included – for example, a map where two inputs are supposed to go to the same output is illegal, because this can never be serviced in a bufferless switch.

Table 1 Applications of optical switches

Application	*Switching time of the switch element itself*	*Commercial status*
DWDM network provisioning	Milliseconds to seconds	Widely deployed
Protection	Milliseconds	Deployed
Patch panel	Milliseconds to seconds	Deployed
Reconfiguration within data center	Microseconds to milliseconds	Emerging
Packet switch capacity growth	Nanoseconds	Development

Reconfigurability is related to blocking – it indicates whether or not existing connections have to be temporarily broken and reconnected when a new connection request arrives.

Switching time is the time between the switch receiving a request for a new connection, and the new lightpath is established and meets all the technical parameters.

Insertion loss (IL) is the fraction of optical signal power lost from the input to the output of the switch, typically expressed in dB.

Path dependent loss is the uniformity of insertion loss over different paths, typically expressed in dB. Optical receivers have finite dynamic range, and the noise penalty from optical amplifiers depends on the optical power level into the amplifier. Hence it is desirable if the path dependent loss is small.

Optical crosstalk is (in its simplest form) the ratio of the power at a specific output port from the desired input port, to the total power leaking from all the other inputs into that output port. A more realistic optical crosstalk calculation has a complicated dependence on the relative wavelengths of the input signals, coherence length and other link engineering parameters.

Multipath interference (MPI) is the fraction of optical power which takes a different path through the switch than the desired path. It arises from multiple back-reflections and leakages within the switch.

Extinction ratio (ER) is the ratio of the output power in the on-state to the output power in the off-state. Typical ER values are 40–50 dB for opto-mechanical switches, while the ER for photonic integrated circuit switches varies widely depending on the technology and architecture with reported values from 10 dB to 60 dB.

Polarization-dependent loss (PDL) is the range of insertion loss for different polarizations of the input light.

Differential group delay (DGD) is the range of time delay through the switch, for different polarizations of the input light. DGD causes deterioration of the optical signal quality, especially when large DGD is combined with large PDL as this is a non-linear effect that cannot be compensated using digital signal processing at the receiver.

Spectral flatness is the variation in the other technical parameters over the overall operating wavelength range of the switch, while ripple is the variation within the pass-band of a given channel. In the case of switches that are not wavelength selective (e.g., Mach-Zehnder switches, MEMS space switches), ripple and spectral flatness are equivalent terms. In the case of switches that are wavelength selective (e.g., WSS or ring-resonator switches), ripple indicates how much an individual optical signal will be corrupted, whereas spectral flatness indicates how the major parameters (insertion loss, PDL, DGD and so on) vary from one wavelength channel to another wavelength channel.

Engineering considerations of size, power consumption, cost, reliability and the ability to monitor the switch are also important parameters.

As a general rule, switches with small optical penalty are preferred, although to an extent the insertion loss and other penalties can be compensated using optical amplifiers and improved transmitters and receivers. The performance requirements vary greatly depending on what services are passing through the switch and the optical link budget from transmitters to receivers. Hence the quantitative requirements on an optical switch are specific to the network in which it is deployed.

Families of Optical Switch Technologies

Table 2 lists the families of optical switching technology that are most commonly used today. Key technologies are described in more detail later in this article.

Table 2 Optical switch technologies

Family	*Propagation medium*	*Mechanism*	*Types of actuator*
Free-space optical	Gas	Angle of the beam is steered by actuators	Tilting MEMS mirror Diffractive beam steerer comprising MEMS or liquid crystal Piezoelectric motor moves fiber relative to lens
Fiber-optic	Fiber	Fiber is physically moved by actuators	Mechanical motor moves fiber
Planar lightwave circuit	Optical waveguides on a chip	Actuators on the chip determine which waveguides the light passes through	Variable interferometer using thermo-optic, carrier injection, or electro-optic effects. Typical materials: silicon, silicon oxide/nitride, III-V semiconductor. Mach-Zehnder and microring interferometers are common. Variable gain in III-V semiconductor amplifier. Waveguide coupler, moved by MEMS.
Tunable laser	Static network of DWDM filters	Actuator in the laser determines wavelength, which determines the route through the filters	Thermo-optic, carrier injection, or movable mirror, in III-V semiconductor laser

Key Application Today: Reconfigurable Optical Add Drop Multiplexer, Using Wavelength Selective Switch

A reconfigurable optical add-drop multiplexer (ROADM) is an optical switch system with wavelength granularity, often used in access ring networks, and in rings and mesh networks that cover cities or larger regions. ROADMs can provide add/drop functions on rings, and cross-connection at hub locations. Initially, ROADMs were controlled by network provisioning software. Currently, many ROADM vendors offer the capability to integrate the ROADM controller into a software defined network.

At the heart of most ROADMs that are deployed today is one or more wavelength selective switch (WSS). A WSS has optical fiber inputs and outputs that each carries multiple wavelengths. The WSS selects which wavelengths from a given input go to a given output. Each output fiber can only carry one signal of a given wavelength. The WSS cannot change the color of the light. The WSS may also provide balancing of optical power across the wavelengths on an output fiber, and monitoring of optical power in each wavelength.

The fibers connected to a WSS are intended to carry multiple wavelengths, and thus network fibers connect directly to the WSS (typically via an optical amplifier to compensate for insertion loss and fiber loss). However, WSS are also used to connect from transceivers to/from the network fibers. A transceiver performs electro-optic conversion on only one wavelength at a time. Therefore between the WSS and optical transceivers there is an optical de/MUX, optical splitter/combiner, and/or an optical cross-connect to separate and direct the individual wavelengths. If the ROADM is used with coherent transceivers, then an external deMUX is not needed, because a coherent optical receiver contains a built-in wavelength filtering function achieved by its local oscillator laser. However, if the ROADM is used with direct detect transceivers, then an external deMUX is needed, because the receiver cannot distinguish light of different wavelengths.

Fig. 2 shows architectural differences between three types of ROADM: classic ROADM, colorless directionless (CD) ROADM, and colorless-directionless-contentionless (CDC) ROADM. The CDC ROADMs offer greater flexibility, but are more complex. For illustration purposes, the classic ROADM (which is colored, has direction, has contention) is shown with direct detect transceivers and therefore has external deMUX, whereas the CDC ROADM is shown with coherent transceivers and thus does not need an external deMUX.

ROADMs have the following attributes:

Degree

The degree is the number of network fiber pairs that connect to a ROADM.

Colorless

The ROADM is colorless if the wavelength that can be added at a given add fiber, or dropped at a given drop fiber, can be changed. Early ROADMs had a fixed wavelength at each add fiber and each drop fiber, and thus were “colored”.

Directionless

Each add fiber/drop fiber can connect to/from any network fiber port. If a given add fiber and drop fiber can only connect to a given network fiber, then the ROADM is not directionless.

Contentionless

In a contentionless ROADM, the ROADM can add and drop as many instances of a given wavelength as it has network fiber pairs. For example, consider a ROADM that has 3 network fiber pairs, as well as add and drop. Imagine that each fiber is carrying signals on the 1549.32 nm wavelength, as well as on other wavelengths. An operator may wish to add and drop all three of the 1549.32 nm signals. If the ROADM can achieve this, it is contentionless. In ROADMs that have contention, instances of the same wavelength would collide at some internal element, and hence cannot be simultaneously added/dropped.

WSS can be implemented using two types of wavelength plan:

Fixed Grid

The wavelengths are on a predetermined grid, typically 100 GHz or 50 GHz, and have a predetermined and static optical pass-band.

Flex-Grid

The separation of the wavelengths and the pass-band shape of each wavelength can be varied.

The most adaptable type of ROADM is a flex-grid CDC (colorless-directionless-contentionless).

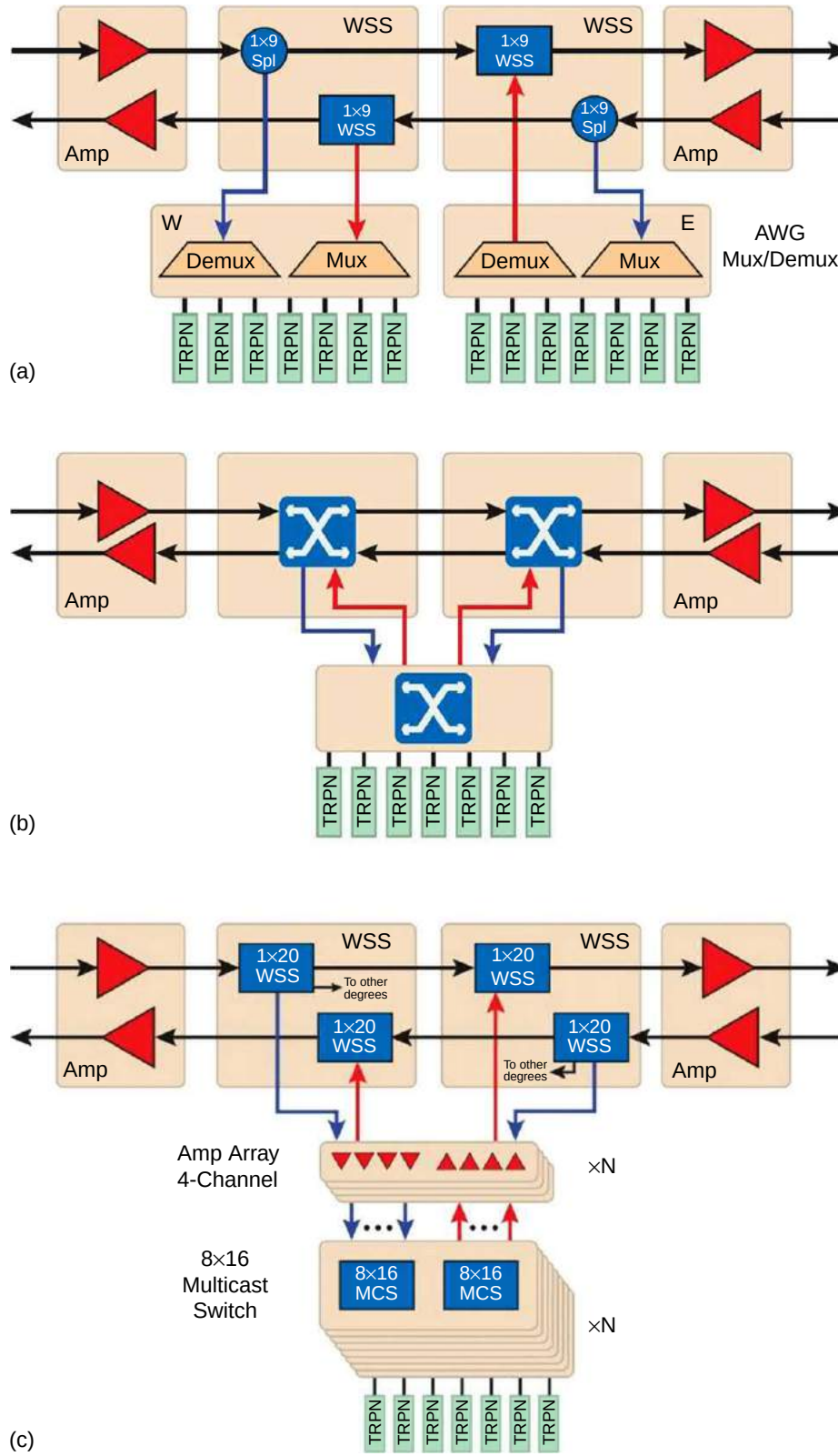


Fig. 2 Types of ROADM (a) classic, (b) CD, and (c) CDC. The CDC ROADMs offer greater flexibility, but are more complex. For illustration purposes, the classic ROADM (colored, has direction, has contention) is shown with direct detect transceivers and therefore has external deMUX, whereas the CDC ROADM is shown with coherent transceivers and thus does not need an external deMUX. (After "CD ROADM – Fact and Fiction" Part 1: Applications, Fujitsu).

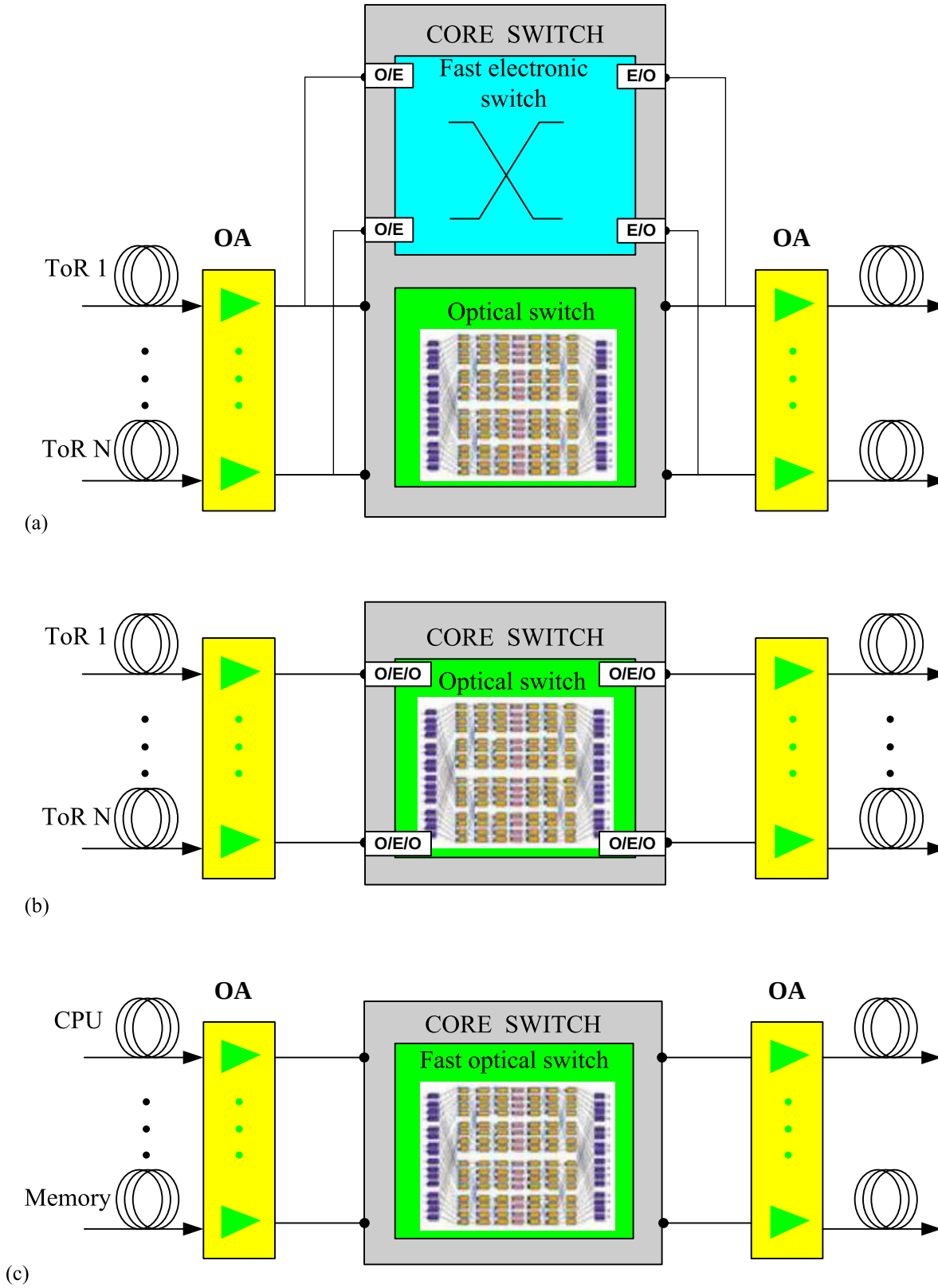


Fig. 3 Applications of optical packet switches: (a) Core router growth, (b) switch core growth, (c) high performance computing switched intraconnect.

Key Proposed Application: Optical Packet Switching

Optical packet switches are the subject of ongoing research, although they are not yet significantly deployed. For the sake of brevity, optical packet switches here include switches capable of switching individual packets, but also optical switches capable of switching custom-length frames or cells, or microsecond-duration dataflow.

Three applications are identified and shown in Fig. 3. Bandwidth requirements continue to increase in these applications. Electronic packet switches are ubiquitous today, but increasing their capacity faces increasing technical challenges related to heat dissipation, power consumption, and signal integrity.

Core Router Growth

At the core routers of a datacenter, an optical packet switch is added alongside an existing electrical router, to offload it. This may be particularly effective to offload longer packets and flows onto the optical packet switch, while leaving short packets in the electrical packet switch, because it is not practical to switch short packets optically.

Switch Core Growth

Conventional OEO switches contain OE and EO interfaces from/to other network elements, and these interfaces are connected through electrical switch chips and an electrical backplane inside the switch. Instead, the internal circuit can be replaced with additional EO and OE interfaces inside the switch, connected through an optical packet switch and an optical backplane.

HPC Switched Intraconnect

Dynamic connection between processing, memory, storage and I/O in a high performance computer (HPC).

In the first application, the optical packet switch switches light that has come from elsewhere in the network – for example, the optical transceivers may be on the other side of a datacenter from the optical packet switch. Hence the optical transceivers are also providing data transport. In the second and third applications, the additional optical transceivers are part of the switch, and are not providing data transport. The second and third applications become increasingly attractive when electrical chips have direct optical I/O, which has been proposed as a solution to the worsening I/O interconnect bottleneck of electronic chips.

Fast optical switch technologies have enormous data throughput, and can be highly efficient in terms of energy per bit. Their switch times are still relatively slow (several nanoseconds at best, for technologies that have good scalability), compared to electronic switches. Optical switches operate strictly as time-of-flight switches, and have no scalable capability for random-access-memory buffers. On the other hand, electrical packet switches switch at sub-nanosecond speed and most such switches today use a memory-based switching architecture, including random-access-memory buffering to handle contention. Hence a network using an optical packet switch cannot simply copy a conventional electronic packet network. Instead, the most practical solutions use a scheduler and ingress buffering to solve time-of-flight contention, and may also retain a conventional electrical switch for some classes of traffic (Fig. 4). Ingress buffering means that the source of the data is buffered at the network element before the optical

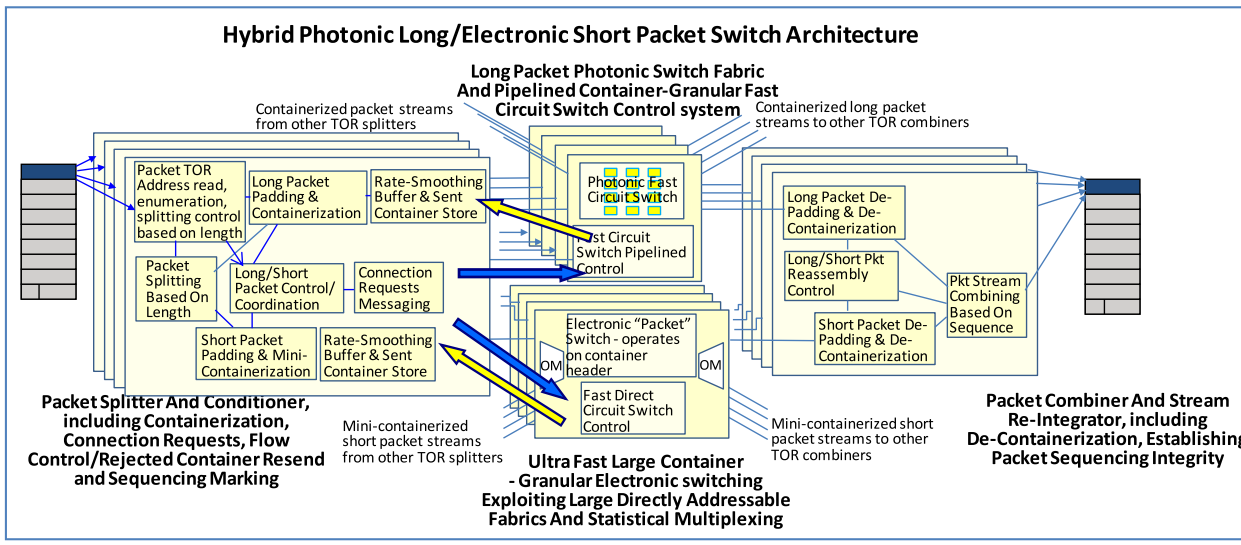


Fig. 4 Hybrid optical packet switch and electrical packet switch, illustrating ingress buffering for data that is handled by the optical packet switch.

packet switch, until the scheduler indicates that it is time to send. This architecture is illustrated in the figure. Thus, it is likely that electrical and optical packet switches will coexist and cooperate.

As will be discussed below in the section on planar lightwave technology, a realistic limit today for a fast photonic switch chip is around 32×32 to 64×64 ports. By a careful choice of parallel architecture, a switch implemented using these chips can implement 1000–4000 I/O ports, each having a throughput of 400 Gb/s or higher. Thus, the overall capacity of an optical packet switch can reach Petabits per second.

Free-Space Optical Switch Technologies

This section describes optical switches that operate by steering beams of light in air (or some other suitable gas). These switches are suitable for circuit switching as they switch relatively slowly (milliseconds), although some LCoS optical switches can achieve microsecond switching time and have been used for experiments with switching short-duration data flows. Generally, free-space optical switches have excellent optical performance. They are precision opto-mechanical assemblies, and so their manufacturing is relatively time-consuming. Careful engineering of vibration control and actuator feedback is used to achieve good alignment.

The 2D-MEMS, 3D-MEMS and piezoelectric optical switches described in this section have fiber switching granularity. If the application calls for switching individual wavelengths, then fiber-coupled wavelength deMUX and MUX components are needed before and after the optical switch. Single-mode and multi-mode versions have been built.

The WSS described in this section have wavelength switching granularity. They contain built-in diffraction gratings that perform wavelength deMUX and MUX. They always use single-mode fibers.

Liquid-Crystal on Silicon Wavelength-Selective Switch

Wavelength selective switches have a collimating lens after each input fiber, and collimate the incoming light onto a diffraction grating. The diffraction grating disperses each signal wavelength into a beam at a different angle. The beams then fall upon an actuator, which reflects them to a desired output fiber focusing lens. Generally, a diffraction grating is used to recombine signal wavelengths (which may be from different inputs) into the output fiber.

The form of the actuator has evolved over time. Early implementations used a MEMS-actuated mirror to tilt each beam, or used a small number of liquid crystal pixels to reflect the beam. By applying a controllable voltage to each liquid crystal pixel, its refractive index is changed. A set of liquid crystal pixels forms a phased-array reflector, which steers the reflected beam. These implementations are limited in their spectral resolution, and typically have fixed grid wavelength granularity at 50 GHz or more channel separation.

Liquid-crystal on silicon (LCoS) WSS were introduced commercially in the mid-2000s and are particularly intended for flexible grid operation. A concept demonstration of a WSS with LCoS optical sensor is shown in Fig. 5. The liquid crystal medium is directly on top of a silicon CMOS control chip. The size of the pixels in an LCoS is very small compared to the size of the beam. The number of liquid crystal pixels hit by the beam depends on the spectral width of the signal in the beam. The control system chooses the appropriate set of pixels depending on the present set of wavelength channels in each fiber, which is learned from network operating software. LCoS granularity down to 6 GHz channel separation is possible. Commercial LCoS WSS with up to 20 fibers are commonly available today.

3D Micro-Electro-Mechanical Optical Switch

3D-MEMS optical switches became well-known around the year 2000. They are currently deployed mostly in patch panel and reconfiguration applications.

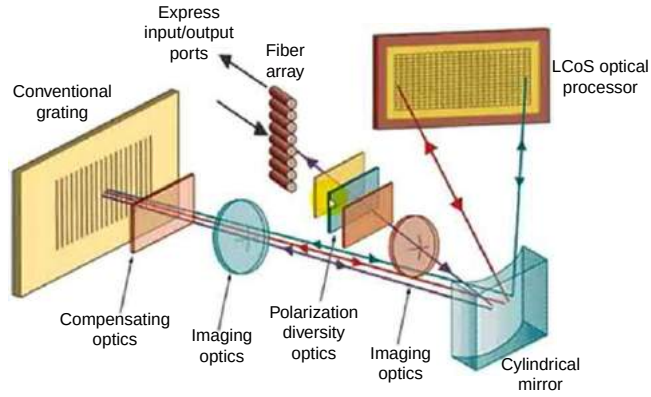


Fig. 5 Wavelength selective switch using liquid crystal on silicon.

A 3D-MEMS, as schematically demonstrated in Fig. 6, has a collimating lens on each input fiber that produces a collimated beam in a gas. The beam reflects off a mirror in a first plane of mirrors, which sends the beam to a mirror in a second plane of mirrors that steers the beam into the acceptance angle of a focusing lens of an output fiber. The beam travels a distance of centimeters through the gas. Each mirror array is a silicon substrate, with tiltable micro-mirrors that are driven by MEMS actuators. A controller tilts each mirror by a variable amount to achieve the desired beam alignment. Typically, an optical monitor in the output fiber provides feedback to the controller.

Most 3D-MEMS optical switches have from 64 to a few hundred ports, and the largest built had more than 1000 ports. Insertion loss is typically less than 3 dB, return loss and crosstalk are low, and switching time is in the order of milliseconds to 10's of milliseconds. The optical switch is non-blocking, operates over a wide wavelength range, and may be bi-directional (depending on the control system). The controller requires careful engineering to achieve and maintain good alignment, and to compensate for environmental vibration.

For a 3D-MEMS optical switch with N inputs and N outputs, there are $2N$ mirrors. For example, a 300×300 switch needs 600 mirrors.

The figures illustrate the 3D-MEMS concept. In practice, the optical path may be folded using static mirrors, to reduce physical volume and to allow the input and output fibers to be assembled in the same plane.

2D Micro-Electro-Mechanical Optical Switch

2D-MEMS optical switches were fashionable in the late 1990s, attracting significant research and industry investment. However, in more recent years they have attracted less attention than 3D-MEMS.

In a 2D-MEMS device, as conceptually illustrated in Fig. 7, each optical input fiber has a lens which creates a collimated beam, and each output fiber has a focusing lens and is at 90° to the input. There is a 2D array of mirrors built on a silicon substrate by MEMS manufacturing processes. To connect an input to an output, one mirror is inserted into the beam, and the other mirrors are moved out of the way. The mirror has two very well defined states; it is either "not inserted" or is "inserted". Several insertion

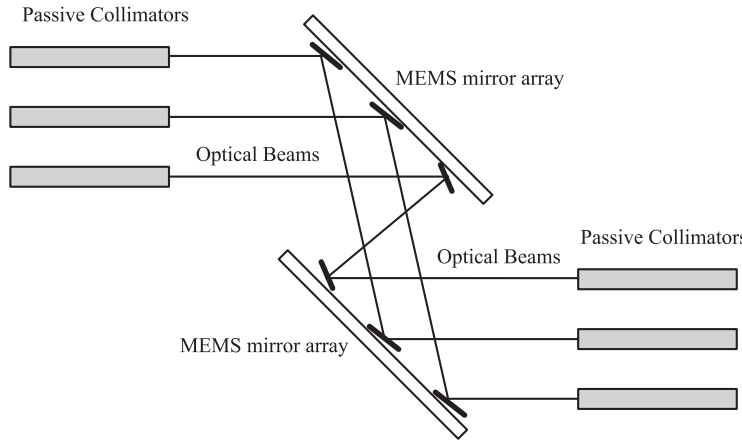


Fig. 6 Schematic of a 3D-MEMS optical switch.

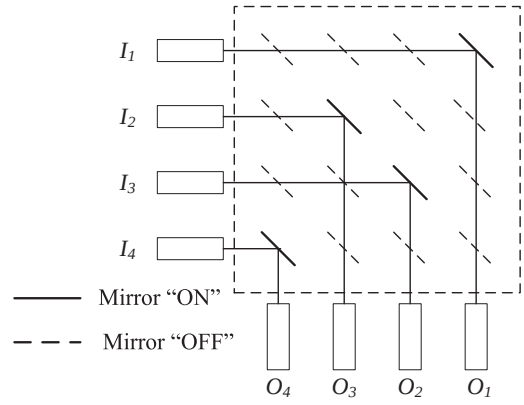


Fig. 7 Schematic of a 4×4 2D-MEMS optical switch.

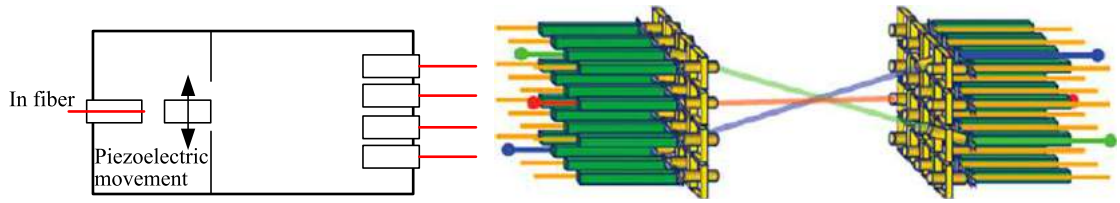


Fig. 8 Schematic of a piezoelectric optical switch. Reproduced from Jencotech, Available at: http://www.jencotech.com/documents./PolatisOverviewRev1.pdf.

mechanisms have been used: the mirror is stood up/laid down, the mirror is slid sideways into the beam, or the mirror is lowered into the beam. Actuators are electrostatic or micro-motors. The mirror is typically silicon, with a reflective coating. Generally, the angle of the mirror is defined by the manufacturing and is not adjustable. For an optical switch with N inputs and N outputs, there are N^2 mirrors.

Switching time is typically in the order of sub-milliseconds, which is 5–10 times faster than the 3D-MEMS devices. The advantage is due to simple direct digital control operation of the mirror movement from a rest position to an active position.

A 2D-MEMS has excellent optical properties because the optical path contains just 2 lenses and one mirror. However, it becomes difficult to achieve good optical beam control at large port count, as the distance between the lenses is larger and thus angular alignment during manufacturing becomes more critical. Typical larger elements are around 16×16 ports.

Piezoelectric Moving-Fiber Optical Switch

A piezoelectric optical switch (Fig. 8) is used in similar applications as a 3D-MEMS switch. It has similar capabilities, performance, physical size, and number of ports as 3D-MEMS, but it uses a different optomechanical mechanism.

In a piezoelectric optical switch, each input fiber has a collimating lens, and each output fiber has a focusing lens and there is a gas region between the lenses the middle, much like the 3D-MEMS. The position of each fiber is moved slightly with respect to its lens, using piezoelectric actuators, so that the beam from the collimating lens is steered toward a desired focusing lens, and the acceptance angle of the focusing lens is tilted so as to couple the beam into the output fiber.

Planar Lightwave Circuit Optical Switch Technologies

Planar lightwave circuit (PLC) optical switches comprise optical waveguides which are manufactured on wafers using lithographic processes, then diced into chips and packaged with optical fibers for optical I/O. Each connection from an input to an output on the chip is called a lightpath. To set up the lightpaths, electrical controllers connect to the planar lightwave circuit. The electrical controllers are usually off-chip. In almost all cases, PLC optical switches are single mode and therefore are compatible only with single mode fiber.

Materials and Physical Principles

A wide variety of materials have been used for PLC optical switches, including silicon-on-insulator, silicon dioxide, silicon oxynitride, GaAs, InP and related III-V semiconductors, polymers, electro-optic polymers, and electro-optic crystals of which lithium niobate and PLZT are the most popular.

PLC optical switches operate using one of the following physical principles:

Interferometer

The electrical signal creates a change in refractive index of the optical waveguide. Each optical switch cell is arranged as an interferometer with 2 inputs and 2 outputs. Changing the refractive index changes which input is connected to which output. The most common forms of interferometer are Mach-Zehnder and microring resonator. To change the refractive index, the electrical signal changes the temperature of the waveguide (thermo-optic effect in dielectric, semiconductor or polymer), changes the density of electrical carriers (carrier injection in semiconductors), or changes the electric field (in materials with an electro-optic coefficient).

Moving waveguide coupler

The electrical signal causes a waveguide to move physically. The moving waveguide is part of a directional coupler with 2 inputs and 2 outputs. Moving the waveguide changes which input is connected to which output. See the section on waveguide-MEMS.

Blocker

The electrical signal changes the optical waveguide from absorbing to having gain. The optical switch cell is a gate with 1 input and 1 output. The gate blocks or transmits light. See the section on SOA switches.

All these PLC optical switches are spectrally broadband and therefore switch all wavelengths within a 10's-nm wide spectrum, except the microring resonator which is narrowband and switches a single WDM wavelength.

Scaling up a Planar Lightwave Optical Switch From Small Building Blocks

Many planar lightwave optical switch technologies can only implement cells with 1 or 2 inputs and outputs. To achieve scalability, these cells are connected together in a switch matrix, preferably with as many cells as possible on one die. In addition to the external characteristics of the overall optical switch, the internal topology of the switch matrix can be described using the following characteristics:

Symmetric

In a symmetric topology, every lightpath uses the same number of cells. In an asymmetric topology, the number of cells in a lightpath depends on the lightpath.

Deterministic

In a deterministic topology, there is exactly one path from a given input to a given output. In a non-deterministic topology, there is more than one possible path, and a routing algorithm considers all the requested connections.

Stages

In a single stage switch, one cell is actuated per lightpath. In a multi-stage switch, multiple cells are actuated per lightpath.

Some important topologies are listed in the Table 3 and illustrated in Fig. 9. For simplicity, the table shows the case where the number of input ports N is equal to the number of output ports, which is known as an $N \times N$ switch. Multi-stage switches are most efficient when the number of ports is a power of 2. Any of these switch matrices can, in principle, be cascaded to make a Clos architecture, although currently actual deployments of a Clos optical switch are rare to nonexistent.

The **crossbar** fabric is a type of switching technology where each node is connected to every other node. The interconnection between the inputs and the outputs is achieved by appropriately setting the states of the 2×2 switches. For a $N \times N$ switch, the total number of 2×2 switching cells is N^2. Fig. 9(a) demonstrates connectivity in a 4×4 switch matrix. To connect input i to output j, the light traverses the 2×2 switches in row i until it reaches the column j and then traverses the switches in column j until it reaches output. The shortest light path length is 1, and the longest is $2N - 1$, resulting in large variation in path length. The crossbar architecture is non-blocking, has a simple layout, and small cell count at small N. Drawback of this topology is the large number of crossings and large cell count at large N.

The **switch and select** architecture consists of an input switch array, an output switch array, and a passive crossover network. The example of Fig. 9(b) demonstrates a 4×4 switch matrix. The number of switches is uniform in all optical paths and follows logarithm scaling instead of the linear scaling. The passive crossover network connects the N^2 outputs from the input switch array to the N^2 inputs of the output switch array. The large cell count at large N and the large number of on-chip waveguide crossings limit the potential for scalability and introduces different insertion loss on the different paths.

Table 3 Switch topologies used with planar lightwave circuit switch cells

Topology	*# Cells per path*	*# Cells total*	*Type*	*Blocking*	*Strength*	*Weakness*
Crossbar	1 to $2N-1$	N^2	Single stage, symmetric deterministic	None	Simple layout Small cell count at small N No blocking	Many crossings Large variation in path length Large cell count at large N
Switch & select	$2 \log_2 N$	$2\ N\ (N-1)$	Multi-stage symmetric deterministic	None	One plane of crossings	Large cell count at large N
Path independent loss	$2 \log_2 N + 2$	$2N^2$	Multi-stage symmetric non-deterministic	None	Equal loss on all paths	Large cell count at large N
Banyan	$\log_2 N$	$(N \log_2 N)/2$	Multi-stage symmetric non-deterministic	Blocking	Small number of cells	High crosstalk Blocking
Benes	$2 \log_2 N - 1$	$N \log_2 N - N/2$	Multi-stage symmetric non-deterministic	Almost none	Small number of cells	Very high crosstalk
Dilated Benes	$2 \log_2 N + 2$	$2N\ (\log_2 N + 2)$	Multi-stage symmetric non-deterministic	Almost none	Low crosstalk Efficient routing algorithm	Larger number of cells in path

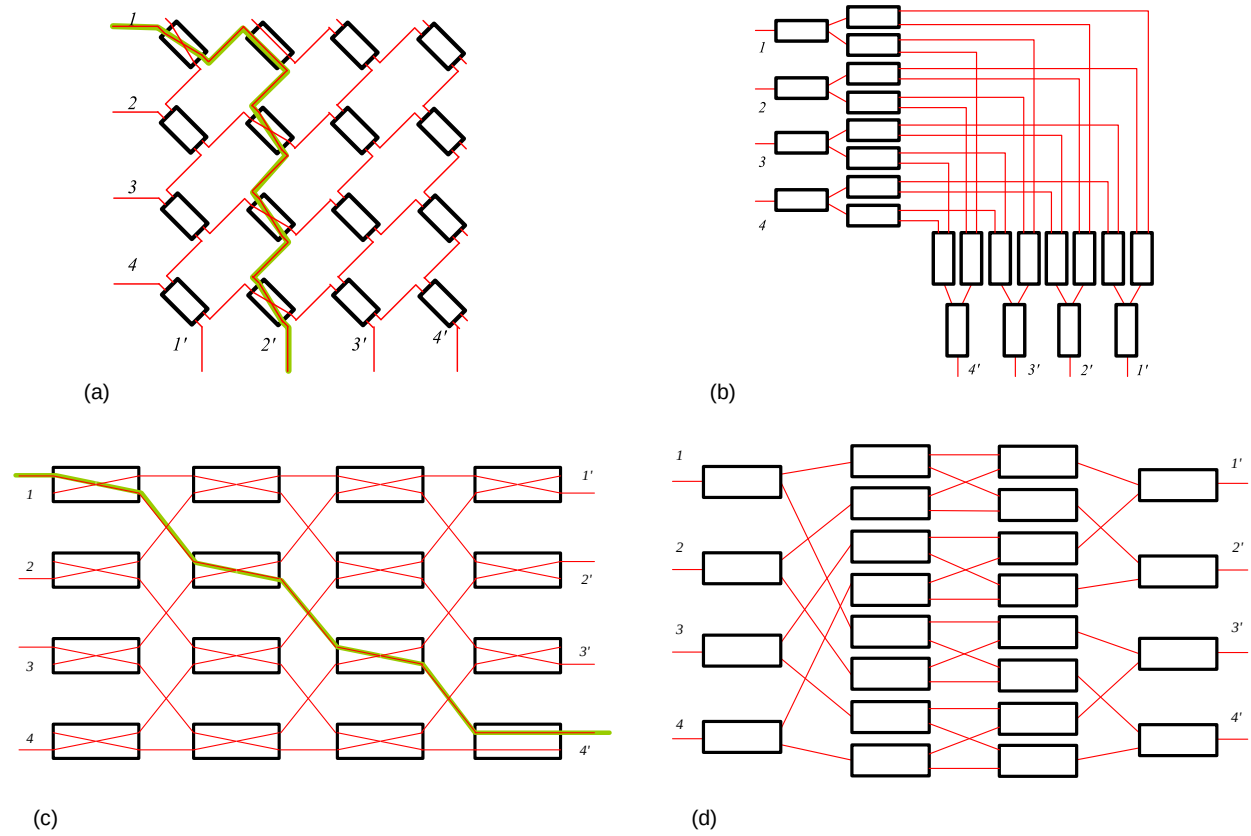


Fig. 9 Schematic of four topologies constructed from small switches: (a) crossbar switch, (b) switch and select switch, (c) path-independent loss (PILOSS) switch, and (d) dilated Benes switch.

PILOSS is another standard switch topology for building $N \times N$ switches. It consists of a matrix of 2×2 element switches as shown in Fig. 9(c). It usually is designed to have the element switches 'normally' OFF (cross-state), therefore reducing the total power consumption of the switch matrix. In this topology, light always passes through N element switches and $N - 1$ intersections, thus experiences the equal attenuation for all the N^2 – ways of paths. The maximum number of ON state element is N, making the PILOSS topology energy-efficient. For this reason, it is believed to be one of the most promising towards larger N.

A **hybrid dilated Benes** with crosstalk suppression is shown in Fig. 9(d), comprised of an ingress column of 1×2 cells, two vertically replicated 4×4 Benes in the middle, and an egress column of 2×1 cells. By replacing the central column of 2×2 cells each with a 2×2 dilated Banyan, crosstalk noise can be suppressed. In addition to crosstalk suppression, use of wavelength constraint routing (routing of lightpaths within the switch matrix to minimize how many time different lightpaths meet within the switch matrix) reduces output crosstalk.

Materials for Planar Lightwave Optical Switches

Silica-on-silicon is a dielectric that is widely used in passive planar lightwave circuits. The core and cladding are composed of silicon dioxide with different dopants that modify the refractive index slightly. The fabrication process is inexpensive, connection to single-mode optical fiber is relatively easy and low loss, and optical propagation loss in the waveguide is low. Silica has a thermo-optic effect which can be used as a switch actuator. However, bends are very large, due to the small refractive index step between the core and cladding of the waveguide. Therefore it has very poor scalability for an optical switch.

Silicon oxynitride (SiON) is similar to silica-on-silicon, except that the waveguide is made of different compositions of SiON. The refractive index of the SiON can be adjusted in the range 1.45–2.01 depending on the composition, which is wider than the refractive index variation in silica. This allows smaller bend radius. SiON is increasingly used for passive planar lightwave circuits, but its use for optical switches is relatively uncommon.

Perovskite crystals include lithium niobate and lead lanthanum zirconium titanate (PLZT). Waveguides are formed by doping the material, typically by diffusing metal ions into the material. Perovskites are used for very fast switches with up to 8 ports. They have an electro-optic effect, whose switch time is limited only by the electronic controller. The electro-optic effect is much weaker per unit length than thermo-optic and carrier injection effects, so a perovskite optical switch cell is large. Therefore, perovskites do not scale up to very large port count.

III-V semiconductors that are epitaxially grown on a GaAs or InP wafer can create thermo-optic interferometer optical switches, carrier-injection interferometer optical switches and SOA optical switches. The waveguide is formed using III-V materials of differing composition. Further, when using III-V semiconductor interferometric optical switches, non-switching SOAs can be monolithically integrated to provide optical gain to compensate for optical losses in the switch circuit. III-V semiconductors are ubiquitous for optical transmitters and receivers, but activity on III-V semiconductor optical switches has declined in the 2010s.

Silicon photonics (SiPh) is fabricated in the silicon layer of a silicon-on-insulator substrate, with silicon dioxide as the waveguide cladding. SiPh can create thermo-optic interferometer optical switches, carrier-injection interferometer optical switches and waveguide-MEMS optical switches. Activity on SiPh semiconductor optical switches has grown in the 2010s as SiPh manufacturing technology becomes more widespread. Silicon is transparent over the optical telecommunication wavelength band (1300 nm–1600 nm). It has high refractive index contrast n[Si]/n[SiO_2]=3.48/1.45 which allows very dense circuit layouts. It is cheap and can be manufactured using legacy silicon electronics processes that have exceptionally good process control. The small waveguide size and large index step causes large propagation loss, which can be overcome using very wide waveguides for long-distance routes on the chip.

Interferometric Optical Switch

Mach-Zehnder interferometer optical switch

A Mach-Zehnder interferometer optical switch cell (Fig. 10) comprises a 50/50 input splitter, two internal optical waveguides and a 50/50 output combiner. Commonly used 50/50 couplers are multi-mode interferometers (MMI) and directional couplers. The two internal optical waveguides are nominally the same length. An electrically-controlled phase shifter on one or both arms changes the relative phase of the light on the two arms as it enters the output combiner, which causes the output light to move from one output waveguide to the other. When the phase difference between the arms is $2N\,\pi$, the switch is in cross state, which means that input 1 is connected to output 2 and input 2 is connected to output 1. When the phase difference between the MZ arms is $(2N+1)\,\pi$, the switch is in bar state, which means that input 1 is connected to output 1 and input 2 is connected to output 2.

This figure shows only the phase shifter(s) that perform the switching function. However, it is impossible to manufacture the two waveguide arms to have equal optical path length to within a small fraction of a wavelength. Therefore, carrier injection switches also have a static thermo-optic phase shifter to bias the initial phase. If care is not taken, the power consumption of this bias thermo-optic phase shifter dominates the power consumption of the whole switch.

Microring resonator optical switch

Microring optical switches use the thermo-optic effect or the carrier injection effect, and switch a single wavelength of light. They have been proposed for add/drop bus switches, but not to-date been commercially deployed for this purpose.

A microring resonator (MRR) couples light from an input waveguide to an output waveguide when the wavelength of the light resonates with the round-trip of the microring. Microring resonators are usually made in silicon photonics. The diameter of the ring is from several micrometers to tens of micrometers. A single microring has a sharp resonance peak with a Lorentzian lineshape. This lineshape is not compatible with most WDM transmission formats. Multiple microrings can be coupled to produce a flatter-top lineshape that is compatible with a conventional WDM signal.

A significant challenge is that MRR are highly temperature dependent, very sensitive to manufacturing variations, and difficult to maintain alignment with the desired wavelength. In addition, a single-MRR with a sharp Lorentzian transmission is not favorable in a high-speed data transmission system where flat-top response is preferred to minimize the signal waveform distortion and thus the data transmission errors.

Fig. 11 shows a schematic view of an optical add/drop bus, formed using microring optical switch cells. Each microring couples one wavelength from the through waveguide to the add/drop waveguides. Each microring contains a phase shifter. Varying the phase shifter using the thermo-optic effect or the carrier injection effect determines which wavelength is added/dropped.

Microring optical switches have also been used to create switch matrices, such as the crossbar illustrated in Fig. 12. Each microring couples light from a west-to-east input waveguide to a north-to-south output waveguide. The wavelength that is coupled is controlled by the phase shifter in the microring.

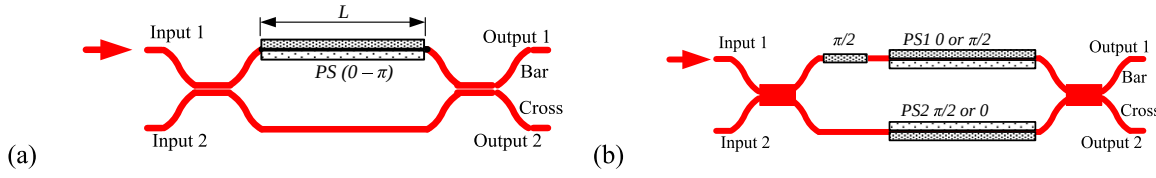


Fig. 10 Mach-Zehnder interferometer optical switch cell: (a) single-arm drive, (b) dual arm drive (push-pull).

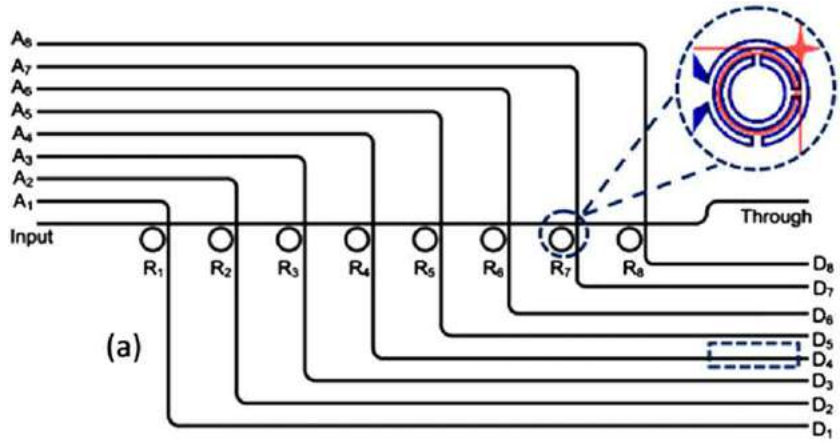


Fig. 11 Add/drop bus of an optical switch with MRRs. Reproduced from Danning Wu, Yuanda Wu, Yue Wang, Junming An, Xiongwei Hu, 2016. Reconfigurable optical add-drop multiplexer based on thermally tunable micro-ring resonators. Optics Communications 367, 44–49 (Elsevier).

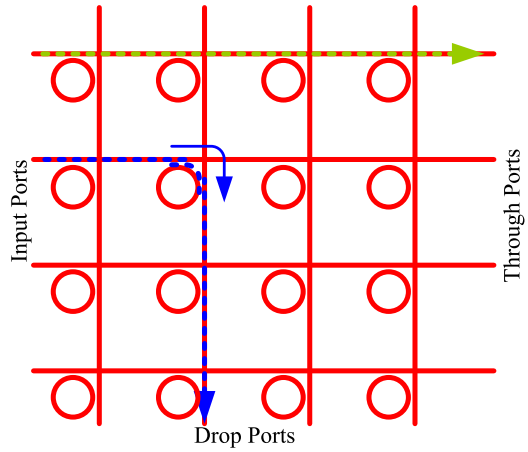


Fig. 12 Microring optical switch, arranged as a crossbar.

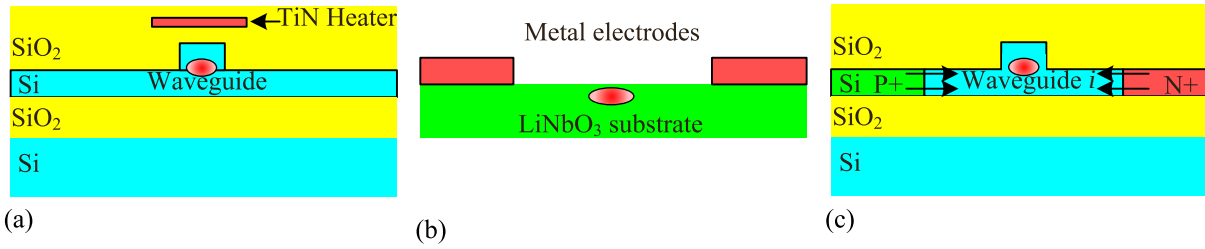


Fig. 13 Cross-section of a (a) thermo-optic phase shifter in silicon photonics (b) electro-optic phase shifter in lithium niobate substrate, and (c) carrier-injection phase shifter in silicon photonics or III-V semiconductor.

Optical phase shifters

Cross sections of typical phase shifters are shown in Fig. 13. The phase change induced through the length of the phase shifter is $\Delta\phi = (2\pi\Delta n)/\lambda L$ where Δn is the change in the effective index of refraction of the guided mode due the electrical control, λ is the free-space wavelength of the propagating light, and L is the length of the phase shifter.

A goal of optical switching is to reduce the energy consumed per bit, compared to an electrical switch. However, a simple thermo-optic phase shifter in silicon photonics consumes around 30 mW of electrical power for π-phase shift (P_π), which is much larger than 1 pJ/bit for a complex circuit such as a 32-port matrix at a line rate of 100 Gb/s per port. A thermal undercut technique greatly improves the power consumption of a thermo-optic switch. Fig. 14 demonstrates a thermo-optic switch cell in silicon

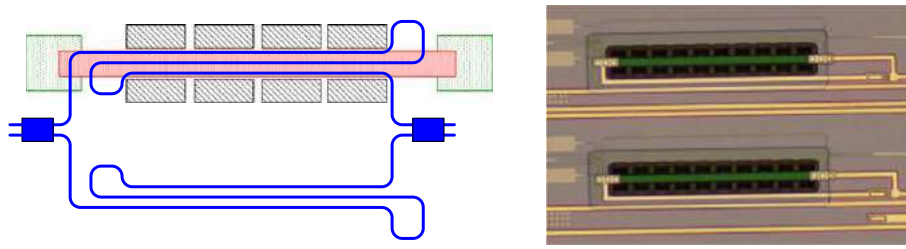


Fig. 14 Thermo-optic Mach Zehnder silicon photonic optical switch cell with thermal undercut: (a) schematic layout where a resistive heater (pink) located above one of the waveguide arms (blue) is used to change the index of refraction causing a phase shift in the propagating light wave, (b) microscope image of the fabricated switch.

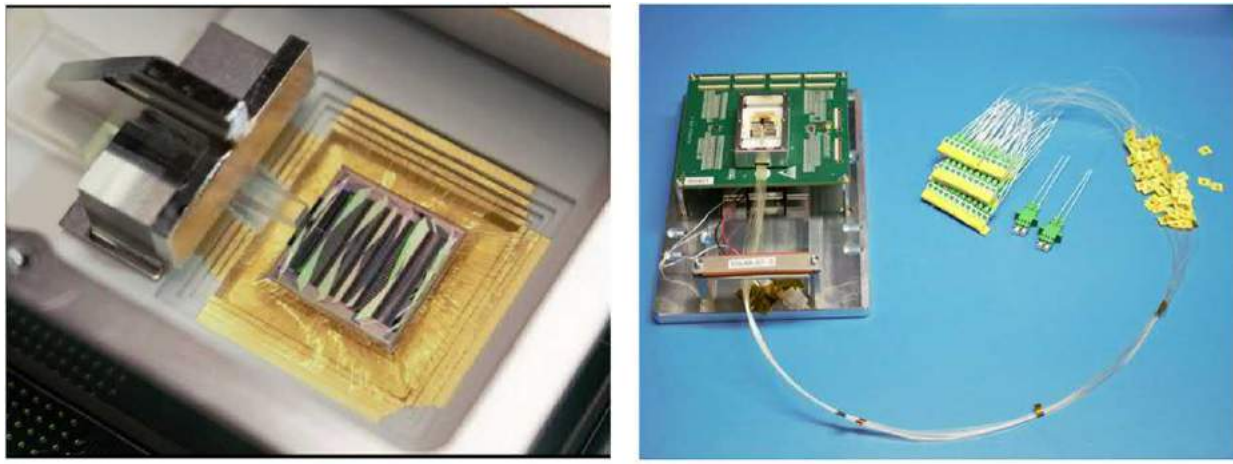


Fig. 15 32×32 SiPh switch (a) packaged die containing 448 cells with 900 photodiodes, wire-bonded in CBGA and silica waveguide mode adapter couples 64-fiber ribbon to SiPh, and (b) system assembly comprising optical package on breakout board, fiber ribbon, detector A/D boards (on back of breakout board), and thermal test plate. Note: the FPGA and heater driver board are not shown. After Dumais, P., Goodwill, D., Kiaei, M., *et al.*, 2017. Silicon Photonic Switch Subsystem With 900 Monolithically Integrated Calibration Photodiodes and 64-Fiber Package, OFC.

photonics with thermal undercut. Deep trench windows are etched vertically on each side of the thermo-optic phase shifter, followed by anisotropic wet chemical etching of the silicon substrate locally under the heated region. This creates a suspended heater and waveguide, which has low thermal conductivity to the silicon substrate, reducing the power consumption by a factor of 20. In addition, folded waveguide arms under the heater increases the interaction length between the heater and waveguide, thus reducing the power consumption by a factor of the number of folds, which is typically 3. The energy consumption then falls to 100 fJ/bit for the entire optical switch matrix, which is competitive with electrical switches.

A demonstration of a fully-assembled, fully-functional thermo-optic 32×32 silicon photonic optical switch matrix with thermal undercut for reduced power consumption is shown in Fig. 15.

A better-isolated heater region consumes less power, but needs more time to tune. Fig. 16 shows the time-response, when the switch is driven by an electrical square wave with amplitude $P\pi$. The 10%–90% switching time increases from 70 μs without undercut, to 1.34 ms with undercut. Nonetheless, as described earlier, millisecond response time is appropriate for many applications.

In a carrier injection phase-shifter, the mechanism to actuate the switch is the free-carrier plasma dispersion effect in a lateral *p-i-n* junction, which is centered on the optical waveguide. A forward-biased current is passed through the junction, which increases the carrier density. The lifetime of the carrier density is 1–2 nanoseconds, and hence the switch time of the phase shifter is several nanoseconds. Therefore, this mechanism is appropriate for optical packet switches. The electrical consumption is typically 1 to 3 mA at about 1 V. The increase in carrier density creates a decrease in refractive index, which saturates as the carrier density increases. However, it also creates an increase in absorption which increases monotonically with carrier density. In silicon photonics, the refractive index change before saturation is large enough to create a push-pull switch with only 0.5 dB absorption loss per switch cell and extinction ratio of 25 dB, which is sufficient for a scalable optical switch. However, a push-only switch has

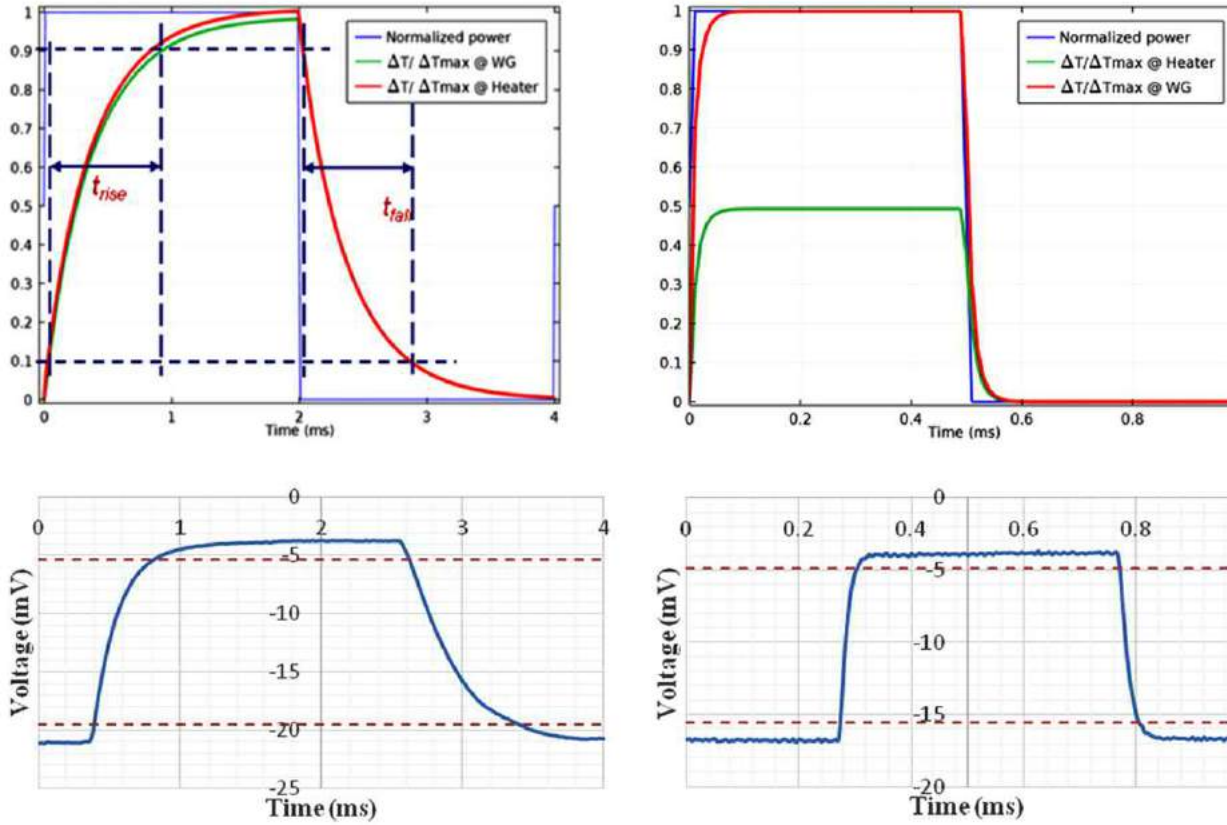


Fig. 16 Switching time of thermo-optic Mach Zehnder silicon photonic optical switch cell: (a) modeled and (b) measured switching time response, with (left) and without (right) thermal undercut. After Celo, D., Goodwill, D.J., Jiang, J., *et.al.*, 2016. Thermo-optic silicon photonics with low power and extreme resilience to over-drive. IEEE Optical Interconnects Conference (OI), pp.26–27.

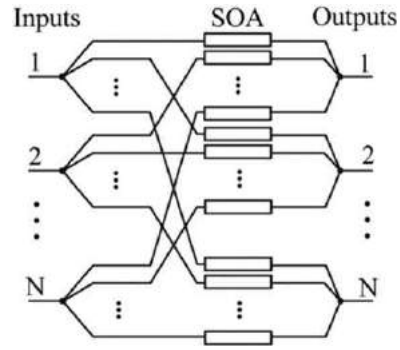


Fig. 17 Semiconductor optical amplifiers (SOA) used in a split-and-select optical switch.

to be driven too hard into saturation, and the absorption loss is 2 dB per switch cell with extinction ratio of 18 dB or less, and therefore push-only carrier injection switches are not effective for scalable optical switches.

Semiconductor Optical Amplifier Switch

Semiconductor optical amplifiers (SOAs) are used in optical switches as an *ON–OFF* switch by varying the bias voltage. SOAs provide nanosecond switching times and broadband operation, have relatively high gain (15–28 dB) with saturation power in the range of +8 to +15 dBm, are small (100's micrometers to few millimeters in length), and can be integrated with other semiconductor and optical components. The typical architecture of an SOA optical switch is a split-and-block topology, integrating a bank of passive splitters, a bank of SOA, and a bank of passive combiners, as shown in Fig. 17. Larger port counts may use several stages. SOA optical switches demonstrated to date have up to 32×32 ports.

An SOA gate can be switched from a high-gain state to a high-loss state within a nanosecond, by changing the drive current from an electronic driver. The difference between large amplification in the *ON*-state and large absorption in the *OFF*-state provides very high extinction ratio, which is needed for the split-and-select architecture.

SOA optical switches have some performance weaknesses. They have relatively high power consumption, of 10's to 100's mW per SOA in the *ON* state, although the *OFF* state power consumption may be small or zero. Many SOAs are polarization dependent, thus requiring polarization diversity lightpaths or single-polarization networks. SOA are best used to carry a single wavelength, because they have large nonlinearity that causes mixing between signals on different wavelengths.

In an alternative use, SOAs serve as optical power amplifiers to compensate for optical losses in other types of planar lightwave optical switches. III-V semiconductor gain elements can be heterogeneously integrated onto a silicon photonic circuit, with the Si and InP chips independently optimized for high performance, reliability and yield.

Waveguide-MEMS Optical Switch

A waveguide-MEMS switch is a planar lightwave optical switch, using waveguides that are moved by micro-mechanical actuators. Each moving waveguide is one half of an adiabatic waveguide coupler. Each switch cell comprises a pair of adiabatic couplers joined by a 90° bend (Fig. 18(a)). In the *OFF* state, the coupler waveguides are moved away from the bus waveguides, and there is no light coupled between the bus waveguides. Upon application of a voltage to the actuator electrodes, the two movable waveguide couplers are brought in close vicinity of the bus waveguide, creating optical coupling between the fixed and the movable waveguides. In the *ON* state, the optical signal is thus routed from the input port to the desired drop port. The switching time is 10's of microseconds.

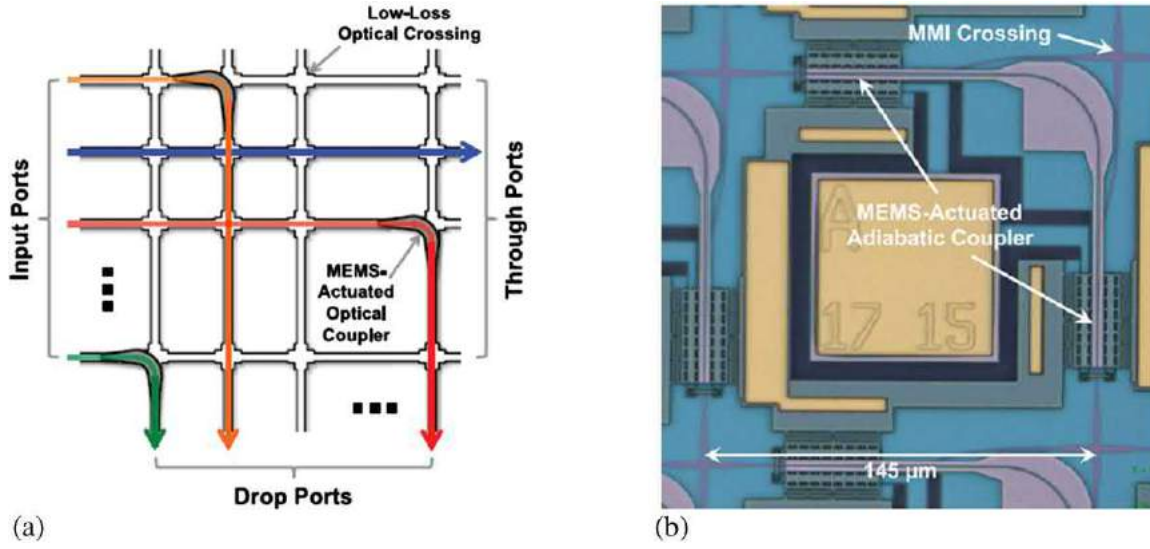


Fig. 18 Optical switch with MEMS-actuated adiabatic waveguide couplers (a) schematic, and (b) microscope image of a representative unit cell. Reproduced from Seok, T.J., Quack, N., Han, S., Muller, R.S., Wu, M.C., 2016. Highly scalable digital silicon photonic MEMS switches. Journal of Lightwave Technology 34(2), 365–371.

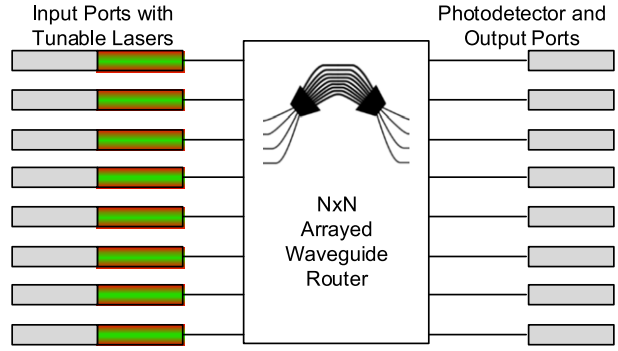


Fig. 19 Tunable laser optical switch – schematic representation of the system including tunable lasers and $N \times N$ AWGR (for $N = 4$).

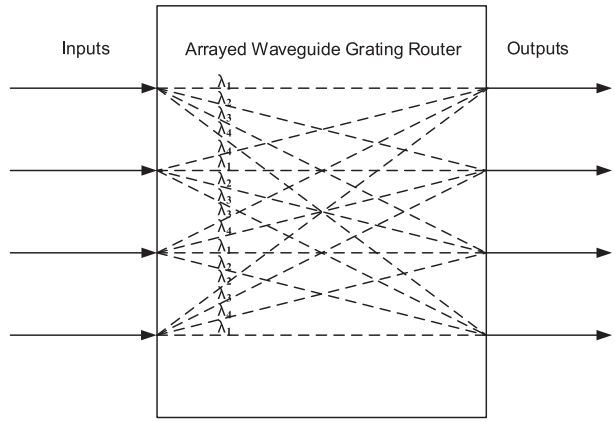


Fig. 20 Details of arrayed waveguide grating router (AWGR) behavior, as used in a tunable laser optical switch. An AWGR is a passive, static planar lightwave circuit, typically made in silica.

A 50×50 optical switch has been demonstrated, using an optical crossbar. The insertion loss compares favorably to other technologies.

Tunable Laser Optical Switch Technologies

A tunable laser optical switch comprises a rapidly-tuned laser at the source, and passive routing using an arrayed waveguide grating routing (AWGR). The switch is illustrated in **Fig. 19** and the behavior of the AWGR is illustrated in **Fig. 20**. To select a destination, the wavelength of a laser is tuned so that the light comes out from the desired output port of the AWGR.

The actuator is in the transmitter – in this manner, the tunable laser optical switch is fundamentally different to all other types of optical switch that are described in this article, as in all other cases the actuator is in the core switch.

An AWGR is similar to the more widely-used optical demultiplexing AWG, where an optical input fiber carries light of multiple wavelengths. The light power is split, passes through an array of curved waveguides of unequal length, and then recombines in a star coupler. The star coupler has multiple output waveguides, and the phase relationship of the curved waveguides is constructed so that light from each wavelength interferes constructively at only one of the output waveguides. An AWGR is similar, except that the input coupler has multiple input fibers, and the geometry is organized so that the mapping of input wavelength at output ports is cyclical. A typical AWGR has -25 dB adjacent channel crosstalk, 7 dB insertion loss, and 3 dB uniformity.

Tunable laser optical circuit switches have been used experimentally for core router growth and for switch core growth (see earlier section for definition of these applications). In both cases, the switching time is limited by the time that it takes to accurately switch the wavelength of the laser. Microsecond switching time is possible, but control is very challenging at nanosecond switch time. Scalability is limited to 40 ports, and AWGRs cannot be cascaded to form large port count systems. Also, AWGRs require DWDM transmitters, which are not commonly used within data centers or high performance computers as they are expensive and consume a large power.

Further Reading

Ben Yoo, S.J., 2006. Optical packet and burst switching technologies for the future photonic internet. Journal of Lightwave Technology 24 (12),
Celo, D., Goodwill, D.J., Jiang, J., *et al.*, 2016. Thermo-optic silicon photonics with low power and extreme resilience to over-drive. IEEE Optical Interconnects Conference (OI), pp. 26–27.
Dharmaweera, M.N., Parthiban, R., Sekercioglu, Y.A., 2015. Toward a power-efficient backbone network: The state of research. IEEE Communication Surveys & Tutorials 17 (1), 198–227.
Dumais, P., Goodwill, D., Kiaei, M., *et al.*, 2017. Silicon Photonic Switch Subsystem With 900 Monolithically Integrated Calibration Photodiodes and 64-Fiber Package, OFC.
Dupuis, N., Rylyakov, A.V., Schow, C.L., *et al.*, 2016. Nanosecond-scale Mach-Zehnder-based CMOS photonic switch fabrics. Journal of Lightwave Technology 35 (4), 615–623.
Han, S., Seok, T.J., Quack, N., 2015. Large-scale silicon photonic switches with movable directional couplers. Optica 2 (4), 370.
Hinton, H.S., 1993. An introduction to photonic switching fabrics. In: Lucky, R.W. (Ed.), Applications of Communications Theory. New Jersey: Bellcore, Red Bank.
Kachris, C., Tomkos, I., Bergman, K., *et al.*, 2013. Optical Interconnects for Future Data Center Networks. New York, NY: Springer-Verlag.
Stabile, R., Albores-Mejia, A., Rohit, A., Williams, K.A., 2016. Integrated optical switch matrices for packet data networks. Microsystems & Nano-Engineering 2, 15042.
Strasser, T.A., Wagener, J.L., 2010. Wavelength-selective switches for ROADM applications. IEEE Journal of Selected Topics in Quantum Electronics 16 (5), 1150–1157.
Yeow, T.W., Law, K.L.E., Goldenberg, A., 2001. MEMS optical switches. IEEE Communications Magazine 39 (11), 158–163.
Zhang, Z., You, Z., Chu, D., 2014. Fundamentals of phase-only liquid crystal on silicon (LCOS) devices. Light: Science & Applications 3, 1–10.

Indium Phosphide Photonic Integrated Circuits

Yuliya Akulova, Lumentum, Milpitas, CA, United States

Why Do We Need Photonic Integration?

To overcome fiber transmission impairments at higher modulation rates – and to overcome speed limitations of optical and electrical components – optical transceiver architectures increasingly rely on transmission components with multiple functional elements (lasers, modulators, power splitters/combiners and wavelength multiplexers). For example, 100, 200, and 400 Gb/s optical transceiver architectures for the local area network applications combine 25 Gbaud on-off keying (OOK) or four-level pulse amplitude modulation (PAM4) with multiplexing four or even eight wavelengths into one single-mode fiber (LAN/MAN Standards Committee of the IEEE Computer Society, 2010). In the metro, and long-haul space, demand for increased line-card density and reduced power dissipation is being supported by 10 Gb/s SFP + transceivers, either operating at a fixed wavelength or tunable over the C-band. This application utilizes traditional OOK modulation format and requires an externally modulated laser to meet the requirements for dispersion tolerance. In addition, starting from 40 Gb/s, long-haul and metropolitan area telecom applications adopted advanced modulation formats, including polarization multiplexing and multilevel quadrature amplitude modulation (Winzer, 2012). A polarization-multiplexed advanced-modulation-format transmitter requires a laser and two quadrature modulators, typically realized using a nested Mach-Zehnder (MZ) architecture. Widespread deployment of the coherent technology in the datacenter interconnects and metro space requires development of new photonic components that are able to support cost-effective, compact, high efficiency pluggable coherent transceivers.

The requirements for increased optical transmission speed and complexity combined with stringent constraints on cost, size, and power dissipation can be met only by implementing photonic integration to replace several packaged components with one.

What Is Photonic Integration?

Generally speaking, photonic integration refers to any technology that combines two or more functional optical elements to replace several separately packaged components with one. Photonic integration can be categorized into several types: Monolithic integration combines multiple active and passive optical elements fabricated on a single semiconductor chip to form a Photonic Integrated Circuit (PIC). Heterogeneous integration refers to a technology that uses dissimilar materials to provide full set of optical functions (e.g., heterogeneous integration of InP-based epitaxial material with Si photonics enables optical gain function). Finally, hybrid integration co-packages several active optical components that are optically coupled using micro-optics or planar lightwave circuits. In practice, a combination of hybrid and monolithic integration is frequently used. As the title of this article suggests, we will focus on monolithic PICs technology in InP.

Monolithic integration provides a drastic reduction in system footprint, inter-element optical coupling losses, power dissipation, and assembly and test cost. To illustrate the reduction in size, **Fig. 1** compares optical components that are required to realize transmission function using discreet laser and modulator with the same functions integrated on an InP PIC. In addition, monolithic integration frequently results in improved reliability since no relative shifts of different optical elements are possible.

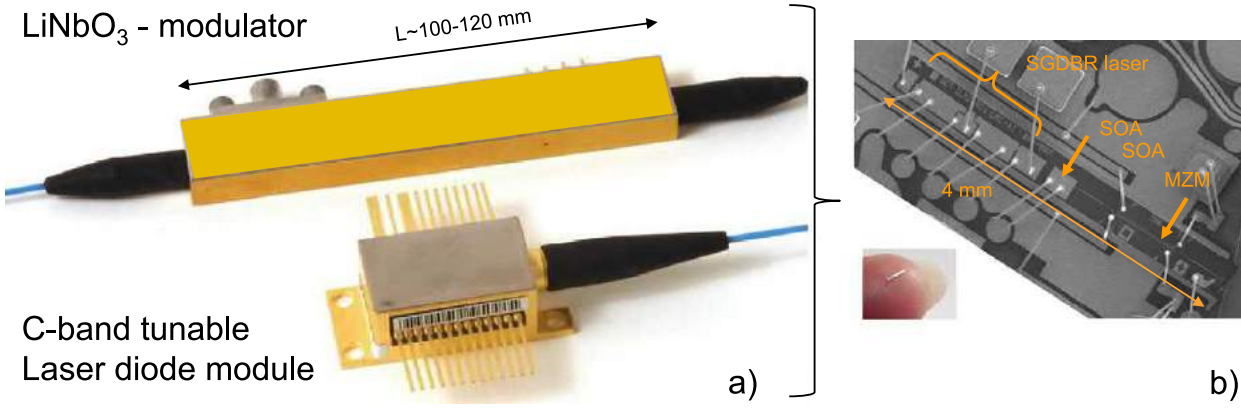


Fig. 1 Optical components required to realize 10 Gb/s C-band tunable transmitter: (a) C-band tunable laser diode module and Lithium Niobate modulator; (b) InP photonic integrated circuit with the same functionality – C-band tunable laser integrated with Semiconductor Optical Amplifier and Mach-Zehnder modulator.

 doi:10.1016/B978-0-12-803581-8.09770-8

Functional Requirements

The primary functions required to implement high-functionality PICs include (1) light-generation and amplification, (2) high-speed phase and/or amplitude modulation, (3) photodetection, and (4) optical routing. Optical routing functions may include only passive elements such as splitters, combiners, etc. and variety of functional elements that rely on the active refractive index tuning to adjust the optical phase. Fortunately, fundamental material properties of indium phosphide (InP) and related quaternary indium gallium arsenide phosphide (InGaasP) and indium aluminum gallium arsenide (InAlGaAs) compounds enable all required functionalities.

In most cases, all functional sections of an InP PIC are *p-i-n* heterostructure diodes, with i-waveguide core containing bulk quaternary material or multiquantum well (MQW) layers. In general, different materials have to be employed to optimize the performance for each of the functional elements as schematically illustrated in **Fig. 2**. Specifically, active regions of a high performance laser typically require several relatively narrow compressively strained QWs with the fundamental transition energy close to the target photon energy. The laser MQW active waveguide is typically surrounded by a separate confinement hetero-structure to maximize overlap between the optical mode and injected carriers. The upper and lower InP cladding regions are relatively heavily doped to minimize resistive heating. Electroabsorption (EA) and electro-optic MZ modulators in InP material systems fundamentally rely on the quantum-confined Stark effect, which produces red shift of the MQW absorption edge and change of the refractive index when reverse bias voltage is applied to p-i-n hetero-structure. In contrast to the laser, active regions of high efficiency, high speed EA and MZ modulators are blue shifted relative to the operating wavelength and typically require wider QWs to increase the efficiency of the quantum confined Stark effect. For EA modulators the difference between operating wavelength and modulator absorption edge wavelength, known as detuning, is chosen such that the absorption is negligible in the absence of reverse bias but increases rapidly as reverse bias is applied. MZ modulators are designed with larger detuning to achieve efficient index tuning without introducing excess loss. It is always desirable to be able to select the number of the QWs in a modulator active material for a specific performance targets: modulation bandwidth and drive voltage. The optical routing sections of the PIC require material with a bandgap larger than operating photon energy and low doping to minimize propagation loss. They can employ bulk or MQW material. The refractive index tuning can be accomplished through carrier injection mechanism by forward biasing the p-i-n diodes, voltage induced refractive index change under the reverse bias as outlined above, or thermal index tuning mechanism using local resistive heaters. Finally, light detection components require materials with a bandgap smaller than the photon energy to ensure efficient optical absorption.

To ensure high performance of the PIC based devices an integration technology has to provide flexibility in design optimization for different functional elements, efficient optical coupling between them and absence of parasitic reflections that can detrimentally affect laser wavelength stability, noise, and chirp. It is also desirable to minimize the transition regions between different functional sections. Ideally the spacing between different functional sections is selected based on the required electrical and thermal isolation and not constrained by the dimensions of transitions.

InP Integration Technologies

Active-Passive Axial Waveguide Integration

The compelling advantages and possibility of monolithic photonic integration in InP have been recognized from the early 1980s, and multiple approaches to integrate regions with different absorption/gain properties together along a single waveguide have been developed and are widely used in InP PIC fabrication (Coldren *et al.*, 2013).

Fig. 3 illustrates five most common techniques employed for monolithic integration of regions with different absorption/gain properties together along a single waveguide. For Offset Quantum wells **Fig. 3(a)** and twin waveguide **Fig. 3(b)** integration technologies different optical functions are separated into vertically displaced waveguides and the integration is realized by selective removal of the unwanted upper waveguides. Selective Area Growth (SAG) technology **Fig. 3(c)** relays on variation of the local material growth rates and hence the transition energy of the MQWs by changing the width of the pre-patterned dielectric masks before material growth. Quantum well intermixing technology **Fig. 3(d)** blue shifts the MQW region absorption edge by

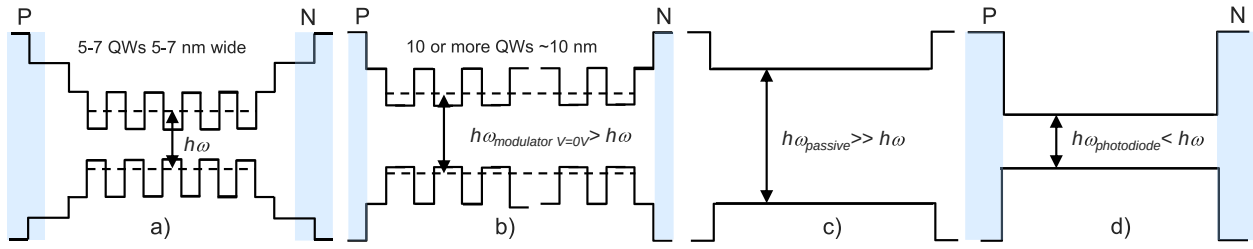


Fig. 2 Schematic band diagrams of the different functional sections of an InP PIC: (a) laser, (b) electro-absorption or Mach-Zehnder modulator, (c) passive waveguide, (d) photodiode.

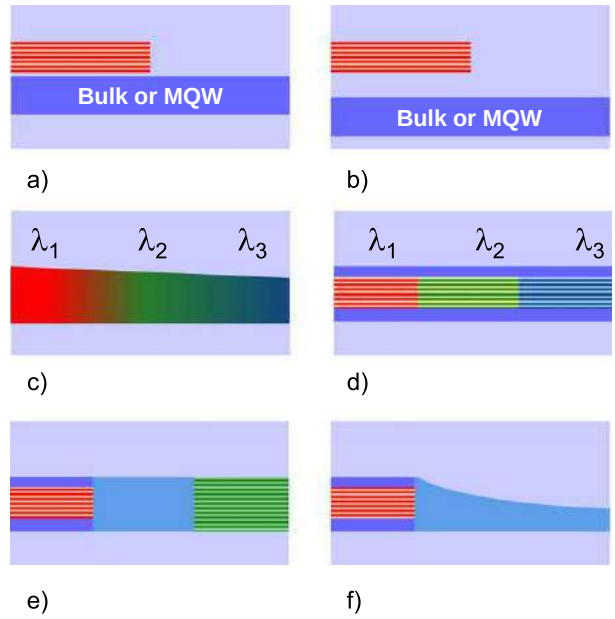


Fig. 3 Active/passive waveguide integration approaches: (a) Offset Quantum wells, (b) twin waveguide, (c) Selective Area Growth, (d) Quantum well intermixing, (e) Material regrowth, (f) Material regrowth combined with selective area growth.

impurity or point defect assisted atomic inter-diffusion between the wells and barriers at a high annealing temperature. In the Material regrowth technology **Fig. 3(e)**, different optical functions are created by sequential material removal and growth in different regions of the wafer. Several of the integration technologies can be further combined to achieve desired functionality. For example, as illustrated in **Fig. 3(f)** material regrowth can be combined with SAG by tailoring the dielectric mask shape and thus local waveguide thickness to match laser or modulator waveguide on the left and thin core waveguide for a spot-size converter on the right.

The 'offset quantum-well' technology and its variance, the 'dual-quantum-well' structure is the most simple integration approach. In this technology the waveguide is subdivided into a wider band-gap section at the bottom and a narrower bandgap MQW stack to be used as a gain section for a laser or an SOA at the top. The bottom section of the waveguide can be made of bulk quaternary alloy or contain a MQW stack embedded in the bulk waveguide. These closely spaced sections are grown in one epitaxial run and form single waveguide that supports only fundamental mode. The gain MQW material is selectively removed in the regions intended for modulators and passive waveguides and, if required, grating is etched into the passive waveguides. A single blanket regrowth of the InP cladding and InGaAs contact layer completes the manufacturing process. Offset quantum-well technology results in relatively low gain confinement factor that limits performance of lasers and SOAs with exception of booster amplifiers where low confinement factor is desirable for achieving high saturation power. However, this technology is very attractive due to its simplicity and manufacturability and has been successfully employed to demonstrate a variety of high performance PICs including widely tunable lasers integrated with SOAs, EA and MZ modulators (see reference Coldren *et al.* (2011) and references therein).

Another approach for monolithic photonic integration was developed based on twin-guide (TG) technology. As illustrated in **Fig. 3(b)**, a TG structure consists of active and/or passive devices formed in separate, vertically displaced waveguides. In difference from the offset quantum well structures, the waveguide layers in TG structures are spaced further apart. Such waveguide structure can support two optical modes (even and odd). Enhanced version of TG integration technology uses asymmetric twin guides (ATG) structure designed with different effective indexes of the active and passive waveguides. Light is transferred between the waveguides via lateral, adiabatically tapered mode transformers, allowing different optical functions to be realized in the different waveguides. The design of ATG structures and adiabatic tapers were described in details in Xia *et al.* (2005). The optimum design of the taper (shape and length) depends on the asymmetry of the ATG and the authors' analysis shows that using the optimal taper shape, a transfer efficiency of 90% can be achieved with the taper length of 290 μm. However, regardless of the fabrication simplicity, AGT technology carries significant drawbacks. These include long tapers, ~0.5 dB coupling loss between neighboring sections, and absorption in the gain tapers if left unpumped. Pumping can be extended to include the tapers but even then the gain and saturation power of the SOAs can be significantly compromised in the output taper. Alternative approach is to use QWI to blue-shift the absorption edge of the active material in the taper. Overview of the PICs that have been realized using ATG technology is provided in reference (Menon *et al.*, 2005). To list a few, these include EMLs, SOA/p-i-n detectors, integrated TG lasers, and integrated arrayed waveguide grating/p-i-n detector.

Quantum well intermixing (QWI) is an effective post-growth technique for increasing the bandgap energy of QWs in the desired locations on the wafer thus integrating laser gain with modulator and passive sections. QWI techniques has been used for

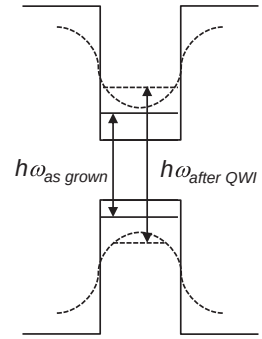


Fig. 4 Schematic representation of a QW potential profile as grown (solid) and after inter-diffusion of atomic species at QW/barrier heterointerfaces (dash).

the fabrication of various PICs in InGaAsP and InAlGaAs MQW materials and relays on inter-diffusion of atomic species at QW/barrier hetero-interfaces at elevated temperatures. Intermixing of the as-grown rectangular wells and barriers results in a graded potential profile and blue shift of the transition energy for the fundamental confined state for electrons and holes as schematically illustrated in **Fig. 4**. The inter-diffusion process is enhanced dramatically in the presence of point defects or vacancies in the interface region. This enhancement of the intermixing is used to provide spatial selectivity of the band gap shift that is required for monolithic integration. Several intermixing techniques including impurity induced disordering, impurity-free vacancy enhanced intermixing, and ion-implantation enhanced intermixing have been developed.

We will describe the latter two techniques in more details. In the vacancy enhanced QWI process the areas of the wafer that require wider bandgap are capped with a dielectric mask and the wafer is subjected to Rapid Thermal Annealing (RTA) typically at the temperature range of 700–800°C for approximately 30 s. The out-diffusion of group III elements at the semiconductor/dielectric interface create vacancies that rapidly propagate into QW region enhancing the rate of QW intermixing. Vacancy induced QW intermixing has been employed to achieve large bandgap shifts (>100 nm) in InGaAsP and AlGaAs MQW structures.

The most complex and high performance PICs were realized using implant-enhanced QWI process. This method has been employed for InGaAsP and InAlGaAs materials and produces good spatial resolution and controllable band-gap shifts using anneal time, temperature, and implant dose. Furthermore, any number of QW band edges can be achieved on the same wafer using selective removal of the catalyst.

We will describe implant-enhanced QWI process in more details using widely tunable sampled-grating DBR (SG-DBR) laser-EA modulator PIC reported in Skogen *et al.* (2005) as an example. In the first step, passive and modulator device regions are defined by selectively implanting phosphorous ($P+$) to introduce point defects into a sacrificial InP implant buffer layer grown above the MQW active region. Next, an RTA step is performed to shift the MQW band edge to that desired in the EAM regions (~ 35 nm). Once the EAM band edge is reached, the point defects are removed in the EAM region by selective removal of the InP implant buffer layer. The annealing is then continued until the desired band gap shift is reached for the tuning sections of the laser (~ 95 nm). Finally, blanket regrowth of the upper cladding and contact layer completes the formation of the waveguide in vertical direction.

Similarly to offset QW integration technology, implementation of PICs based on QWI requires only two epitaxial growth steps. However, QWI technology offers several significant advantages that include ability to place active QWs in the center of the waveguide and therefore maximize gain confinement factor, absence of multiple hetero-barriers and low doped material under the gain sections, ability to use band-gap shifted QWs in the carrier injection tuned phase sections therefor achieving higher index tuning efficiency due to band-filling effect in the step-function density of states. Finally, since QWI does not change the average material composition in the optical waveguide but only slightly changes the compositional profile, the index discontinuity and mode shape miss match at the interface between adjacent sections are small. This results in very low optical coupling loss and eliminates parasitic reflections between adjacent functional sections.

The drawbacks of the QWI techniques include shared number of QWs throughout the PIC and requirement for careful calibration of the intermediate band-gap shifts. The intermixing must also be done in the absence of other rapidly diffusing species in the wafer, such as zinc. Hence, regrowth of the p-doped cladding has to be done after the intermixing step resulting in somewhat poor controllability of the doping profile just above the active regions of the laser and modulator.

SAG has been employed in photonic integration using Metal-Organic Chemical Vapor Deposition (MOCVD) for InGaAsP and AlGaInAs bulk and MQW materials. With SAG technology different band-gap materials can be defined simultaneously using single epitaxial growth with patterned dielectric mask. In the absence of deposition on the mask, extra material is available over the masked zone resulting in the lateral concentration gradient. The gradient leads to extra lateral diffusion of species around the mask and to a local enhancement of the growth rate in the vicinity of the mask. In addition, since the diffusion coefficient in the vapor phase and sticking rate constant to the wafer surface are different for different species, the growth enhancement is accompanied by composition, band-gap, and strain variation as shown schematically in **Fig. 5**. The region between the mask stripes is used for the laser section and has thicker MQW stack, narrower band-gap, and slight compressive strain. Note, that in the case of MQWs both the composition and QW width are contributing to the red shift of the effective band-gap. Band-gap shifts up to 150 nm for MQW

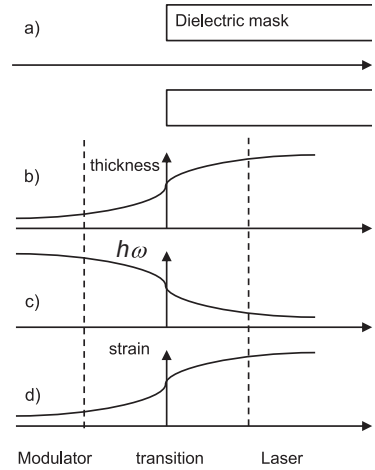


Fig. 5 Selective area growth: (a) top view of the mask layout and illustration of variation of waveguide thickness (b), material band-gap (c), and strain (d).

material and growth enhancement factor of ~2.5 can be achieved with appropriate design of the mask pattern. Examples of PICs fabricated using SAG include DFB-EA, tunable DBR-EA, wavelength selectable DFB arrays, and spot-size converters.

Three dimensional vapor phase diffusion models have been successfully developed and can accurately describe spatial variation of the thickness, composition, bandgap, and biaxial strain of the materials grown with different dielectric mask shapes. The models were validated by employing optical interferometer microscopy, high-resolution micro-photoluminescence, and micro X-ray. The models allow for constructive engineering of the waveguide thickness and composition through the manipulation of dielectric mask shape and also can predict undesired spatial variation that can occur due to neighboring mask cells in the case of high mask density.

The best applications for SAG technology include multi-wavelength laser arrays and spot size converters. Multi-wavelength laser arrays require adjustment of the MQW band-gap of 40 and 60 nm to cover C-band and CWDM O-band grid, respectively. Such band-gap engineering is well within SAG capability. In the case of a spot size converter both waveguide thickness and refractive index can be tapered to produce lossless interface with the rest of the PIC and the desired mode field diameter at the SSC output. In the case of a laser-electroabsorption modulator integration SAG can produce sufficient band-gap shift but limits freedom to optimize the material for each functional section. The design has to relay on the same number of QWs and doping throughout the PIC. In addition, SAG produces wider QWs for the laser and narrower QWs for the modulator imposing severe trade-off between laser and modulator efficiency and adds relatively long (~100 μm) transition regions. Such design trade-offs have proven to be acceptable for 2.5 and 10 Gb/s components. However, as the industry moves to 100 and 400 Gb/s transmission, more stringent requirements imposed on the modulation speed and PIC output power will require independent optimization of each functional section and therefore favor regrowth integration.

Among various integration technologies, regrowth integration is the only technology that enables flexibility in the design optimization of each PIC section with minimum transition length. In the regrowth integration technique, the laser gain layer stack is typically grown first on a planar substrate and then protected with a dielectric mask where it is required in the PIC. The exposed gain layers are then etched away, and the layers for the next functional section are grown. This process can be repeated several times to produce optimized layer stack for each PIC section. The primary advantage of the regrowth integration technology is that the layer composition (bulk or MQW) and doping can be changed abruptly, and therefore optimum design can be independently selected for each of the functional sections. The coupling loss between the neighboring sections are primarily determined by vertical alignment of the waveguides which can be controlled to very high accuracy by using selective etch. Even with the perfect vertical waveguide alignment some band-gap engineering is usually required to improve mode overlap and index matching at the transition. Well-designed regrowth transition can consistently produce less than 0.1 dB loss.

InP Waveguides

After the active-passive waveguides are defined along the light propagation direction lateral waveguide geometry has to be chosen based on the desired functionality of the PIC elements. **Fig. 6** illustrates several choices that are available in InP technology. They are arranged from highest to lowest index contrast as indicated in the figure. The deep- and shallow-ridge waveguides are formed by etching after all epitaxial steps are completed. The deep-ridge is fabricated using dry etch technique and due to high index contrast is suitable for bends with small radius of curvature. This type of waveguide is most frequently used as a routing guide for large scale PICs when multiple 90 degree bends, U-turns, and waveguide crossings are required to minimize the chip footprint. The deep-ridge waveguide is also the technology of choice for high-speed Mach-Zehnder modulators since in this geometry the active

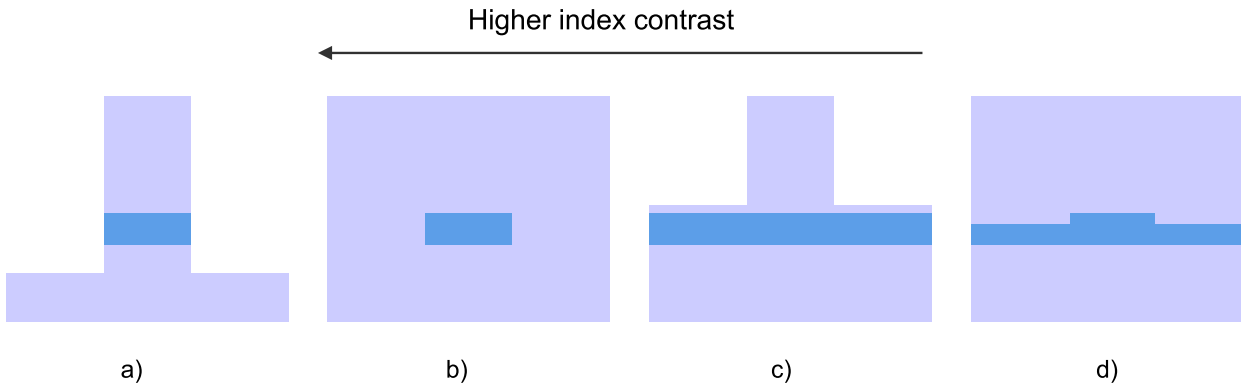


Fig. 6 Lateral waveguide cross-sections available in InP technology: (a) deep ridge, (b) buried channel, (c) shallow or surface ridge, (d) buried rib.

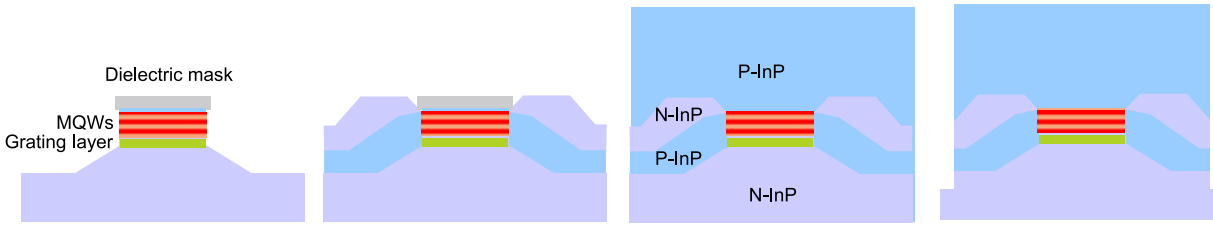


Fig. 7 Fabrication flow for planar buried hetero-structure laser.

region capacitance is determined only by the area of the waveguide without parasitic contribution of the electric field fringing effect in the surrounding material.

The shallow ridge waveguide (also called surface ridge) is fabricated through a combination of dry and selective wet etching. Selective wet etch produces very smooth sidewalls with the ridge profile depended on the crystallographic orientation. Due to the selectivity of the wet etch, the depth of the shallow ridge waveguide is controlled precisely. This type of lateral waveguide is a popular choice for many functional section of InP PICs. It is especially well suited for fabrication of CW lasers (FP, DFB, and DBRs), SOAs, and electro-absorption modulators. In contrast to MZ modulators, EAMs are usually short and very high modulation bandwidth can be attained even with larger junction capacitance of the shallow ridge structure.

The buried channel and buried rib waveguides must be defined prior to epitaxial regrowth. If such waveguides are intended for low loss routing they can be capped with thick undoped layer of InP. Buried channel waveguide is also used for manufacturing of high performance CW and directly modulated lasers (DMLs) and laser arrays by using PN or PNIN regrowth to block carrier flow around the active waveguide. Fig. 7 illustrates the fabrication sequence for a planar buried heterostructure laser. The process starts with the growth of the grating layer followed by the grating etch and regrowth of a spacer and MQW laser active. The lateral waveguide is formed using a combination of dry and wet etch and several layers of InP with different doping are re-grown with the mask still in place. Finally, the mask is removed and the top cladding and contact layers are grown. PN blocking structure is effective for CW lasers but produces very high parasitic capacitance, so Fe-doped InP i-region sandwiched between n-doped InP layers forming PNIN blocking structure is used as an alternative for manufacturing of high-speed DMLs or electro-absorption modulated lasers (EMLs). The advantages of this technology include low threshold current and high differential gain for the lasers due to reduced active region volume and the fact that carriers are confined to the center of the optical mode, relatively symmetric output beam, and good thermal properties. However, this technology adds critical regrowth step with regrowth interfaces positioned along the whole cavity length between MQW active and the blocking. Since the carries can diffuse rapidly in the QW layers, non-radiative recombination at these interfaces can be detrimental to laser performance and reliability. Excessive parasitic capacitance can limit modulation bandwidth even if PNIN blocking structure is used due to high dielectric constant of InP. Parasitic capacitance can be somewhat reduced by etching trenches to narrow the width of the blocking structure as shown in Fig. 7. Finally, it is difficult to engineer low loss, low reflection couplers between a buried heterostructure laser and deep-ridge routing waveguides and high-speed MZ modulators if such laser is used a part of a PIC. This is due to the significant mismatch in the optical mode sizes and effective refractive indexes and the fact that such transitions have to rely on the precise lithography since the mask used to fabricate buried waveguide section is removed prior to the regrowth of the cladding.

Transitions between different waveguide types are essential elements of the integration technology. Transition between deep- and shallow ridge waveguide is the most frequently used transition in large scale PICs that incorporate lasers and SOAs together with high speed MZ modulators and high functionality routing sections. Fortunately, since both types of waveguides are formed after regrowth of the top cladding, this transition can be fabricated using the same dielectric mask to define the waveguide along

the light propagation direction. In this self-aligned process, the coupling efficiency is not impacted by the accuracy of the lithographic alignment and, as described in reference (Coldren *et al.*, 2011), the waveguide mask can be tapered to ensure better overlap between the modes in the shallow and deep ridge sections. Transitions between buried ridge and other ridge types have to rely on an accurate alignment of the surface ridge waveguide mask to the underlying buried waveguide. However, even in this case coupling efficiency in excess of 95% with large manufacturing tolerance ($+/-0.3$ μm) can be obtained by tapering both waveguides at the junction.

Low loss InP waveguides

Losses in InP photonic waveguides are dominated by scattering loss and absorption provided that the material composition is chosen such that the contribution of the direct band-gap absorption is negligible. Smooth sidewalls are essential for achieving low loss single mode waveguides. As process control improves, the p-doped cladding becomes the most important contribution to loss through inter-valence band absorption. The n-type free carrier absorption is dominated by intra-valley transitions and requires interaction with phonon or ionized impurity scattering to provide conservation of momentum and thus much less significant than p-type absorption. This analysis implies that low loss waveguides on an InP PIC have to be fabricated with undoped cladding or p-doping has to be setback far enough to have negligible overlap with the optical mode. Optical propagation loss below 0.3 dB/cm has been demonstrated for shallow and deep ridge waveguides with doping setback. Excess loss below 0.1 dB per 90 degree turn with 100 μm radius of curvature can be readily achievable in deep ridge waveguides.

Since InP PICs typically include laser and SOA functions, special care has to be taken to suppress parasitic reflections that can occur due to abrupt changes in optical cross-sectional areas or/and effective index of the different components. Such reflections can detrimentally affect noise properties of the laser, introduce adiabatic chirp if reflection occurs after a modulator, and create gain ripples in the SOAs. Tilting the transitions with refractive index discontinuities and introducing tapered features at the transitions from shallow to deep waveguides and between single and multimode regions enables to maintain sufficiently low reflection level without introducing excess loss. Detailed description of splitters, polarization devices, microbends, and de/multiplexers developed for high-functionality InP PICs can be found in Williams *et al.* (2015).

Finally, ion implantation can be employed for electrical isolation between different functional sections of the chip. Hydrogen implant is typically used for passivation of p-doped material. Helium implant can be used for electrical isolation of n-doped material. In both cases the implantation dose and energy need to be tailored to provide the desired isolation without introducing excess optical loss.

We encourage the readers to follow up on the references provided above to learn more about high performance InP PICs developed in various integration platforms. In the last part of this article we will discuss design and performance for several InP PICs fabricated using regrowth integration. However, it is important to understand that the choices that PIC designers have to make are not limited to the integration technology but also include appropriate selection of the architecture (e.g., EA vs. MZ modulator) and additional functional elements that can be included in the PIC to further alleviate design trade-offs and increase fabrication tolerances, enhance functionality, and achieve sufficient margin relative to all performance specifications to enable high manufacturing yield and ultimately low cost. As we describe the PICs, we will point out why certain integration technology, device architecture, and features were adopted for each specific application.

Integrated (C-Band Tunable) Laser Mach-Zehnder Modulator (ILMZ) PIC

Development of ILMZ PIC drastically reduced footprint and power consumption of widely tunable optical transmitters and enabled transition of 10 Gb/s DWDM metro and LH transmission architecture from discreet implementations on a linecard or in 300 pin transponders into the realm of small form factor pluggable modules such as T-XFP and SFP+.

As illustrated in **Fig. 8(a)**, ILMZ PIC consists of a four-section SG-DBR laser, an SOA, and a MZ modulator. The SG-DBR laser is an efficient widely tunable laser source that can be easily integrated with other optical components. This laser was pioneered at UCSB and was used in a variety of high functionality PICs. An SG-DBR laser is a PIC in itself and consists of a MQW gain section

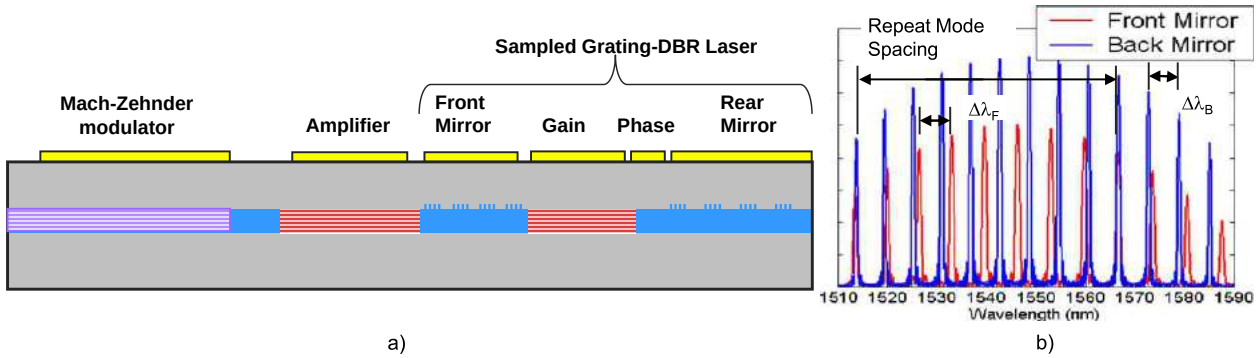


Fig. 8 (a) Schematic cross-section of ILMZ PIC. (b) reflectivity spectra of an SGDBR laser.

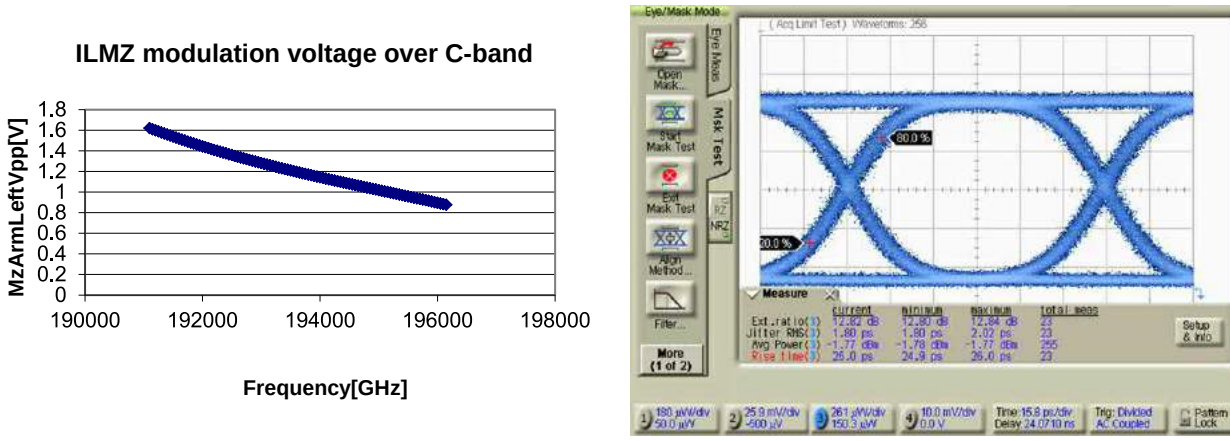


Fig. 9 Performance characteristics of ILMZ PIC based transmitter: (a) modulation voltage vs wavelength across C-band, (b) eye diagram at 10.7 Gb/s.

positioned between sampled-grating DBR mirrors. SG-DBR laser utilizes Vernier effect to achieve wavelength tuning in excess of 40 nm with only moderate index tuning in the mirrors. The detailed description of an SG-DBR laser is given in Coldren *et al.* (2004). In brief, by imposing additional periodicity on the hollographically defined grating the reflectivity spectra of the mirrors are transformed into combs of reflectivity peaks as shown in **Fig. 8(b)**. The spacing between adjacent peaks is inversely proportional to the grating sampling period. Selecting different sampling periods for the front and back mirrors ensure that only one set of reflectivity peaks can coincide for a particular set of mirror tuning currents resulting in lasing at the specified wavelength. The laser is tuned by adjusting currents in the mirrors to bring the desired reflectivity peaks into alignment. The phase section is used for alignment of the cavity mode to the peak of mirrors' reflectivity. Continuous tuning can be achieved by tuning both mirrors simultaneously. The specific design of SG-DBR laser implemented in ILMZ PIC uses carrier injection index tuning mechanism. Carrier injection is very efficient and fast refractive index tuning mechanism but it introduces cavity losses caused by free carrier absorption. Therefore it is highly desirable to integrate an SOA as an independent power adjustment knob. The integrated SOA compensates modulator loss and cavity losses caused by free carrier absorption in the tuning sections and allows wavelength independent power leveling, beam blanking during wavelength switching, and variable optical attenuator functionality.

Finally, MZ modulator is selected since the device has to operate over full C-band with high extinction ratio and good control of the chirp. Chirp of a differentially driven InP MZ is determined by the build in phase shift and split ratio and thus can be optimized (negative or zero) for the target application.

The PIC is fabricated using regrowth technology for lossless integration of the laser and SOA MQW active regions with the bulk quaternary mirror and phase sections, and independently optimized MQW MZ modulator section. Shallow ridge waveguide is used throughout the PIC. As described above, shallow ridge waveguide is the most simple and robust choice among the possible waveguide structures. From manufacturability point of view, shallow ridge is optimum technology choice for applications that require a single MZ with a moderate bandwidth. **Fig. 9** shows some of the key performance characteristics of ILMZ based optical transmitter. Notably, the differentially driven lumped MZ modulator requires less than 1.6 Vpp of modulation voltage per arm. Such low drive voltage is compatible with low power dissipation modulator drivers and is another very important factor that enabled full C-band tunable pluggable modules.

Narrow Linewidth High Power SG-DBR Laser

As described in the introduction, the LH, metro, and datacenter interconnects adopted coherent transmission format and continue to evolve towards higher bit rates by increasing the baud rate and the level of QAM modulation. This trend results in more stringent requirements imposed on the output power of the tunable lasers (>16 dBm) and, even more importantly, on the laser linewidth (<100 kHz). As discussed above, performance of widely tunable SG-DBR lasers with carrier injection tuning was adequate to meet full system requirements for 10 Gb/s OOK applications but additional tuning-induced loss is a challenge preventing injection tuned widely-tunable lasers from achieving narrow linewidth across the C-band. To overcome this limitation, Larson *et al.* (2015) developed thermally tuned SG-DBR laser and employed thermal engineering tailoring the impedance of the tuning sections to minimize power consumption and maintain millisecond-timescale tuning speed.

In addition, flexible-grid and colorless/directionless/contention-less network architectures employing power-combined launch require transmission lasers with high spectral purity in order to minimize noise interference on other channels. In particular, side mode suppression ratios (SMSR) greater than 47–50 dB are needed. SMSR of monolithic widely tunable DBR lasers integrated with SOAs is typically limited to 40–45 dB due to gain tilt and amplification of the spontaneous emission generated by the SOA as it is reflected from the back mirror. Addition of the thermally tuned broad-band filter into the PIC enabled to overcome this limitation, thereby simultaneously achieving >50 dB SMSR and <70 kHz linewidth at $+17$ dBm fiber coupled output power across C-band.

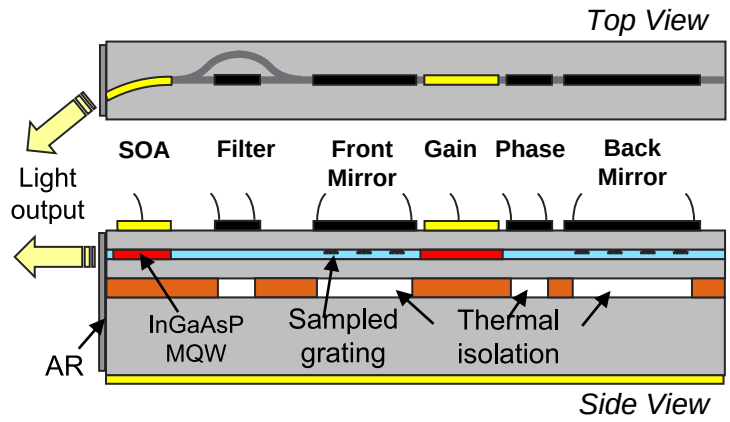


Fig. 10 (a) Schematic layout and cross-section of thermally tuned SG-DBR laser with integrated filter and SOA, (b) Instantaneous linewidth measure for 104 ITU channels at 16.5 and 17.5 dBm fiber coupled power.

The thermally tuned SG-DBR laser PIC comprising of the laser, delay interferometer tunable filter, and SOA is schematically shown in **Fig. 10**. The device uses regrowth integration technology similar to the injection tuned SG-DBR described above with the micro-heaters deposited above the shallow ridge structure in all tuning sections. The thermal impedance engineering is accomplished by placing sacrificial InGaAs layer underneath the entire PIC. As illustrated in **Fig. 10**, the sacrificial layer is selectively removed through lateral undercut etching in the thermally tuned sections. These section become thermally isolated from the substrate and the heat generated in the micro heaters must flow laterally before reaching the substrate through unetched InGaAs. For all other sections the InGaAs remains as a spacer layer through which heat can flow vertically to the substrate/heat sink.

The thermally-tuned SGDBR laser described above is high performance optical source for high bit rate coherent transmitters and local oscillators and is widely adopted in coherent links as a discreet component utilized in combination with stand-alone modulators and coherent receivers. In addition, this all- monolithic laser design lends itself to integration with high-speed IQ modulators to produce high performance fully monolithic dual polarization coherent transmitter PIC.

Dual Polarization IQ Transmitter PIC

A polarization-multiplexed multilevel quadrature amplitude modulation transmitter requires a narrow linewidth laser and two IQ modulators. Compact IQ modulators with nested Mach-Zehnder (MZ) architecture have been successfully demonstrated using InP and Si modulator technologies. Power consumption of such transmitter is directly proportional to the active load (laser and SOA drive currents, modulator photocurrent and termination) and the required modulation voltage. Since the dual-polarization IQ modulator requires four drivers the power efficiency of the transceiver can be significantly improved by lowering modulation voltage.

Electro-optic MZ modulators in InP material systems fundamentally rely on the quantum-confined Stark effect, which produces red shift of the multi-quantum well (MQW) absorption edge and change of the refractive index when reverse bias voltage is applied to p-i-n hetero-structure. The modulation efficiency can be enhanced by optimizing the material band-gap, width of the QWs and using the material with larger conduction band discontinuity such as InAlGaAs. Modulator design optimization also includes the device length, depletion region width and electrode configuration to achieve desired modulation bandwidth. For a 32Gbaud MZ modulator the V_π voltage is typically below 2 V across C-band making InP modulator technology extremely attractive for coherent applications.

Another important consideration for minimizing transmitter power consumption and enabling high optical output power is low loss optical coupling between the laser and modulator. This can be accomplished only by monolithic integration of a narrow linewidth laser with an IQ modulator in a single PIC. Monolithic InP coherent PICs have been developed in multi-channel (Evans *et al.*, 2011) and tunable configurations (Binetti *et al.*, 2012; Akulova, 2016). In this article we will review the design, integration technology, and performance of the PIC reported in Akulova (2016).

The monolithic InP PIC is shown schematically in **Fig. 11**. The chip integrates C-band thermally tuned narrow linewidth SG-DBR laser with two IQ MZ modulators and three semiconductor optical amplifiers (SOAs). Light exiting the front mirror of the laser is split into three tributaries for X, Y, and Local Oscillator (LO) optical ports. Each optical path integrates an SOA for independent optical power leveling across C-band and variable optical attenuation functionality. The IQ modulators are realized using conventional nested MZ architecture. The signal electrodes of the IQ modulators are implemented using traveling wave design. Differential dc phase tuning sections are integrated into each of the IQ modulators and into individual MZs. The integration of multiple active and passive functional elements in one PIC has been accomplished using regrowth integration technology which is a platform technology for ILMZ and narrow linewidth SG-DBR laser PICs described above. The nested MZ modulator are fabricated using deep ridge waveguides. The performance characteristics of the monolithic DP-IQ transmitter PIC include V_π voltage of less than 2 V and the bandwidth in excess of 38 GHz (**Fig. 12**).

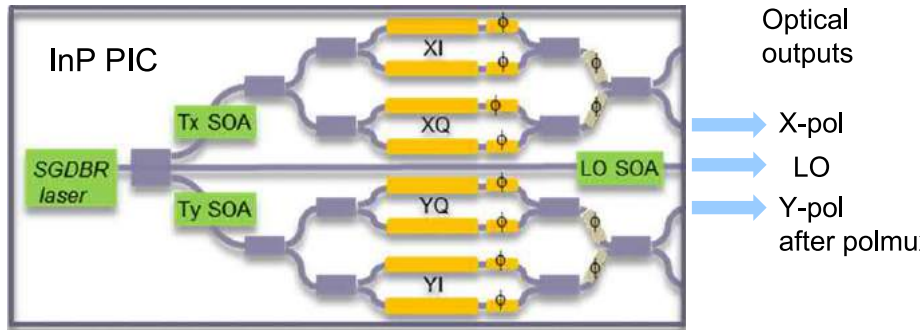


Fig. 11 Schematic layout of monolithic InP PIC consisting of thermally tuned narrow linewidth C-band tunable SGDBR laser, dual-polarization IQ modulators, and three SOAs.

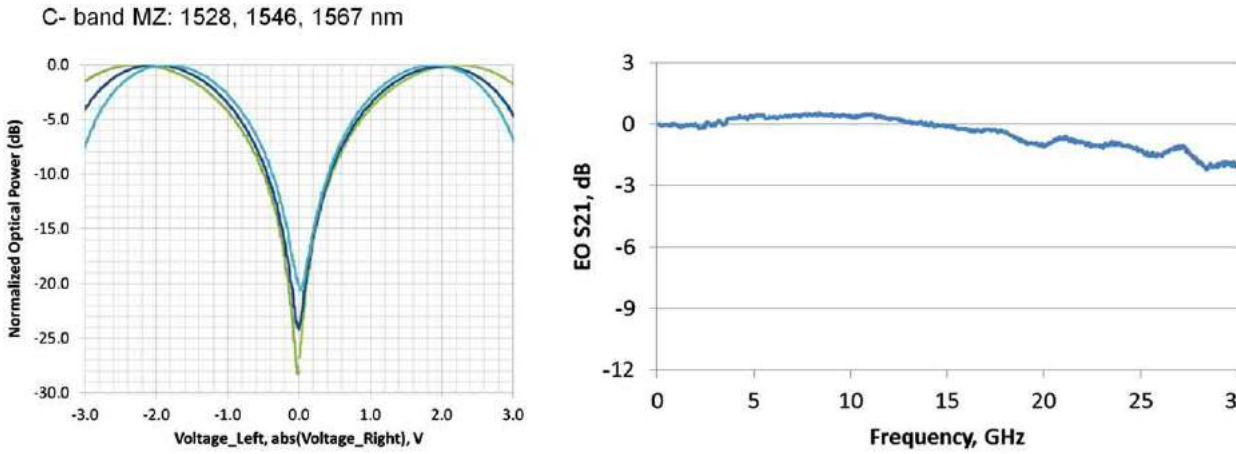


Fig. 12 Performance characteristics of monolithic widely tunable coherent transmitter PIC: (a) dc transfer function of MZ modulator measured at several wavelengths across C-band; (b) small signal modulation bandwidth.

High-efficiency of InP MZs modulators combined with the capability of lossless monolithic integration with tunable lasers and SOAs positions InP PIC technology as a primary candidate for realization of cost-effective, compact, high-efficiency optical transmitter engines for pluggable coherent modules for 100 and 200 Gb/s and thus enabling wide deployment of coherent transmission in metro and datacenter interconnects links. With the potential for higher bandwidth, high modulation efficiency, and increased scale of photonic integration, the InP PIC technology will continue to scale enabling compact, cost effective solutions for 400 Gb/s and beyond.

Electro-Absorption Modulated Lasers

An EML – the smallest scale InP PIC- consists of a DFB laser and electro-absorption modulator. EMLs have been extensively used in DWDM and ZR metro links 2.5 and 10 Gb/s for decades. They have been commercialized in 100 Gb/s transceivers starting from 2014 and currently being designed into 200 and 400 Gb/s small form factor pluggable transceivers. Wide adoption of the EMLs for these high bit rate application is driven by the transceiver architectures and link budget requirements.

Transmission component architectures adopted by IEEE for the inter-datacenter optical links (500 m to 2 km) and local area networks (up to 40 km) at 100 Gb/s combines 25 Gb/s OOK modulation with multiplexing four wavelengths into one single-mode fiber. These links utilize O-band wavelength range in the minimum fiber dispersion window. Similarly, standards that are currently in development for 200 and 400 Gb/s transmission rely on four and eight wavelengths, respectively, with 50 Gb/s pulsed amplitude modulation (PAM4) format which does not require increase in modulation bandwidth relative to 100 Gb/s architecture. Alternatively, 400 Gb/s transmission can be implemented with only four wavelength using PAM4 modulation at 100 Gb/s per wavelength. In all the cases implementation of components has to fit into power dissipation budget and footprint of small form factor pluggable modules such as QSFP28 and QSFP-DD. These links are un-amplified and have to support 6–18 dB link loss budget (including fiber and connectors loss and transmitter and dispersion penalty) resulting in stringent requirements imposed on transmitter output power and extinction ratio.

Links in the minimum fiber dispersion window relax the requirement on transmitter chirp and can be implemented using directly modulated lasers (DMLs) and EMLs. Several high-performance DMLs and DML array integrated with multiplexer PICs

have been reported (see, e.g., Kanazawa *et al.* (2016) and Matsui *et al.* (2016) and references therein). However, it is advantageous to separate light generation and modulation functions to enable uncooled operation with large margin relative to all performance specifications and therefor high manufacturing yield. EML based transmitters enable independent optimization of the laser for the target output power and low relative intensity noise while EA modulator can easily meet the bandwidth and ER requirements and offers superior dispersion tolerance as compared to DMLs, and therefor can be used up to 40 km applications.

Fig. 13 presents schematic x-section of an integrated DFB laser with a high speed EA modulator. The EML is fabricated using regrowth integration and shallow ridge technology. With exception of the modulator MQW regrowth step, the fabrication process for an EML is similar to the flow for the fabrication of a stand alone DFB laser. Interestingly, the fabrication process for 25 Gb/s DMLs is frequently more complex than for EMLs. In the effort to meet the requirements for modulation bandwidth, DMLs have to be designed with very small active volume and the fabrication might include buried heterostructure process and integration of passive sections to maintain manufacturability as the laser cavity length is reduced to below 150 μm.

EMLs are compact and can be co-packaged or monolithically integrated with multiplexers to realize small form factor transmitter optical sub-assemblies (TOSAs) compatible with QSFP28 transceivers. **Fig. 14** shows typical transfer functions of an EA

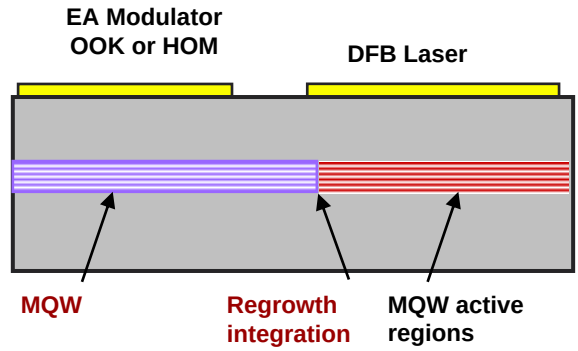


Fig. 13 Schematic cross-section of an EML chip.

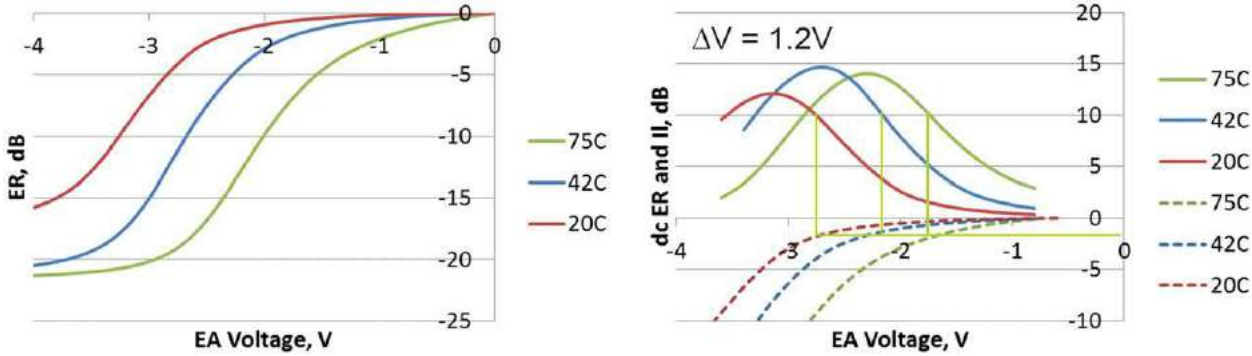


Fig. 14 (a) Transfer function of uncooled EML; (b) Extinction ratio vs insertion loss for fixed modulation voltage (V_{pp} = 1.2 V).

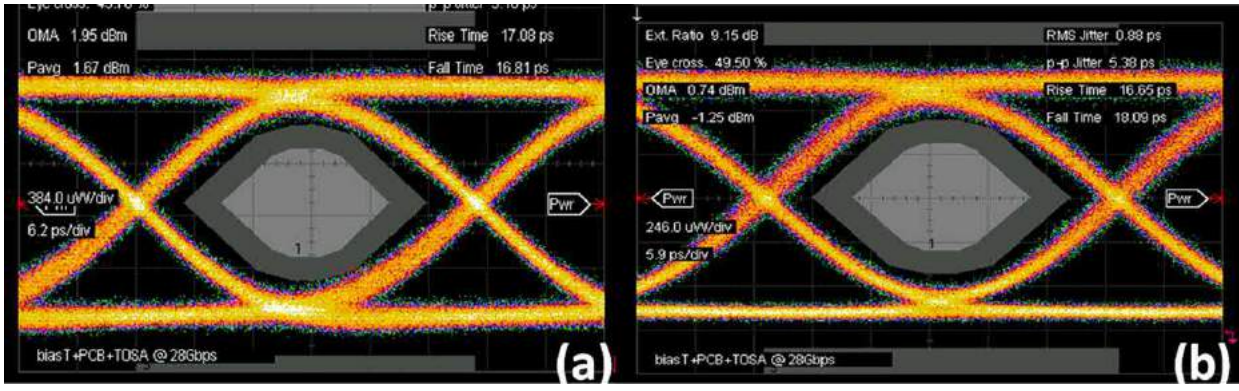


Fig. 15 Eye diagrams measured at 28 Gb/s: (a) ER > 5 dB, MM > 40%, Vpp = 1.2 V; (b) ER > 9 dB, MM > 40%, Vpp = 1.5 V.

modulator measured at several temperatures and ER versus insertion loss calculated from the measured transfer function by changing dc bias with fixed modulation voltage. The ER in excess of 10 dB with insertion loss below 2 dB is demonstrated with 1.2 V modulation voltage across wide temperature range. The appropriate bias and modulation voltage is selected based on the target extinction ratio during transceiver calibration. **Fig. 15** presents eye diagrams measured at 28 Gb/s. Extinction ratio in excess of 5 and 9 dB is demonstrated with the modulation voltage of 1.2 and 1.5 V, respectively. Such low requirements on the drive voltage ensure compatibility with low power dissipation modulator drivers and improve the power budget of the transceivers.

In summary, we have reviewed available InP integration technologies and discussed how optical link performance requirements drive the selection of PIC functional sections and integration approaches. We presented design, integration technology, and performance for several InP PICs that have been commercialized in Datacom, Metro, and LH transmission systems. InP PICs performance continues to scale to deliver higher transmission capacity at lower cost and power dissipation. In addition, InP technology enables integration of wide range of photonic functions and impacts broad range of market segments including photonic solutions for sensing, imaging, and high-speed signal analysis.

References

Akulova, Y., 2016. Advances in integrated widely tunable coherent transmitters. In: Proceedings of Optical Fiber Communications Conference W4H.1, IEEE.

Binetti, P.R.A., Lu, M., Norberg, E.J., *et al.*, 2012. Indium phosphide photonic integrated circuits for coherent optical links. Journal of Quantum Electronics 48, 279–291.

Coldren, L.A., Corzine, S.W., Mashanovitch, M.L., 2013. Diode lasers and photonic integrated circuits, second ed. John Wiley & Sons.

Coldren, L.A., Fish, G.A., Akulova, Y., *et al.*, 2004. Tunable semiconductor lasers: A tutorial. Journal of Lightwave Technology 22, 193–202.

Coldren, L.A., Nicholes, S.C., Johansson, L., *et al.*, 2011. High performance InP based photonics ICs – A tutorial. Journal of Lightwave Technology 29, 554–570.

Evans, P., Fisher, M., Malendevich, R., *et al.*, 2011. Multi-channel coherent PM-QPSK InP transmitter photonic integrated circuit (PIC) operating at 112 Gb/s per wavelength. In: Proceedings of Optical Fiber Communications Conference PDPC7, IEEE.

Kanazawa, S., Kobayashi, W., Ueda, Y., *et al.*, 2016. 30-km Error-free transmission of directly modulated DFB laser array transmitter optical sub-assembly for 100-Gb application. Journal of Lightwave Technology 34, 3646–3652.

Larson, M.C., Bhardwaj, A., Xiong, W., *et al.*, 2015. Narrow linewidth sampled-grating distributed bragg reflector laser with enhanced side-mode suppression. In: Proceedings of Optical Fiber Communications Conference M2D.1, IEEE.

LAN/MAN Standards Committee of the IEEE Computer Society, 2010. IEEE Std 802.3ba-2010 and IEEE Std 802.3bs- draft. Available at: www.ieee802.org.

Matsui, Y., Pham, T., Sudo, T., *et al.*, 2016. 28-Gbaud PAM4 and 56-Gb/s NRZ performance comparison using 1310-nm Al-BH DFB laser. Journal of Lightwave Technology 34, 2677–2683.

Menon, V.M., Xia, F., Forrest, S.R., 2005. Photonic integration using asymmetric twin-waveguide (ATG) technology: Part II – Devices. IEEE Journal of Selected Topics in Quantum Electronics 11, 30–42.

Skogen, E.J., Raring, J.W., Morrison, G.B., *et al.*, 2005. Monolithically integrated active components: a quantum-well intermixing approach. IEEE Journal of Selected Topics in Quantum Electronics 11, 343–355.

Williams, K.A., Bente, E.A.J.M., Heiss, D., *et al.*, 2015. InP photonic circuits using generic integration. Photonics Research 3, B60–B68.

Winzer, P.J., 2012. High-spectral-efficiency optical modulation formats. Journal of Lightwave Technology 30, 3824–3835.

Xia, F., Menon, V.M., Forrest, S.R., 2005. Photonic integration using asymmetric twin-waveguide (ATG) technology: Part I – Concepts and theory. IEEE Journal of Selected Topics in Quantum Electronics 11, 17–29.

CMOS Transceiver Circuits for Optical Interconnects

Samuel Palermo, Texas A&M University, College Station, TX, United States

Optical Interconnects

Optical interconnects for computing systems use electronic driver circuitry to transmit digitally-modulated light from a source laser over an optical channel, most commonly an optical fiber, to a photodetector where the signal is converted back into the electrical domain with a receiver circuit. Relative to electrical interconnects, these optical communication systems overcome key interconnect bottlenecks and greatly improve data transfer efficiency due to the optical fiber's flat channel loss over a wide frequency range and also relatively small crosstalk and electromagnetic noise. Another important feature of optical interconnects is the ability to combine multiple data channels on a single waveguide via wavelength-division-multiplexing (WDM) and greatly improve bandwidth density. These key features have motivated the application of optical interconnect systems in high-performance computing and data center systems where transmission distances can exceed 1 km.

The dominant optical interconnect transmitter technology for these applications has been directly-modulated multi-mode vertical-cavity surface-emitting lasers (VCSELs) due to these devices being inexpensive and the relaxed alignment tolerances offered by the multi-mode fiber system. However, modal dispersion limits performance as distances reach the km-range and data rates climb above 25 Gb/s. This motivates single-mode solutions that allow for increases in both transmission distances and bandwidth density via WDM. These single-mode system often utilize external modulation of a continuous-wave (CW) laser operating at a constant power level in order to overcome laser bandwidth limitations and optical spectrum broadening due to modulated laser chirp. Common external modulator devices include electroabsorption modulators (EAMs) and refractive modulators, such as the Mach-Zehnder and ring resonator modulators.

At the receive side, a photodiode is typically utilized to sense the high-speed optical power and produce an input current. This photocurrent is then converted to a voltage and amplified sufficiently for data resolution. In order to support high data rates, sensitive high-bandwidth photodiodes are necessary. High-speed p-i-n photodiodes are typically used in optical receivers due to their high responsivity and low capacitance. In these device structures, the light is absorbed in the intrinsic region and the generated carriers are collected at the reverse bias terminals, thereby causing an effective photocurrent to flow. The amount of current generated for a given input optical power is set by the detector's responsivity. For devices where the light is normally incident, there is an inherent tradeoff between responsivity and bandwidth set by the intrinsic layer width. A wider intrinsic region will result in an increased amount of absorbed photons and higher responsivity, but will also cause increased carrier transit times that results in reduced detector bandwidth. This responsivity-bandwidth tradeoff is broken in waveguide p-i-n structures where the light travels horizontally down the intrinsic region and the electric field is formed orthogonal. These structures allow for both a thin intrinsic region for short carrier transit times and a sufficiently long region for high responsivity.

Energy efficient optical interconnect systems require low-power and area circuits that interface to these optical sources and detectors. This article focuses on CMOS driver and receiver circuits, as this technology allows for low-voltage and low-power operation and the potential to integrate the optical front-ends with the main data processing integrated circuits (ICs). While high-performance silicon-germanium (SiGe) or other compound semiconductor technologies may offer superior performance relative to CMOS implementations, these processes are often more expensive and/or not competitive in terms of energy efficiency for short distance optical interconnect applications.

Transmitter Circuitry

Directly modulated lasers and optical modulators, both electroabsorption and refractive, have been proposed as high bandwidth sources for optical interconnects, with these different sources displaying tradeoffs in both device and circuit driver efficiency. Vertical-cavity surface-emitting lasers are an attractive candidate due to their ability to directly emit light with low threshold currents and reasonable slope efficiencies; however their speed is limited by both electrical parasitics and carrier-photon interactions. A device which doesn't display this carrier speed limitation is the electroabsorption modulator, based on either the quantum-confined Stark effect (QCSE) or the Franz-Keldysh effect, which is capable of achieving acceptable contrast ratios at low drive voltages over tens of nm optical bandwidth. Ring resonator modulators are refractive devices that display very high resonant quality factors and can achieve high contrast ratios with small dimensions and low capacitance, however their optical bandwidth is typically less than 1 nm. Another refractive device capable of wide optical bandwidth (>100 nm) is the Mach-Zehnder modulator, however this comes at the cost of a large device and high voltage swings. All of the optical modulators also require an external source laser and incur additional coupling losses relative to a VCSEL-based link. The following describes CMOS circuitry optimized for these particular optical transmitter devices.

 doi:10.1016/B978-0-12-803581-8.09771-X

VCSEL Transmitters

A VCSEL, shown in Fig. 1(a), is a semiconductor laser diode which emits light perpendicular from its top surface. These surface emitting lasers offers several manufacturing advantages over conventional edge-emitting lasers, including wafer-scale testing ability and dense 2D array production. The most common VCSELs are Gaas-based operating at 850 nm, with longer wavelength devices manufactured in various material structures also available that operate near 1000, 1310, and 1550 nm. Current-mode drivers are often used to modulate VCSELs due to the device's linear optical power-current relationship. A typical CMOS VCSEL output driver is shown in Fig. 1(b), with a differential stage steering current between the optical device and a dummy load, and an additional static current source used to bias the VCSEL sufficiently above the threshold current, I_{TH}, in order to ensure adequate bandwidth.

Total VCSEL bandwidth is limited by a combination of electrical parasitics and the electron-photon interaction dynamics. The laser diode's dominant electrical time constant comes from the bias-dependent junction *RC*. In addition to the bias-dependent junction resistance, there is also significant series resistance due to the large number of distributed Bragg reflector (DBR) mirrors used for high reflectivity. VCSEL optical bandwidth is regulated by two coupled differential equations which describe the electron-photon interaction. Derived from these rate equations, the VCSEL relaxation oscillation frequency ω_R, which is proportional to the effective bandwidth, is directly proportional to the square-root of the injected current above the threshold current I_{TH}.

$$\omega_R \propto \sqrt{I - I_{TH}} \tag{1}$$

Output power saturation due to self-heating and also device lifetime concerns restrict excessive increase of VCSEL average current levels to achieve higher bandwidth. VCSEL reliability potentially poses a series impediment to very high speed modulation, as the mean time to failure (MTTF) is inversely proportional to the current density squared. The conflicting dependencies of VCSEL bandwidth and reliability on device current yield a very steep tradeoff.

In order to ease this tradeoff, equalizing output stages have been proposed to extend the data rate for a given average current. As an example, Fig. 2 shows a VCSEL transmitter with a four-tap finite impulse response (FIR) equalizer consisting of one pre-cursor, one main, and two post-cursor taps implemented by summing current sources at the output node. Five parallel data bits, D[4:0], are routed to the taps, where they are shifted one bit time with respect to the clock phases to implement the necessary filter delays. At each tap, a multiplexer serializes the five parallel input bits and drives a differential output stage which steers current between the VCSEL and dummy diode-connected thick-oxide nMOS devices that are connected to a separate 2.8 V LVdd supply. This higher supply is necessary to support the 1.5 V VCSEL knee voltage. Eight-bit current mirror DACs bias the output stages to the desired current value. Because of the smaller current requirements of the pre/post-cursor taps, their muxes and output stages are set to one-fourth the size of the main tap to save power. A static DC current source, I_{DC}, is also used to bias the VCSEL above the threshold current to insure adequate bandwidth. This bias current and the leakage current from the tap driver transistors, I_{leak}, provide sufficient voltage drop across the VCSEL and dummy load to prevent excessive voltage stress on the output stage transistors. While the VCSEL's varying frequency response with current limits the performance of this linear equalizer for large extinction ratio modulation, the frequency response variations diminish with increasing average current due to the square-root relationship and a linear equalizer is effective in canceling intersymbol interference (ISI) for extinction ratios near 3 dB. In order to further improve performance, non-linear equalizers have been proposed with asymmetric current profiles for the different data transitions.

Electroabsorption Modulator Transmitters

An electroabsorption modulator is typically made by placing an absorbing quantum-well region in the intrinsic layer of a reverse-biased p-i-n diode. In order to produce a modulated optical output signal, light originating from a continuous-wave source laser is absorbed in an EA modulator depending on electric field strength through electro-optic effects such as the quantum-confined Stark effect or the Franz-Keldysh effect. These devices are often implemented as a waveguide structure where light is coupled in and

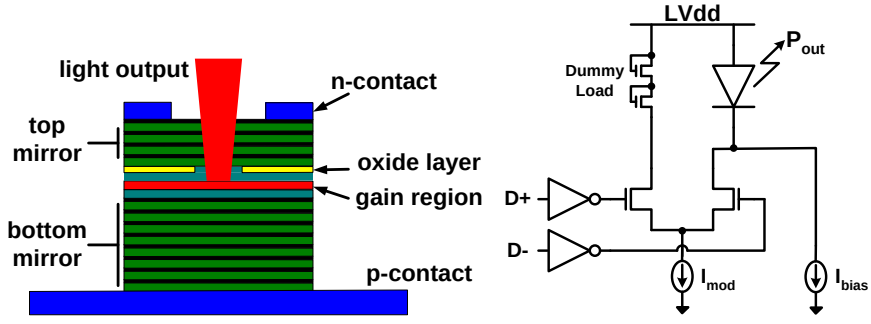


Fig. 1 Vertical-Cavity Surface-Emitting Laser: (a) device cross-section and (b) driver circuit.

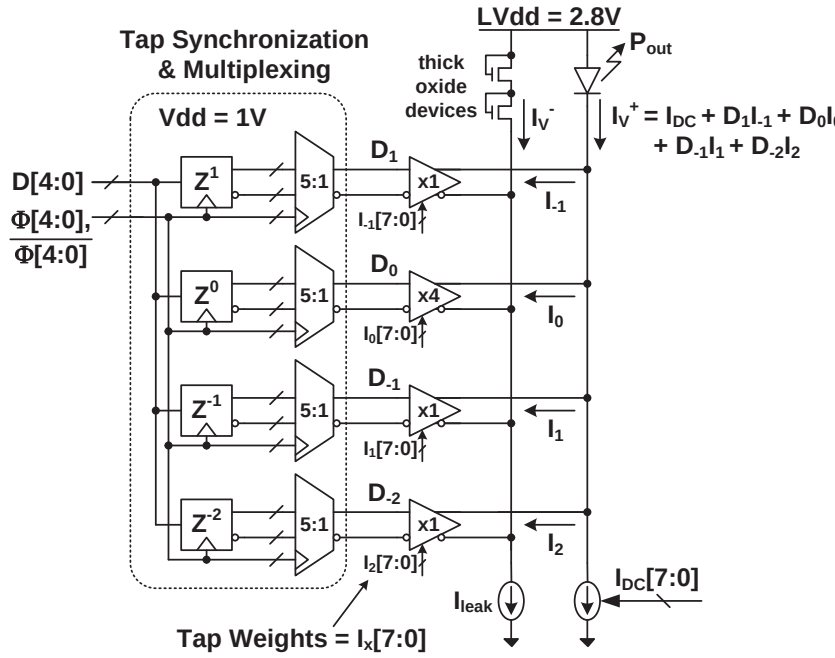


Fig. 2 VCSEL transmitter with a four-tap FIR equalizer.

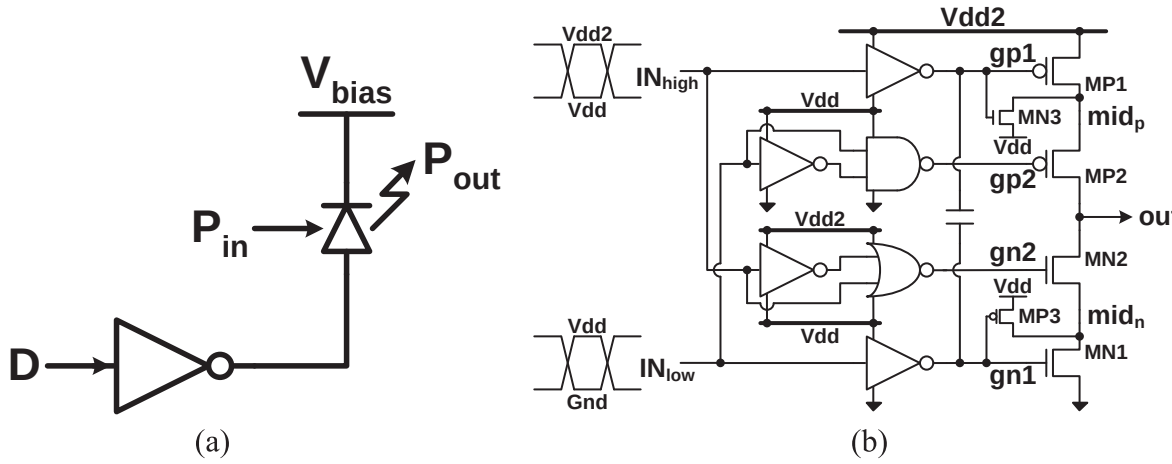


Fig. 3 Electroabsorption modulator drivers: (a) simple CMOS inverter and (b) high-swing pulsed-cascode output stage.

travels laterally through the absorbing multiple-quantum-well (MQW) region. Modulation is typically achieved by applying a static positive bias voltage to the n-terminal and driving the p-terminal with the high-speed signal.

Assuming tight integration with the driver circuitry, the EAM can be treated as a lumped-element device consisting of the diode capacitance and a parallel voltage-dependent photocurrent source. While conceptually the high-speed driver could be a simple CMOS inverter (Fig. 3(a)), the sub-1 V swings available in modern CMOS processes limit the achievable extinction ratio for relatively short lumped-element EAMs. Improved extinction ratios are possible with longer EAM device lengths. However, achieving high-speed operation with these longer devices necessitates traveling-wave driver topologies that often require low-impedance termination and a relatively large amount of switching current. Another option is to use a lumped element driver with a higher output swing than the nominal CMOS supply. The ability to drive the smaller devices as an effective lumped-element capacitor offers a huge power advantage when compared to longer controlled-impedance structures, as the CV^2f power is relatively low due to small device capacitance. One circuit which allows for this in a reliable manner is a pulsed-cascode driver, which offers a voltage swing of twice the nominal supply while using only core devices for maximum speed.

Fig. 3(b) shows the pulsed-cascode output stage which accepts both a "low" input IN_{low} that swings between the Gnd and the nominal chip Vdd and a "high" input IN_{high} with the same data value that has been level-shifted to swing between Vdd and Vdd2, where Vdd2 is nominally twice the voltage of Vdd. Static-voltage overstress is eliminated in the output-stage cascode structure by

equally splitting the output voltage across the series transistors. Pulsing the gates of the cascode transistors (MN2 and MP2) during transitions with NAND- and NOR-pulse gates, respectively, allows this driver to eliminate the transient drain–source voltage (V_{DS}) overstress present in static-biased cascode drivers and prevents transistor degradation from hot-carrier injection.

Ring Resonator Modulator Transmitters

As shown in Fig. 4(a), a ring resonator consists of a top waveguide which couples a portion of the incoming continuous wavelength (CW) light into a ring waveguide. At the resonance wavelength, most of the incident light will be trapped due to the interference between the ring and top waveguide. This results in minimal power at the through port, as shown in Fig. 4(b). Adding a second bottom waveguide to the ring allows for the light coupled into the ring from the input port to be coupled out onto the bottom waveguide at the resonance wavelength, implementing a drop filter. By changing the ring waveguide's effective refractive index via the plasma dispersion effect, the resonance wavelength can be shifted to allow for modulation. Fig. 4(c) shows that if the resonance wavelength is aligned with the incident light wavelength, an adequate output extinction ratio is possible. However, if the high-Q device's resonance is off due to temperature sensitivity or manufacturing variations, the performance greatly diminishes (Fig. 4(d)).

The two most common silicon ring resonator modulators, carrier-injection and carrier-depletion devices, are implemented by forming a junction across the ring waveguide. Carrier-injection devices have an intrinsic ring waveguide surrounded by p+ and n+ regions that form the modulator's two electrical terminals. As they operate primarily in forward bias, this allows for significant changes in the intrinsic-region carriers with varying drive signals. While this provides large refractive index changes and high modulation depths, dynamic operation is limited by long minority carrier lifetimes. A pn junction is formed directly in the ring waveguide in carrier-depletion modulators. Higher speeds relative to a carrier-injection ring are generally achievable due to the

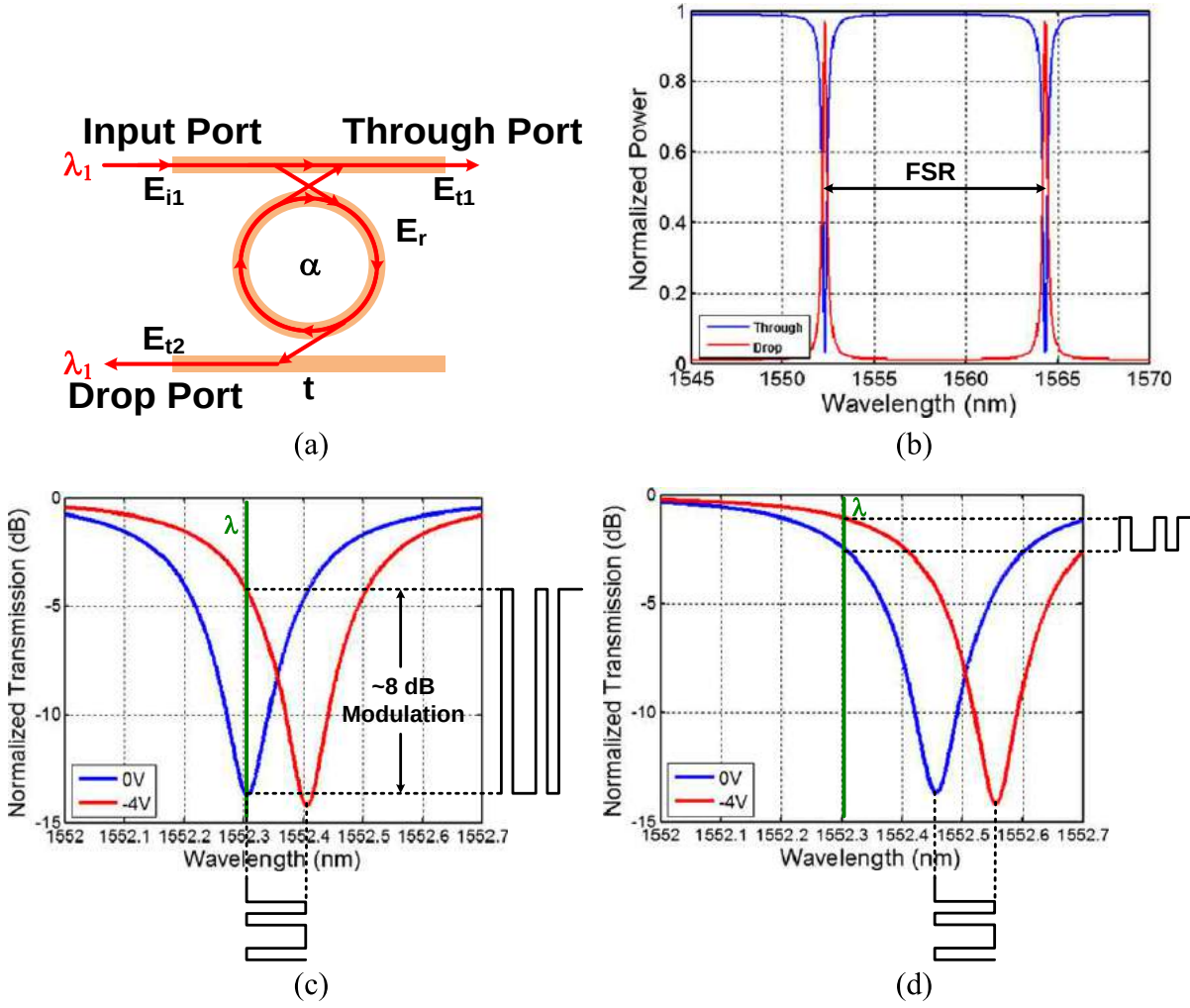


Fig. 4 Ring resonator device: (a) schematic, (b) through and drop port power versus wavelength, (c) modulator operation with 0 V bias resonance equal to the incident light wavelength, (d) modulator operation with a 160 pm resonance shift.

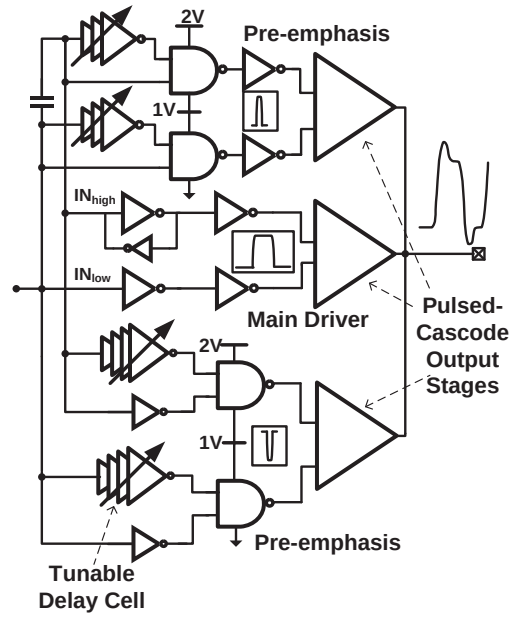


Fig. 5 Non-linear pre-emphasis ring modulator driver with $2V_{pp}$ output stage.

ability to rapidly change the depletion width. However, the modulation depth is limited due to the low doping concentration in the waveguide to avoid excessive loss.

The differing characteristics of carrier-injection and carrier-depletion ring modulators necessitate variations in the output stage design to enable high data rate operation. While carrier-injection modulators are capable of high extinction ratios, the operating speed is limited by relatively slow carrier dynamics in forward-bias and parasitic contact resistance in reverse-bias. This necessitates relatively large amounts of non-linear pre-emphasis to compensate for the differing time constants during rising and falling transitions. A single-ended driver optimized for carrier-injection modulators is shown in Fig. 5. This design utilizes a pulsed-cascode output stage for robust high output swing and segments this output stage into main and independent rising/falling edge pre-emphasis drivers. Here the per-edge pre-emphasis strength is set with tunable delay elements that allow optimization of the transient response to be decoupled from the steady-state extinction ratio value. A 65 nm CMOS implementation of this design achieved an output swing of $2V_{pp}$ and a 9.2 dB extinction ratio at 9 Gb/s when driving a carrier-injection ring modulator.

Depletion-mode modulators have shorter carrier lifetimes and can easily achieve data rates >10 Gb/s. However, they typically exhibit low pn junction tunability that necessitates higher modulation voltages, require negative DC-biasing for high-speed operation, and need pre-emphasis optimized for non-linear optical dynamics important at data rates >20 Gb/s. One circuit which can achieve this is a differential version of the Fig. 5 driver with on-die AC-coupling capacitors that allow for a constant negative-bias operation over large-swing dynamic operation. A 65 nm CMOS implementation of this design achieved $4.4V_{pp}$ differential swing and implemented asymmetric 2-tap equalization in the output stage to operate at 25 Gb/s with a 7 dB ER.

In order to ensure reliable operation over temperature and fabrication variations, resonance wavelength tuning can be achieved by adjusting the ring resonator's refractive index thermally, electrically, or by employing both techniques. Thermal tuning with integrated resistor heaters is most commonly employed, as it does not impact ring loss, but the tuning speed is limited by thermal time constants. Alternatively, electrical or bias-based tuning offers the potential for rapid tuning at the cost of some reduction in modulator extinction ratio. This approach has been shown to be effective with carrier-injection modulators, but is generally not considered effective for carrier-depletion modulators due to the reduced tunability.

Fig. 6 shows an average-power-based dynamic thermal tuning loop where a drop-port with a waveguide PD has been added to a ring modulator to monitor the ring's power levels. An FSM automatically locks the ring to the optimal thermal bias point by comparing the average optical power with an on-chip reference voltage to produce error information for the control of a thermal DAC that drives a resistive heater placed close to the modulator. Successful demonstration of this dynamic tuning loop with through-port monitoring was achieved for 25 Gb/s operation of carrier-depletion modulators with integrated 1 kΩ resistor heaters. This tuning loop was also modified to also tune the drop filters in a 25 Gb/s receiver front-end, with ring power monitoring achieved by adding peak detectors at the TIA output in order to not compete with the receiver's offset correction loop. A similar control loop was used for bias-based tuning of carrier-injection ring modulators, with the thermal DAC replaced by a bias DAC driving the modulator's anode terminal and the high-speed modulation applied to the cathode. This opens the possibility for dual-loop tuning with carrier-injection modulators, which involves a coarse-step thermal DAC and fine tuning with the bias DAC.

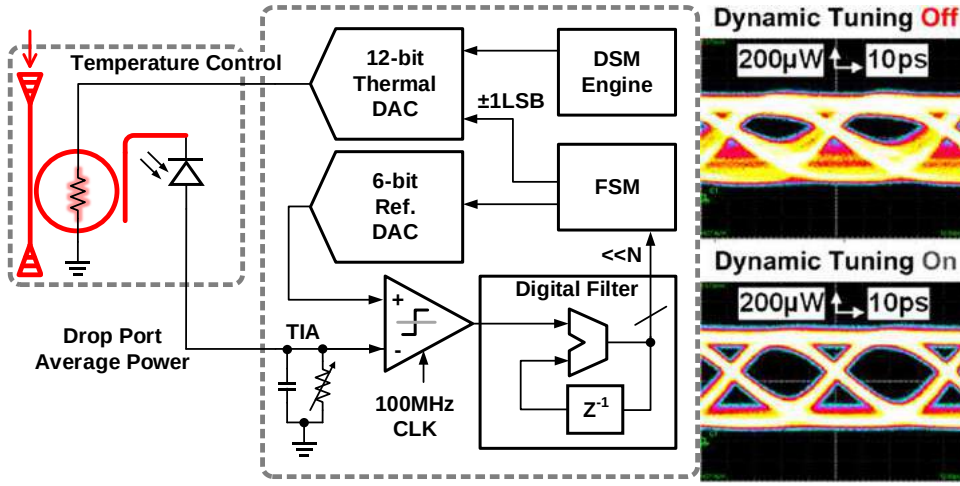


Fig. 6 Average-power-based dynamic thermal tuning loop to stabilize the ring resonance wavelength.

Mach-Zehnder Modulator Transmitters

Mach-Zehnder modulators (MZMs) are also refractive modulators which work by splitting the light to travel through two arms where a phase shift is developed that is a function of the applied electric field. The light in the two arms is then recombined either in phase or out of phase at the modulator output to realize the modulation. MZMs which use the free-carrier plasma-dispersion effect in pn diode devices to realize the optical phase shift have been integrated in CMOS processes and have demonstrated operation in excess of 10 Gb/s. The modulator transfer characteristic with half wave voltage $V\pi$ is

$$\frac{P_{out}}{P_{in}} = \frac{1}{2}\left(1 + \sin\frac{\pi V_{swing}}{V_{\pi}}\right) \tag{2}$$

Unlike smaller modulators which are treated as lumped capacitive loads, the relatively low modulation efficiency requires MZM phase shifters that commonly exceed 1 mm in silicon photonic implementations. Thus, the lumped-element drivers that are commonly utilized for ring resonator and electroabsorption modulators are not well suited for MZM operation at high speed. Instead, as shown in Fig. 7(a), traveling-wave drivers are often used with a differential electrical output signal that is distributed along the MZM phase shifter length using a pair of transmission lines terminated with a low impedance. The best high-speed output signal integrity is achieved when the transmission line electrodes are designed such that the propagation velocity of the electrical signal matches that of the optical signal in the MZM waveguides. While conceivably large contrast ratios are possible by simply increasing the MZM phase shifter length, this is limited by high-frequency loss of the long transmission lines and also increased optical insertion loss caused by the long doped waveguides. In order to achieve the required phase shift and reasonable contrast ratio, long devices and large differential swings are required; often necessitating a separate voltage supply MVdd. Thick-oxide cascode transistors are used to avoid stressing driver transistors with the high supply.

A driver which provides more flexibility in electrode design and allows for high-swing lumped element drivers is the multistage topology shown in Fig. 7(b). This design segments the total length of the MZM phase shifters into shorter regions which are driven by high-swing stacked inverter-style drivers. The segment length is chosen to have a self-resonance frequency higher than the signal bandwidth. As the high-frequency loss of these short segments is relatively small, higher contrast ratios are possible with high-swing drivers. A combination of passive wire delay elements and tunable-delay circuitry are employed to match the multistage driver delay with the optical propagation velocity. In this design, the speed and power efficiency is limited by distributing CMOS-level signals across long distances and the power consumed by the tunable delay elements. This driver topology generally achieves better power efficiency at low to moderate data rates due to the dominant CV^2f switching power component. While at the highest data rates, the limited contrast ratio traveling wave design becomes more power efficient.

Receiver Circuitry

Optical receiver front-ends generally determine the overall optical link performance, as their sensitivity sets the maximum data rate and amount of tolerable channel loss. Typical optical receivers use a photodiode to sense the high-speed optical power and produce an input current. This photocurrent is then converted to a voltage and amplified sufficiently for data resolution. In order to achieve increasing data rates, sensitive high-bandwidth receiver front-end circuits are necessary.

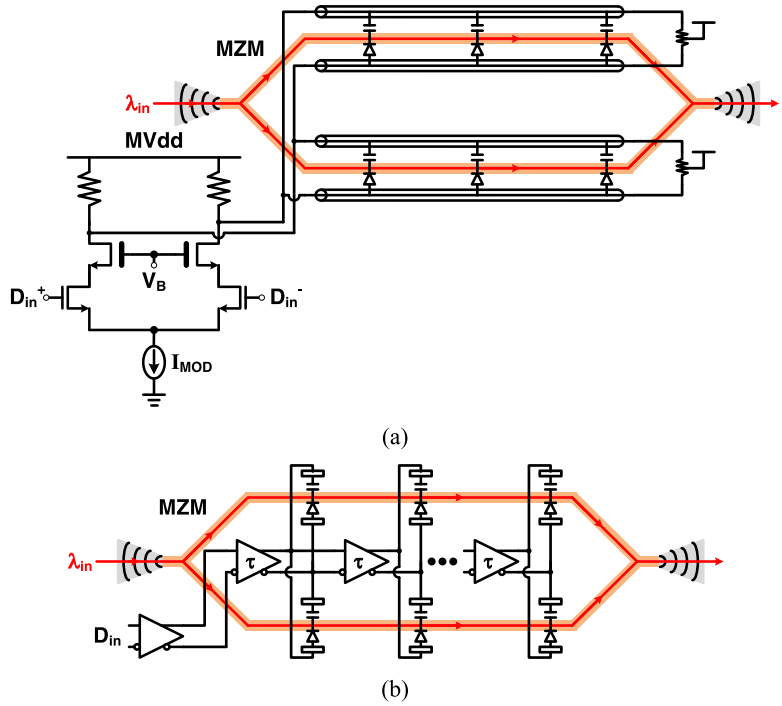


Fig. 7 Mach-Zehnder modulator drivers: (a) traveling-wave driver and (b) segmented driver.

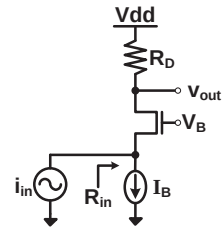


Fig. 8 Common-gate TIA.

While it is possible to convert the photocurrent into a voltage with a simple passive resistive front-end R_{in}, a direct tradeoff exists between input bandwidth and transimpedance gain R_T.

$$\text{Resistive Front-End}: \quad \begin{aligned} R_T &= R_{in} \\ \omega_{3dB} &= \frac{1}{R_{in}C_{in}} \end{aligned} \tag{3}$$

Thus, in order to achieve higher sensitivity, an active front-end circuit is often used. These optical receiver frontends can be generally classified as being either wideband, where the frontend has sufficient bandwidth to induce minimal ISI, or bandwidth-limited, where the input bandwidth is intentionally limited to improve sensitivity and/or power and the signal is recovered after a subsequent equalizer or double-sampling technique.

Wideband Transimpedance Amplifiers

A common-gate amplifier, shown in Fig. 8, is a useful circuit to decouple the input resistance and transimpedance.

$$\text{Common-Gate TIA Front-End}: \quad \begin{aligned} R_T &= R_D \\ \omega_{3dB} &\approx \frac{g_m}{C_{in}} \end{aligned} \tag{4}$$

High bandwidth is possible with sufficient biasing to enable a high input transistor transconductance g_m, while utilizing a high R_D allows for a high transimpedance gain. Unfortunately, as both the photodetector signal current and the transistor DC bias current flow through R_D, voltage headroom constraints in nanometer CMOS technologies limit its maximum value and the achievable gain and noise performance. A popular technique to improve input resistance for a given input bias current involves

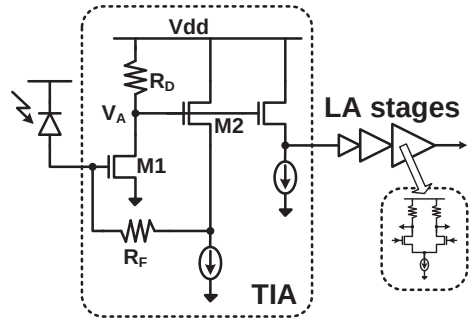


Fig. 9 Feedback TIA.

modifying a conventional common-gate input stage to a regulated cascode (RGC) architecture which employs active negative feedback gain to boost the input transconductance. This reduced input resistance pushes the input pole to a higher frequency, relaxing tradeoffs between TIA gain and bandwidth. However, conventional RGC topologies require additional voltage headroom due to the cascode topology. Moreover, extra power is required in the feedback stage in order to avoid excessive TIA frequency peaking and obtain sufficient noise performance. An interesting alternative approach to improve the input bandwidth involves employing a transformer-based RGC input stage which provides passive negative-feedback gain that enhances the effective transconductance of the input common-gate transistor. This allows for considerable bandwidth extension without significant noise degradation or power consumption.

Another wideband receiver frontend is the shunt-feedback TIA shown in Fig. 9, which decouples the transimpedance gain and input impedance by the amount of amplifier gain.

$$\text{Feedback TIA Front-End}: \quad \begin{aligned} R_T &= R_F\left(\frac{A}{1+A}\right) \\ \omega_{3dB} &\approx \frac{1+A}{R_F C_{in}} \\ &\text{where } A \approx g_{m1}(R_D \| r_{o1}) \end{aligned} \tag{5}$$

The feedback TIA topology has the advantage that only the photodetector signal current flows through R_F, which provides the potential for large transimpedance gain without voltage headroom constraints. This structure allows for potentially both high transimpedance and bandwidth, provided that the amplifier in the TIA has a sufficient gain-bandwidth product, Af_A, to ensure system stability. However, as data rates scale, the TIA's amplifier gain-bandwidth must increase as a quadratic function in order to maintain the same effective TIA gain because of the inherent transimpedance limit.

$$R_T \leq \frac{Af_A}{2\pi C_{in} f_{3dB}^2} \tag{6}$$

TIA performance scaling is further limited by the lack of gain that is achieved in modern CMOS processes at nominal supply voltages due to both voltage headroom constraints and intrinsic transistor gain. This limits the maximum R_F value, which both reduces the transimpedance gain and increases the input-referred current noise. As data rates increase and less gain is realized in the input transimpedance stage, receiver sensitivity is improved with additional voltage or limiting amplifier (LA) stages that follow the TIA. High-performance optical receivers often use four or more differential amplifier stages in order to achieve adequate sensitivity, which can more than double the total optical receiver power consumption.

Bandwidth-Limited Frontends

The aforementioned optical receiver scaling issues have motivated researchers to investigate bandwidth-limited front-ends in order to improve sensitivity and power consumption. A receiver front-end architecture that eliminates linear high gain elements, and thus is less sensitive to the reduced gain in modern processes, is the integrating and double-sampling front-end shown in Fig. 10. The absence of high gain amplifiers allows for savings in both power and area and makes the integrating and double-sampling architecture advantageous for chip-to-chip optical interconnect systems where retiming is also performed at the receiver. The integrating and double-sampling receiver front-end demultiplexes the incoming data stream with parallel segments that include a pair of input samplers, a buffer, and a sense-amplifier. Two current sources at the receiver input node, the photodiode current and a current source that is feedback biased to the average photodiode current, supply and deplete charge from the receiver input capacitance respectively. For data encoded to ensure DC balance, the input voltage will integrate up or down due to the mismatch in these currents. A differential voltage, ΔV_b, that represents the polarity of the received bit is developed by sampling the input voltage at the beginning and end of a bit period defined by the rising edges of the synchronized sampling clocks $\Phi[n]$ and $\Phi[n+1]$ that are spaced a bit-period, T_b, apart. This differential voltage is buffered and applied to the inputs of an offset-corrected sense-amplifier which is used to regenerate the signal to CMOS levels.

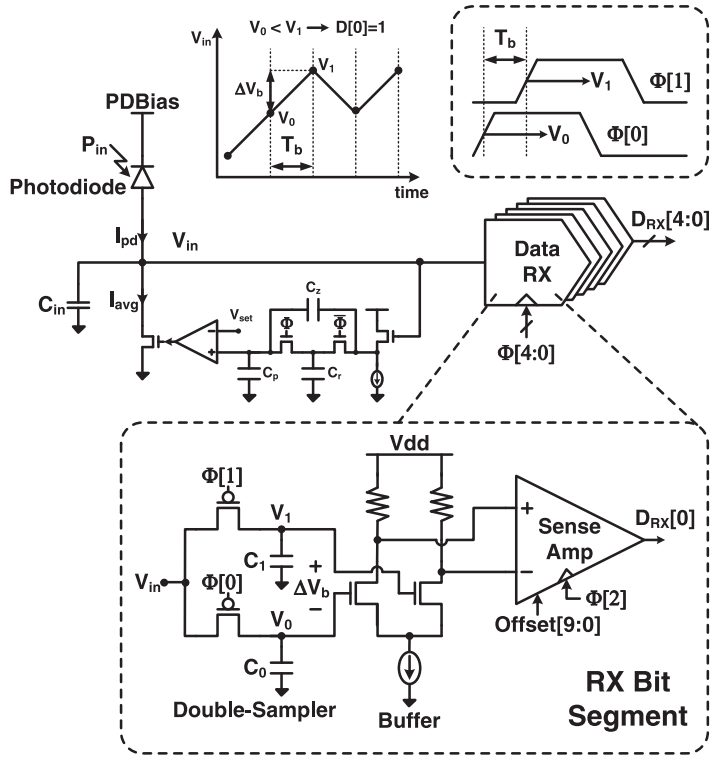


Fig. 10 Integrating and double-sampling receiver front-end.

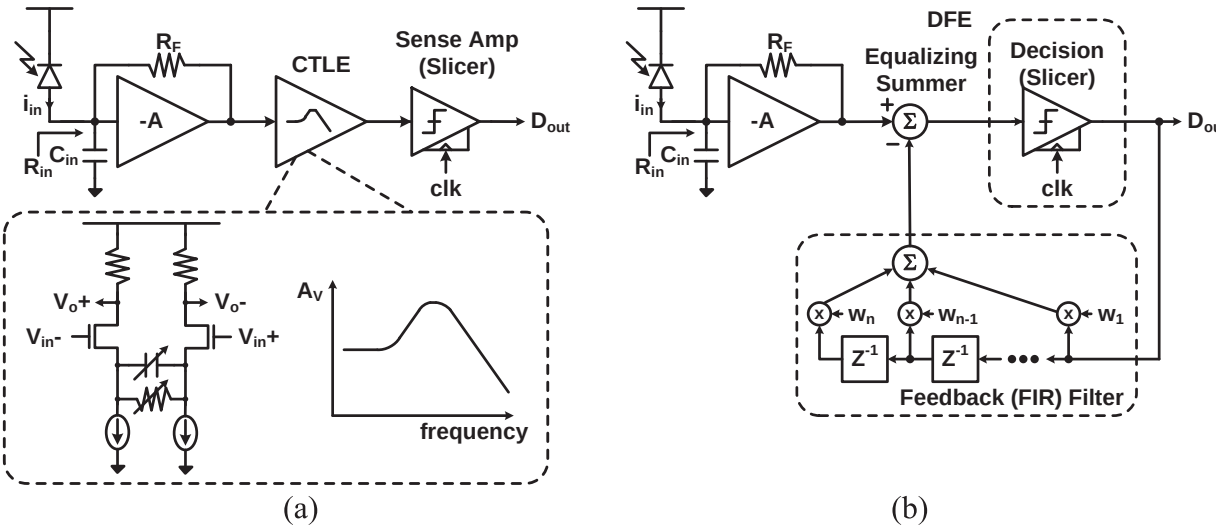


Fig. 11 Low-bandwidth TIA front-end followed by an equalizer: (a) continuous-time linear equalizer and (b) decision feedback equalizer.

One issue with this integrating and double-sampling front-end is that the input node saturates with long runlengths of consecutive bits. A solution to this is to implement a lossy integrator input node by utilizing a relatively large input resistor or low-bandwidth feedback TIA front-end to limit the low-frequency swing to a predictable value. A constant effective differential voltage is achieved by dynamically modulating the sense-amplifier's offset with a discrete-time FIR filter that follows the low-bandwidth input stage.

Fig. 11 shows another receiver front-end architecture which offers sensitivity improvement by utilizing a low-bandwidth feedback TIA followed by an equalizer circuit. At high data rates, the input-referred current noise of wideband feedback TIAs is often dominated by the feedback resistor R_F whose value is limited to achieve sufficient bandwidth. Intentionally increasing R_F results in lower noise which is integrated over a smaller bandwidth. However, significant intersymbol interference (ISI) exists at the

output of this low-bandwidth feedback TIA front-end. This ISI is then compensated by the subsequent equalizer block to achieve adequate voltage and timing margins at the data samplers.

A continuous-time linear amplifier (CTLE) is a simple and low-area equalizer block which implements a high-pass filter transfer function to restore the total receiver front-end bandwidth. While this equalizer doesn't require any additional sampling clocks, it does have to supply gain at frequencies close to the full signal data rate. Also, the CTLE amplifier introduces additional noise. Nonetheless, the ability to reduce the input R_F noise significantly allows for a net sensitivity improvement.

Another equalizer topology is the decision feedback equalizer (DFE). A DFE attempts to directly subtract ISI from the incoming signal by feeding back the resolved data to control the polarity of the equalization taps. Unlike linear equalization, a DFE doesn't directly amplify the input signal noise or cross-talk since it uses the quantized input values. However, there is the potential for error propagation in a DFE if the noise is large enough for a quantized output to be wrong. Also, due to the feedback equalization structure, the DFE cannot cancel pre-cursor ISI. The major challenge in DFE implementation is closing timing on the first tap feedback since this must be done in one bit period or unit interval (UI). Direct feedback implementations require this critical timing path to be highly optimized. Improvements in DFE speed are possible with loop-unrolling architectures which speculative summation with multiple data samplers and additional look-ahead logic to relax the critical timing path.

Further Reading

Analui, B., Guckenberger, D., Kucharski, D., Narasimha, A., 2006. A fully integrated 20-Gb/s optoelectronic transceiver implemented in a standard 0.13-μm CMOS SOI technology. IEEE Journal of Solid-State Circuits 41, 2945–2955.

Cevrero, A., Ozkaya, I., Francese, P.A., *et al.*, 2017. A 64 Gb/s 1.4pJ/b NRZ optical-receiver data-path in 14 nm CMOS FinFET. In: IEEE International Solid-State Circuits Conference, pp. 482–483.

Emami-Neyestanak, A., Liu, D., Keeler, G., Helman, N., Horowitz, M., 2002. A 1.6 Gb/s, 3 mW CMOS receiver for optical communication. In: IEEE Symposium on VLSI Circuits, pp. 84–87.

Li, C., Bai, R., Shafik, A., *et al.*, 2014. Silicon photonic transceiver circuits with microring resonator bias-based wavelength stabilization in 65-nm CMOS. IEEE Journal of Solid-State Circuits 49, 1419–1436.

Li, C., Palermo, S., 2013. A low-power 26-GHz transformer-based regulated cascode SiGe BiCMOS transimpedance amplifier. IEEE Journal of Solid-State Circuits 48, 1264–1275.

Li, D., Minoia, G., Repossi, M., *et al.*, 2014. A low-noise design technique for high-speed CMOS optical receivers. IEEE Journal of Solid-State Circuits 49, 1437–1447.

Li, H., Xuan, Z., Titriku, A., *et al.*, 2015. A 25 Gb/s, 4.4 V-swing, AC-coupled ring modulator-based WDM transmitter with wavelength stabilization in 65 nm CMOS. IEEE Journal of Solid-State Circuits 50, 3145–3159.

Palermo, S., Emami-Neyestanak, A., Horowitz, M., 2008. A 90 nm CMOS 16 Gb/s transceiver for optical interconnects. IEEE Journal of Solid-State Circuits 43, 1235–1246.

Palermo, S., Horowitz, M., 2006. High-speed transmitters in 90 nm CMOS for high-density optical interconnects. In: IEEE European Solid-State Circuits Conference, pp. 508–511.

Park, S., Yoo, H., 2004. 1.25-Gb/s regulated cascode CMOS transimpedance amplifier for gigabit ethernet applications. IEEE Journal of Solid-State Circuits 39, 112–121.

Raj, M., Monge, M., Emami, A., 2016. Modelling and nonlinear equalization technique for a 20 Gb/s 0.77 pJ/b VCSEL transmitter in 32 nm SOI CMOS. IEEE Journal of Solid-State Circuits 51, 1734–1743.

Soref, R., Bennett, B., 1987. Electrooptical effects in silicon. IEEE Journal of Quantum Electronics 23, 123–129.

Temporiti, E., Ghilioni, A., Minoia, G., *et al.*, 2016. Insights into silicon photonics Mach–Zehnder-based optical transmitter architectures. IEEE Journal of Solid-State Circuits 51, 3178–3191.

Young, I., Mohammed, E., Liao, J., *et al.*, 2010. Optical I/O technology for tera-scale computing. IEEE Journal of Solid-State Circuits 45, 235–248.

Yu, K., Li, C., Li, H., *et al.*, 2016. A 25 Gb/s hybrid-integrated silicon photonic source-synchronous receiver with microring wavelength stabilization. IEEE Journal of Solid-State Circuits 51, 2129–2141.

Foundations of Coherent Transients in Semiconductors

Torsten Meier, University of Paderborn, Paderborn, Germany
Stephan W Koch, Philipps University, Marburg, Germany

Introduction

Electromagnetic irradiation of matter generates a coherent superposition of the excited quantum states. The attribute "coherent" relates to the fact that the material excitations originally have a well defined phase dependence which is imposed by the phase of the excitation source, often a laser in the optical regime. Macroscopically, the generated superposition state can be described as an optical polarization which is determined by the transition amplitudes between the participating quantum states.

The optical polarization is a typical non-equilibrium quantity that decays to zero when a system relaxes to its equilibrium state. Coherent effects are therefore only observable in a certain time window after pulsed photoexcitation, or in the presence of a continuous-wave (cw) beam. As *coherent transients* one usually refers to phenomena that can be observed during or shortly after pulsed laser excitation and critically depend on the presence of the induced optical polarization.

Many materials such as atoms, molecules, metals, insulators, semiconductors including bulk crystals, heterostructures, and surfaces, as well as organic and biological structures are studied using coherent optical spectroscopy. Depending on the particular system, the states participating in the optical transitions, the interactions among them, and the resulting time scale for the decay of the induced polarization may be very different. As a result, the time window during which coherent effects are observable can be as long as seconds for certain atomic transitions, or it may be as short as a few femtoseconds (10^{-15} s), e.g., for metals, surfaces, or highly excited semiconductors.

Coherent spectroscopy and the analysis of coherent transients has provided valuable information on the nature and dynamics of the optical excitations. Often it is possible to learn about the interaction processes among the photoexcitations and to follow the temporal evolution of higher-order transitions which are only accessible if the system is in a non-equilibrium state.

The conceptually simplest experiment which one may use to observe coherent transients is to time resolve the transmission or reflection induced by a single laser pulse. However, much richer information can be obtained if one excites the system with several pulses. A pulse sequence with well-controlled delay times makes it possible to study the dynamical evolution of the photoexcitations not only by time resolving the signal but also by varying the delay. Prominent examples of such experiments are pump–probe measurements, which usually are performed with two incident pulses, or four-wave mixing, for which one may use two or three incident pulses.

Besides the microscopic interaction processes, the outcome of an experiment is determined by the quantum mechanical selection rules for the transitions and by the symmetries of the system under investigation. For example, if one wants to investigate coherent optical properties of surface states one often relies on phenomena, such as second-harmonic or sum-frequency generation, which give no signal in perfect systems with inversion symmetry. Due to the broken translational invariance, such experiments are therefore sensitive only to the dynamics of surface and/or interface excitations.

In this article the basic principles of coherent transients are presented and several examples are presented. The basic theoretical description and its generalization for the case of semiconductors are introduced.

Basic Principles

In the absence of free charges and currents, Maxwell's equations show that the electromagnetic field interacts with matter via the optical polarization. This polarization $\mathbf{P}$, or more precisely its second derivative with respect to time $\frac{\partial^2}{\partial t^2}\mathbf{P}$, appears as a source term in the wave equation for the electric field $\mathbf{E}$. Consequently, if the system is optically thin such that propagation effects within the sample can be ignored and if measurements are performed in the far field region, i.e., at distances that significantly exceed the characteristic optical wavelength λ, the emitted electric field resulting from the polarization is proportional to its second time derivative, $\mathbf{E} \propto \frac{\partial^2}{\partial t^2}\mathbf{P}$. Thus the measurement of the emitted field dynamics yields information about the temporal evolution of the optical material polarization.

Microscopically the polarization is determined by the transition amplitudes between the different states of the system. These may be the discrete states of atoms or molecules, or the microscopic valence and conduction band states in a dielectric medium, such as a semiconductor. In any case, the macroscopic polarization $\mathbf{P}$ is computed by summing over all microscopic transitions p_{cv} via $\mathbf{P}=\sum_{c,v}(\boldsymbol{\mu}_{cv}p_{cv}+c.c.)$, where $\boldsymbol{\mu}_{cv}$ is the dipole matrix element which determines the strength of the transitions between the states v and c, and $c.c.$ denotes the complex conjugate. If ε_c and ε_v are the energies of these states, their dynamic quantum mechanical evolution is described by the phase factors $e^{-i\varepsilon_c t/\hbar}$ and $e^{-i\varepsilon_v t/\hbar}$, respectively. Therefore each p_{cv} is evolving in time according to $e^{-i(\varepsilon_c-\varepsilon_v)t/\hbar}$. Assuming that we start at $t=0$ with $p_{cv}(t=0)=p_{cv,0}$, which may be induced by a short optical pulse, we have for the optical polarization $\mathbf{P}(t)=\Theta(t)\sum_{c,v}\left(\boldsymbol{\mu}_{cv}p_{cv,0}e^{-i(\varepsilon_c-\varepsilon_v)t/\hbar}+c.c.\right)$. Thus $\mathbf{P}(t)$ is given by a summation over microscopic transitions which all oscillate with frequencies proportional to the energy differences between the involved states. Hence, the optical

 doi:10.1016/B978-0-12-803581-8.09564-3

polarization is clearly a coherent quantity which is characterized by amplitude and phase. Furthermore, the microscopic contributions to $\mathbf{P}(t)$ add up coherently. Depending on the phase relationships, one may obtain either constructive superposition, interference phenomena like quantum beats, or destructive interference leading to a decay (dephasing) of the macroscopic polarization.

Optical Bloch Equations

The dynamics of photoexcited systems can be conveniently described by a set of equations, the optical Bloch equations, named after Felix Bloch (1905–1983) who first formulated such equations to analyze the spin dynamics in nuclear magnetic resonance. For the simple case of a two-level model the Bloch equations can be written as

$$i\hbar \frac{\partial}{\partial t} p = \Delta\varepsilon p + \mathbf{E} \cdot \boldsymbol{\mu} I \tag{1}$$

$$i\hbar \frac{\partial}{\partial t} I = 2\mathbf{E} \cdot \boldsymbol{\mu}(p - p^*) \tag{2}$$

Here, $\Delta\varepsilon$ is the energy difference and I the inversion, i.e., the occupation difference between upper and lower state. The field $\mathbf{E}$ couples the polarization to the product of the Rabi energy $\mathbf{E} \cdot \boldsymbol{\mu}$ and the inversion I. In the absence of the driving field, i.e., $\mathbf{E} \equiv 0$, Eq. (1) describes the free oscillation of $p \propto e^{-i\Delta\varepsilon t/\hbar}$ discussed above.

The inversion is determined by the combined action of the Rabi energy and the transition p. The total occupation N, i.e., the sum of the occupations of the lower and the upper states, remains unchanged during the optical excitation since light does not create or destroy electrons, it only transfers them between different states. If N is initially normalized to 1, then I can vary between -1 and 1. $I = -1$ corresponds to the ground state of the system, where only the lower state is occupied. In the opposite extreme, i.e., for $I = 1$, the occupation of the upper state is 1 and the lower state is completely depleted.

One can show that Eqs. (1) and (2) contain another conservation law. Introducing $\mathcal{P} = p + p* = 2Re[p]$ and $\mathcal{J} = i(p - p*) = -2Im[p]$, which are real quantities and often named polarization and polarization current, respectively, one finds that the relation

$$\mathcal{P}^2 + \mathcal{J}^2 + I^2 = 1 \tag{3}$$

is fulfilled. Thus the coupled coherent dynamics of the transition and the inversion can be described on a unit sphere, the so-called Bloch sphere. The three terms $\mathcal{P}$, $\mathcal{J}$, and I define the components of the three-dimensional Bloch vector $\mathbf{S} = (\mathcal{P}, \mathcal{J}, I)$. With these definitions, one can reformulate Eqs. (1) and (2) as

$$\frac{\partial}{\partial t}\mathbf{S} = \boldsymbol{\Omega} \times \mathbf{S} \tag{4}$$

with $\boldsymbol{\Omega} = (-2\boldsymbol{\mu} \cdot \mathbf{E}, 0, \Delta\varepsilon)/\hbar$ and $\times$ denoting the vector product.

The structure of Eq. (4) is mathematically identical to the equations describing either the angular momentum dynamics in the presence of a torque or the spin dynamics in a magnetic field. Therefore, the vector $\mathbf{S}$ is often called pseudospin. Moreover, many effects which can be observed in magnetic resonance experiments, e.g., free decay, quantum beats, echoes, etc. have their counterparts in photoexcited optical systems.

The vector product on the right hand side of Eq. (4) shows that $\mathbf{S}$ is changed by $\boldsymbol{\Omega}$ in a direction perpendicular to $\mathbf{S}$ and $\boldsymbol{\Omega}$. For vanishing field the torque $\boldsymbol{\Omega}$ has only a z-component. In this case the z-component of $\mathbf{S}$, i.e., the inversion, remains constant, whereas the x- and y-components, i.e., the polarization and the polarization current, oscillate with the frequency $\Delta\varepsilon/\hbar$ on the Bloch sphere around $\boldsymbol{\Omega}$.

Semiconductor Bloch Equations

If one wants to analyze optical excitations in crystalline solids, i.e., in systems which are characterized by a periodic arrangement of atoms, one has to include the continuous dispersion (bandstructure) into the description. This can be done by including the crystal momentum $\hbar\mathbf{k}$ as an index to the quantities which describe the optical excitation of the material, e.g., p and I in Eqs. (1) and (2). Hence, one has to consider a separate two-level system for each crystal momentum $\hbar\mathbf{k}$. As long as all other interactions are disregarded the bandstructure simply introduces a summation over the uncoupled contributions from the different $\mathbf{k}$ states, leading to inhomogeneous broadening due to the presence of a range of transition frequencies $\Delta\varepsilon(\mathbf{k})/\hbar$.

For any realistic description of optical processes in solids, it is essential to go beyond the simple picture of non-interacting states and to treat the interactions among the elementary material excitations, e.g., the Coulomb interaction between the electrons and the coupling to additional degrees of freedom, such as lattice vibrations (electron-phonon interaction) or other bath-like subsystems. If crystals are not ideally periodic, imperfections, which can often be described as a disorder potential, need to be considered as well.

All these effects can be treated by proper extensions of the optical Bloch equations introduced above. For semiconductors the resulting generalized equations are known as semiconductor Bloch equations, where the microscopic interactions are included at a

certain level of approximation. For a two band model of a semiconductor, these equations can be written schematically as

$$i\hbar\frac{\partial}{\partial t}p_{\mathbf{k}} = \Delta\varepsilon_{\mathbf{k}}p_{\mathbf{k}} + \Omega_{\mathbf{k}}\left(n^{c}_{\mathbf{k}} - n^{v}_{\mathbf{k}}\right) + i\hbar\frac{\partial}{\partial t}p_{\mathbf{k}}|_{corr} \tag{5}$$

$$i\hbar\frac{\partial}{\partial t}n^{c}_{\mathbf{k}} = \left(\Omega^{*}_{\mathbf{k}}p_{\mathbf{k}} - \Omega_{\mathbf{k}}p^{*}_{\mathbf{k}}\right) + i\hbar\frac{\partial}{\partial t}n^{c}_{\mathbf{k}}|_{corr} \tag{6}$$

$$i\hbar\frac{\partial}{\partial t}n^{v}_{\mathbf{k}} = -\left(\Omega^{*}_{\mathbf{k}}p_{\mathbf{k}} - \Omega_{\mathbf{k}}p^{*}_{\mathbf{k}}\right) + i\hbar\frac{\partial}{\partial t}n^{v}_{\mathbf{k}}|_{corr} \tag{7}$$

Here p_k is the microscopic polarization and $n^{c}_{\mathbf{k}}$ and $n^{v}_{\mathbf{k}}$ are the electron populations in the conduction and valence bands (c and v), respectively. Due to the Coulomb interaction and possibly further processes, the transition energy $\Delta\varepsilon_k$ and the Rabi energy Ω_k both depend on the excitation state of the system, i.e., they are functions of the time-dependent polarizations p_k and populations $n^{c/v}_{\mathbf{k}}$. This leads, in particular, to a coupling among the excitations for all different values of the crystals momentum $\hbar\mathbf{k}$. Consequently, in the presence of interactions the optical excitations can no longer be described as independent two-level systems but have to be treated as a coupled many-body system. A prominent and important example in this context is the appearance of strong exciton resonances which, as a consequence of the Coulomb interaction, show up in the absorption spectra of semiconductors energetically below the fundamental band gap.

The interaction effects lead to significant mathematical complications since they induce couplings between all the different quantum states of a system and introduce an infinite hierarchy of equations for the microscopic correlation functions. The terms given explicitly in Eqs. (5)–(7) arise in a treatment of the Coulomb interaction on the Hartree-Fock level. Whereas this level is sufficient to describe excitonic resonances, there are many additional effects, e.g., excitation-induced dephasing due to Coulomb scattering and significant contributions from higher-order correlations like excitonic populations and biexcitonic resonances, which make it necessary to treat many-body correlation effects that are beyond the Hartree-Fock level. These contributions are formally included by the terms denoted as $|_{corr}$ in Eqs. (5)–(7). The systematic truncation of the many-body hierarchy and the analysis of controlled approximations is the basic problem in the microscopic theory of optical processes in condensed matter systems.

Examples

Radiative Decay

The emitted field originating from the non-equilibrium coherent polarization of a photoexcited system can be monitored by measuring the transmission and the reflection as function of time. If only a single isolated transition is excited, the dynamic evolution of the polarization and therefore of the transmission and reflection is governed by radiative decay. This decay is a consequence of the coupling of the optical transition to the light field described by the combination of Maxwell- and Bloch-equations. Radiative decay simply means that the optical polarization is converted into a light field on a characteristic time scale $2T_{rad}$. Here, T_{rad} is the population decay time, often also denoted as T_1-time. This radiative decay is a fundamental process which limits the time on which coherent effects are observable for any photoexcited system. Due to other mechanisms, however, the non-equilibrium polarization often vanishes considerably faster.

The value of T_{rad} is determined by the dipole matrix element and the frequency of the transition, i.e., $T^{-1}_{rad} \propto |\boldsymbol{\mu}|^2\Delta\omega$, with $\Delta\omega = \Delta\varepsilon/\hbar$. The temporal evolution of the polarization and the emitted field are proportional to $e^{-i\Delta\omega t - t/(2T_{rad})}$. Usually one measures the intensity of a field, i.e., its squared modulus, which evolves as $e^{-t/T_{rad}}$ and thus simply shows an exponential decay, the so-called *free-induction decay*, see **Fig. 1**. T_{rad} can be very long for transitions with small matrix elements. For semiconductor quantum wells, however, it is on the order of only 10 *ps*, as the result of the strong light–matter interaction.

The time constant on which the optical polarization decays is often called T_2. In the case that this decay is dominated by radiative processes we thus have $T_2 = 2T_{rad}$.

Superradiance

The phenomenon of superradiance can be discussed considering an ensemble of N two-level systems which are localized at certain positions $\mathbf{R}_i$. In this case Maxwell's equations introduce a coupling among all these resonances since the field emitted from any specific resonance interacts with all other resonances and interferes with their emitted fields. As a result, this system is characterized by N eigenmodes originating from the radiatively coupled optical resonances.

A very spectacular situation arises if one considers N identical two-level systems regularly arranged with a spacing that equals an integer multiple of $\lambda/2$, where λ is the wavelength of the system's resonance, i.e., $\lambda = c/\Delta\omega$, where c is the speed of light in the considered material, see **Fig. 2(a)**. In this case all emitted fields interfere constructively and the system behaves effectively like a single two-level system with a matrix element increased by $\sqrt{N}$. Consequently, the radiative decay rate is increased by N and the polarization of the coupled system decays N-times faster than that of an isolated system, see **Fig. 2(b)**. This effect is called superradiance.

It is possible to observe superradiant coupling effects, e.g., in suitably designed semiconductor heterostructures. **Fig. 2(c)** compares the measured time-resolved reflection from a single quantum well (dashed line) with that of a Bragg structure, i.e., a multiple quantum-well structure which consists of 10 wells that are separated by $\lambda/2$ (solid line), where λ is the wavelength of the

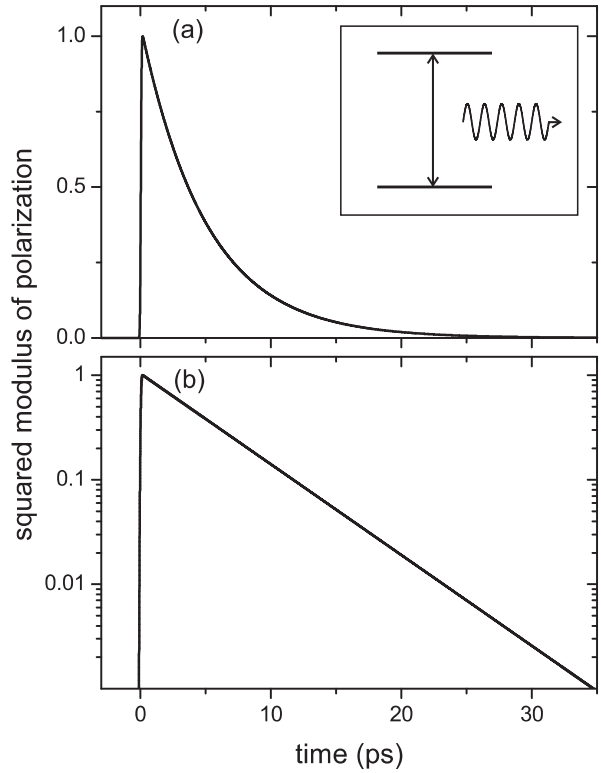


Fig. 1 A two-level system is excited by a short optical pulse at $t=0$. Due to radiative decay (inset) and possibly other dephasing mechanisms, the polarization decays exponentially as function of time with the time constant T_2, the dephasing time. The intensity of the optical field which is emitted as a result of this decay is proportional to the squared modulus of the polarization, which falls off with the time constant $T_2/2$. The decay of the squared modulus of the polarization is shown in (a) on a linear scale and in (b) on a logarithmic scale. The dephasing time was chosen to be $T_2=10ps$ which models the radiative decay of the exciton transition of a semiconductor quantum well.

exciton resonance. For times greater than about 2 *ps* the direct reflection of the exciting pulse has decayed sufficiently and one is left with the exponential decay of the remaining signal. Fig. 2(c) shows that this decay is much more rapid for the Bragg structure than for the single quantum well due to superradiance introduced by the radiative coupling among the quantum wells.

Destructive Interference

We now consider a distribution of two-level systems which have slightly different transition frequencies characterized by the distribution function $g(\Delta\omega-\overline{\omega})$ of the transition frequencies which is peaked at the mean value $\overline{\omega}$ and has a spectral width of $\delta\omega$, see Fig. 3(a). Ignoring the radiative coupling among the resonances, the optical polarization of the total system evolves after excitation by a short laser pulse at $t=0$ proportional to $\int d\omega g(\Delta\omega-\overline{\omega})e^{-i\Delta\omega t}\propto\tilde{g}(t)e^{-\overline{\omega}t}$, where $\tilde{g}(t)$ denotes the Fourier transform of the frequency distribution function. $\tilde{g}(t)$ and thus the optical polarization decays on a time scale which is inversely proportional to the spectral width of the distribution function $\delta\omega$. Thus the destructive interference of many different transitions frequencies results in a rapid decay of the polarization, see Fig. 3(b).

In the spectral domain this rapid decay shows up as inhomogeneous broadening. Depending on the system under investigation, there are many different sources for such an inhomogeneous broadening. One example is the Doppler broadening in atomic gases or disorder effects such as well-width fluctuations in semiconductor quantum wells or lattice imperfections in crystals.

In the nonlinear optical regime it is under certain circumstances possible to reverse the destructive interference of inhomogeneously broadened coherent polarizations. For example, in four-wave mixing a second pulse may lead to a rephasing of the contributions with different frequencies, which results in the photon echo, see discussion below.

Quantum Beats

The occurrence of quantum beats can be understood most easily in a system where the total optical polarization can be attributed to a finite number of optical transitions. Let us assume for simplicity that all these transitions have the same matrix element. In this case, after excitation with a short laser pulse at $t=0$ the optical polarization of the total system evolves proportional to $\sum_i e^{-i\Delta\omega_i t}$. The finite number of frequencies results in a temporal modulation with time periods $2\pi/(\Delta\omega_i-\Delta\omega_j)$ of the squared modulus of the polarization which is proportional to the emitted field intensity. For the case of two frequencies the squared modulus of the

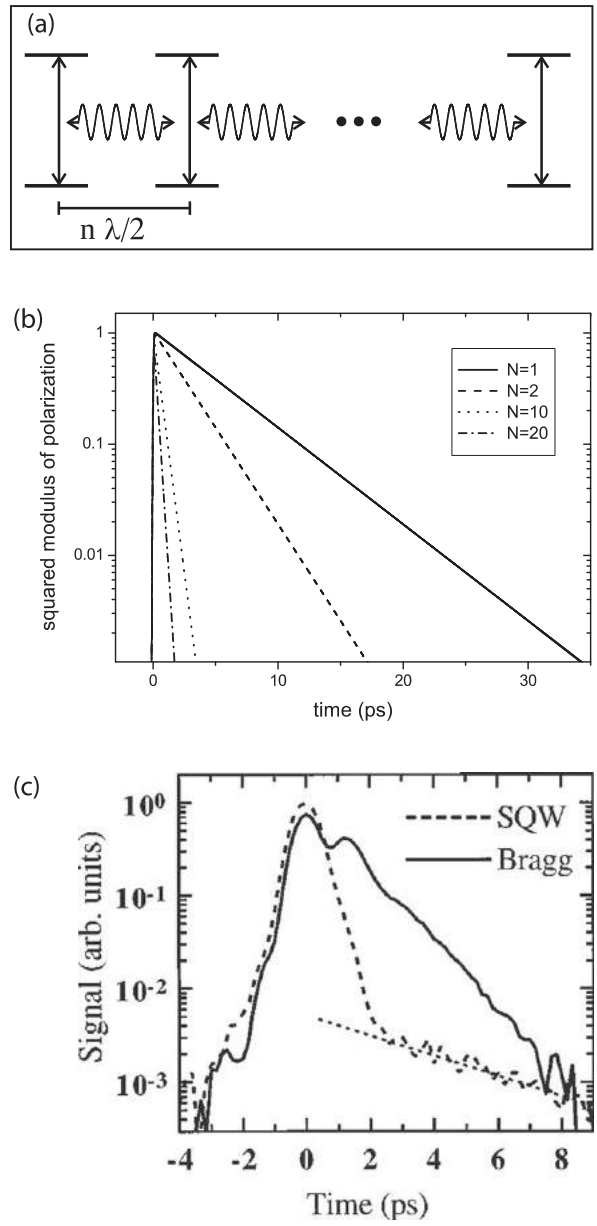


Fig. 2 (a) *N* identical two-level systems are depicted which are regularly arranged with a spacing that equals an integer multiple of $\lambda/2$, where λ is the wavelength of the system's resonance. Due to their coupling via Maxwell's equations all fields emitted from the two-level systems interfere constructively and the coupled system behaves effectively like a single two-level system with a optical matrix element increased by $\sqrt{N}$. (b) The temporal decay of the squared modulus of the polarization is shown on a logarithmic scale. The radiative decay rate is proportional to *N* and thus the polarization of the coupled system decays *N*-times faster than that of an isolated system. This effect is called superradiance. (c) Measured time-resolved reflection of a semiconductor single quantum well (dashed line) and a N=10 Bragg structure, i.e., a multi quantum well where the individual wells are separated by $\lambda/2$ (solid line), on a logarithmic scale. Part (c) is taken from **Fig. 5** of Haas, S., Stroucken, T., Hübner, M., *et al.* 1998. Phys. Rev. B 57, 14860.

polarization is proportional to $[1+cos((\Delta\omega_1-\Delta\omega_2)t)]$, i.e., due to the interference of two contributions the polarization varies between a maximum and zero, see **Fig. 4(b)**.

In the linear optical regime it is impossible to distinguish whether the optical transitions are uncoupled or coupled. As shown in **Fig. 4**, two uncoupled two-level systems give the same linear polarization as a three-level system where the two transitions share a common state. It is, however, possible to decide about the nature of the underlying transitions if one performs nonlinear optical spectroscopy. This is due to the fact that the so-called quantum beats, i.e., the temporal modulations of the polarization of an intrinsically coupled system, show a different temporal evolution as the so-called polarization interferences, i.e., the temporal modulations of the polarization of uncoupled systems. The former ones are also much more stable in the presence of inhomogeneous broadening than the latter ones.

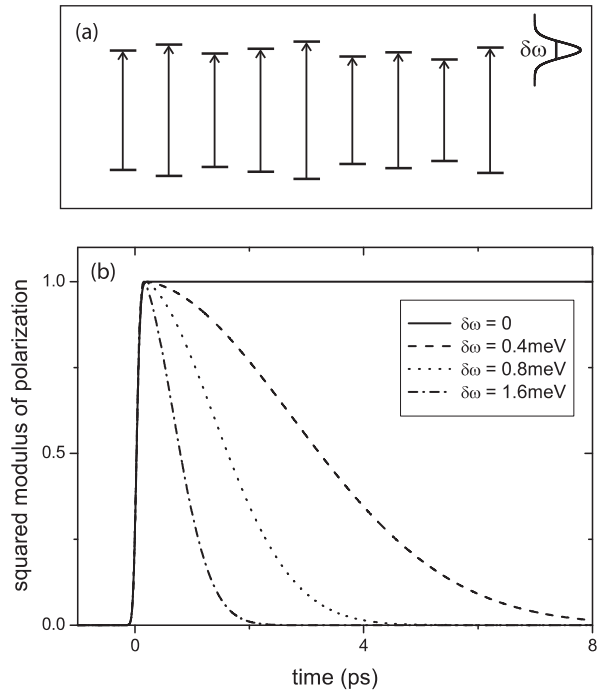


Fig. 3 (a) An ensemble of two-level systems is shown where the resonance frequencies are randomly distributed according to a Gaussian distribution function of width $\delta\omega$ around an average value. (b) The temporal dynamics of the squared modulus of the polarization of ensembles of two-level systems after excitation with a short optical pulse at $t=0$ is shown on a linear scale. Since the dephasing time is set to infinity, i.e., $T_2 \to \infty$, the polarization of an ensemble of identical two-level system ($\delta\omega=0$) does not decay. However, for a finite width of the distribution, $\delta\omega > 0$, the individual polarizations of the ensemble oscillate with different frequencies and, therefore, due to destructive interference the polarization of the ensemble decays as function of time. Since the Fourier transform of a frequency-domain Gaussian is a Gaussian in the time domain, the dashed, dotted, and dashed dotted line have a Gaussian shape with a temporal width which is inversely proportional to $\delta\omega$.

In semiconductor heterostructures, quantum-beat spectroscopy has been widely used to investigate the temporal dynamics of excitonic resonances. Also the coupling among different optical resonances has been explored in pump–probe and four-wave-mixing measurements.

Coherent Control

In the coherent regime the polarization induced by a first laser pulse can be modified by a second, delayed pulse. For example, the first short laser pulse incident at $t=0$ induces a polarization proportional to $e^{-i\Delta\omega t}$ and a second pulse arriving at $t=\tau>0$ induces a polarization proportional to $e^{-i(\varphi)}e^{-i\Delta\omega(t-\tau)}$. Thus for $t\geq\tau$ the total polarization is given by $e^{-i\Delta\omega t}(1+e^{-i\Delta(\varphi)})$ with $\Delta(\varphi)=(\varphi)-\Delta\omega\tau$. If $\Delta(\varphi)$ is an integer multiple of 2π the polarizations of both pulses interfere constructively and the amplitude of the resulting polarization is doubled as compared to a single pulse. If, on the other hand, $\Delta(\varphi)$ is an odd multiple of π, the polarizations of both pulses interfere destructively and the resulting polarization vanishes after the action of both pulses. Thus by varying the phase difference of the two pulses it is possible to coherently enhance or destroy the optical polarization, see **Fig. 5(b)**.

One can easily understand the coherent control in the frequency domain by considering the overlap between the absorption of the system and the pulse spectrum. For excitation with two pulses that are temporally delayed by τ, the spectrum of the excitation shows interference fringes with a spectral oscillation period that is inversely proportional to τ. The positions of the maxima and the minima of the fringes depend on the phase difference between the two pulses. For constructive interference, the excitation spectrum is at a maximum at the resonance of the system, whereas for destructive interference the excitation spectrum vanishes at the resonance of the system, see **Fig. 5(a)**.

Coherent control techniques have been applied to molecules and solids to control the dynamical evolution of electronic wavepackets and also the coupling to nuclear degrees of freedom. In this context, it is sometimes possible to steer certain chemical reactions into a preferred direction by using sequences of laser pulses which can be chirped, i.e., have a time-dependent frequency.

Coherent Photocurrents

Exciting an unbiased inversion symmetric semiconductor with laser light that resonantly above the band gap does not lead to electrical currents. This is due to the fact, that in such a situation one excites the same amount of electrons with positive as with negative (crystal) momentum $\hbar\mathbf{k}$, see **Fig. 6(a)**. Therefore the electrons move with the same average velocity in any direction and

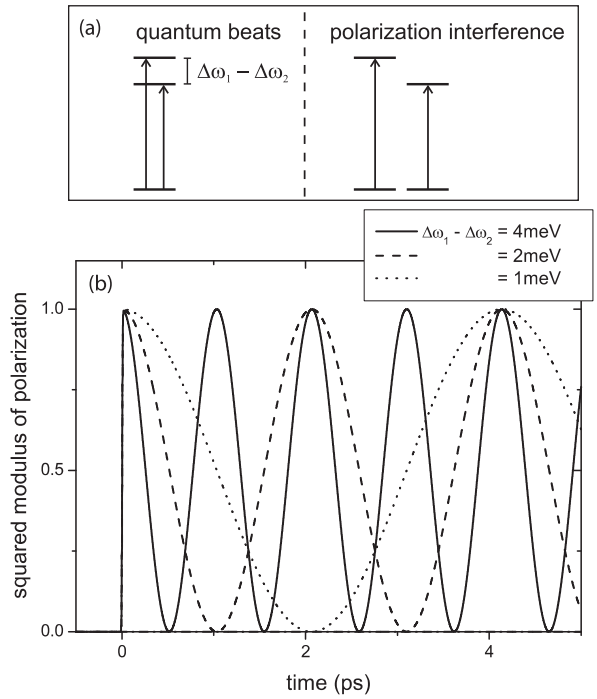


Fig. 4 (a) Two optical resonances with frequency difference $\Delta\omega_1 - \Delta\omega_2$ may be realized by either a three-level system or by two uncoupled two-level systems. After impulsive excitation with an optical pulse the linear polarization of both types of systems show a modulation of the squared modulus of the polarization with the time period $2\pi/(\Delta\omega_1 - \Delta\omega_2)$. For the intrinsically coupled three-level system these modulations are called quantum beats, whereas for an interference of uncoupled systems they are named polarization interferences. Using nonlinear optical techniques, e.g., four-wave mixing and pump probe, it is possible to distinguish quantum beats and polarization interferences since coupled and uncoupled systems have different optical nonlinearities. (b) The temporal dynamics of the squared modulus of the polarization is shown for systems with two resonances and frequency difference $\Delta\omega_1 - \Delta\omega_2$, neglecting dephasing, i.e., setting $T_2 \to \infty$. After excitation with a short optical pulse at $t=0$ the squared modulus of the polarization is periodically modulated with a time period $2\pi/(\Delta\omega_1 - \Delta\omega_2)$.

consequently no net photocurrent is generated since the currents in opposite directions cancel each other. It is, however, possible to coherently generate photocurrents by nonlinear optical excitation with laser fields that contain two frequencies ω and 2ω which fulfill $\hbar\omega < E_{gap}$ and $2\hbar\omega > E_{gap}$ where E_{gap} is the band gap. Such two-color laser fields couple the initial (valence band) and the final (conduction band) states by two quantum-mechanical excitation pathways, i.e., a single-photon transition of the 2ω field and a two-photon transition of the ω field. With this scheme it is possible to coherently control the amount of excitation by the phase difference between the ω and the 2ω fields and to realize asymmetric distributions of electrons and holes in momentum space, see **Fig. 6(b)**, which correspond to the generation of photocurrents.

Such electrical currents generated by two-color laser excitation do not rely on any specific requirements or symmetry of the material and have been observed in several systems. Using the typical selection rules for III–V semiconductors it is possible to show that linearly polarized ω and the 2ω fields generate electrical currents if they are parallel polarized whereas orthogonal polarization directions lead to pure spin currents, i.e., a motion of spin-up and spin-down electrons in opposite directions.

Due to scattering processes, the optically-generated asymmetric momentum-space distributions relax rapidly and consequently the photocurrents decay on short, typically femtosecond, time scales. For the case of a quantum well, see **Fig. 6(c)**, microscopic calculations including Coulomb and electron-phonon scattering predict decay times on the order of $100 - 200$ fs for the charge and spin currents.

Transient Absorption Changes

In a typical pump–probe experiment one excites the system with a pump pulse (E_p) and probes its dynamics with a weak test pulse (E_t), see **Fig. 7**. With such experiments one often measures the differential absorption $\Delta\alpha(\omega)$ which is defined as the difference between the probe absorption in the presence of the pump $\alpha_{pump\ on}(\omega)$ and the probe absorption without the pump $\alpha_{pump\ off}(\omega)$.

For resonant pumping and for a situation where the pump precedes the test (positive time delays $\tau > 0$), the absorption change is usually negative in the vicinity of the resonance frequency $\Delta\omega$ indicating the effect of absorption bleaching, see **Fig. 8(a)** and **(b)**. There may be positive contributions spectrally around the original absorption line due to resonance broadening and, at other spectral positions, due to excited state absorption, i.e., optical transitions to energetically higher states which are only possible if the system is in an excited state. The bleaching and the positive contributions are generally present in coherent and also in incoherent situations, where the polarization vanishes but occupations in excited states are present.

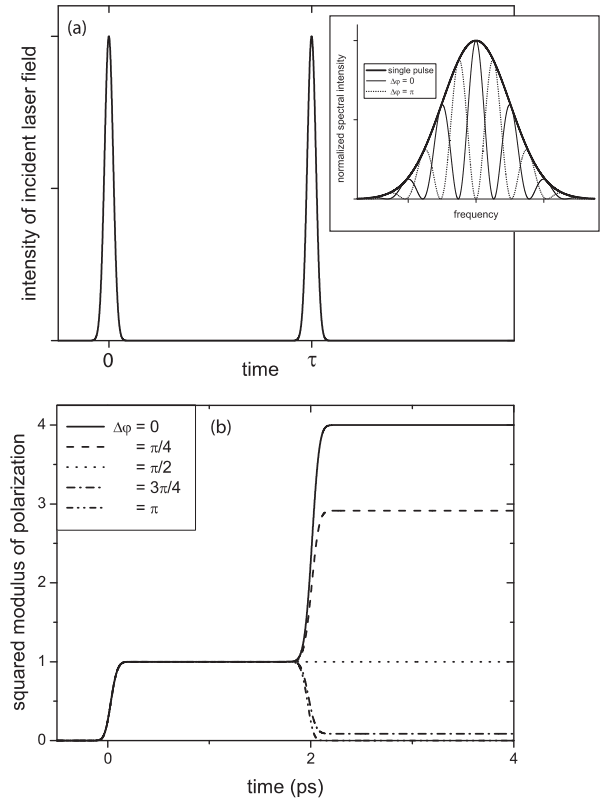


Fig. 5 (a) The interference of two Gaussian laser pulses separated by the time delay τ depends on their phase difference. As shown by the thick solid line in the inset, the spectral intensity of a single pulse is a Gaussian with a maximum at the central frequency of the pulse. The width of this Gaussian is inversely proportional to the duration of the pulse. The spectral intensity of the field consisting of both pulses shows interference fringes, i.e., is modulated with a spectral period that is inversely proportional to τ. As shown by the thin solid and the dotted lines in the inset, the phase of the interference fringes depends on the phase difference $\Delta(\varphi)$ between the two pulses. Whereas for $\Delta(\varphi)=0$ the spectral intensity has a maximum at the central frequency, it vanishes at this position for $\Delta(\varphi)=\pi$. (b) The temporal dynamics of the squared modulus of the polarization is shown for a two-level system excited by a pair of short laser pulses neglecting dephasing, i.e., setting $T_2\rightarrow\infty$. The first pulse excites at $t=0$ an optical polarization with a squared modulus normalized to 1. The second pulse, which has the same intensity as the first one, also excites an optical polarization at $t=2ps$. For $t>2ps$ the total polarization is given by the sum of the polarizations induced by the two pulses. Due to interference, the squared modulus of the total polarization depends sensitively on the phase difference $\Delta(\varphi)$ between the two pulses. For constructive interference the squared modulus of the total polarization after the second pulse is four times bigger than that after the first pulse, whereas is vanishes for destructive interference. One may achieve all values in between these extremes by using phase differences $\Delta(\varphi)$ which are no multiples of π. It is thus shown that the second pulse can be used to coherently control the polarization induced by the first pulse.

For detuned pumping the resonance may be shifted by the light field, as, e.g., in the optical Stark effect. Depending on the excitation configuration and the system, this transient shift may be to higher (blue shift) or lower energies (red shift), see **Fig. 8(a)** and **(b)**. With increasing pump–probe time delay the system gradually returns to its unexcited state and the absorption changes disappear.

As an illustration we show in **Fig. 8(c)** experimentally measured differential absorption spectra of a semiconductor quantum well which is pumped spectrally below the exciton resonance. For a two-level system one would expect a blue shift of the absorption, i.e., a dispersive shape of the differential absorption with positive contributions above and negative contributions below the resonance frequency. This is indeed observed for most polarization directions of the pump and probe pulses. However, if both pulses are oppositely circularly polarized, the experiment shows a resonance shift in the reverse direction, i.e., a red shift. For the explanation of this behavior one has to consider the optical selection rules of the quantum well system. One finds that the signal should actually vanish for oppositely circularly polarized pulses, as long as many-body correlations are neglected. Thus in this case the entire signal is due to many-body correlations and it is their dynamics which gives rise to the appearance of the red shift.

Spectral Oscillations

For negative time delays $\tau<0$, i.e., if the test pulse precedes the pump, the absorption change $\Delta\alpha(\omega)$ is characterized by spectral oscillations around $\Delta\omega$ which vanish as τ approaches zero, see **Fig. 9(a)**. The spectral period of the oscillations decreases with increasing $|\tau|$. **Fig. 9(b)** shows measured differential transmission spectra of a multiple quantum-well structure for different negative time delays. As the differential absorption also the differential transmission spectra are dominated by spectral oscillations around the exciton resonance whose period and amplitude decrease with increasing $|\tau|$.

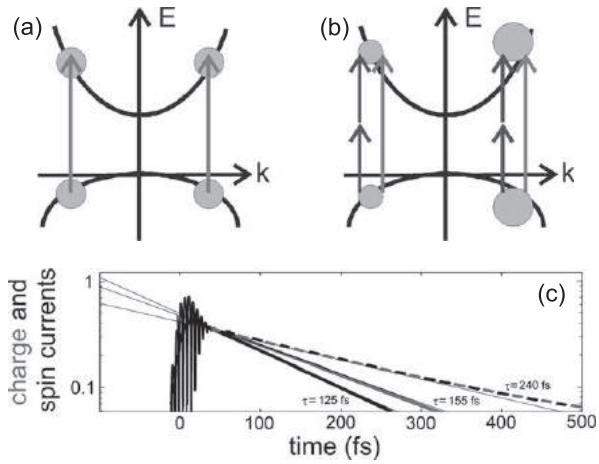


Fig. 6 (a) Schematic illustration of the electron and hole distributions in momentum space that are generated in the conduction and valence band, respectively, by an above band gap optical excitation. Due to the symmetry of the distributions the optical excitation does not lead to a photocurrent. (b) Schematic illustration of the electron and hole distributions in momentum space that are generated by the excitation with a two-color laser field containing the frequencies ω and 2ω. Since the valence and conduction band states are coupled by two quantum-mechanical pathways, i.e., a single and a two-photon transitions, it is possible to coherently control the excitations and to generate asymmetric electron and hole distributions which correspond to the generation of photocurrents. (c) Calculated dynamics of photocurrents that are excited in a GaAs quantum well 150 meV above the band gap by a 20 fs two-color laser field on a logarithmic scale. The simultaneous excitation by the ω and 2ω fields leads to rapid oscillations during the duration of the laser beams. After the excitation, the currents decay approximately exponentially on a 100 – 200 fs time scale. If only electron-LO-phonon scattering is considered, both the charge and spin currents decay with the same time constant, see dashed line. When also electron-electron scattering is included in the analysis, the spin current (dark grey line) decays more rapidly than the charge current (light grey line). Part (c) is taken from **Fig. 2(a)** of Duc, H.T., Meier, T., Koch, S.W., 2005. Phys. Rev. Lett. 95, 086606.

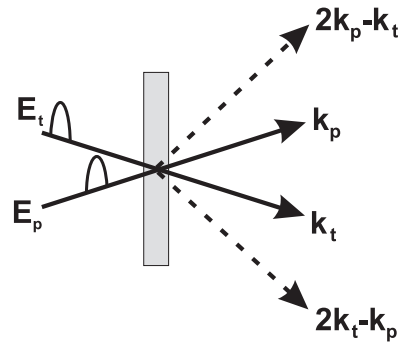


Fig. 7 Transient optical nonlinearities can be investigated by exciting the system with two time-delayed optical pulses. The pump pulse (E_p) puts the system in an excited state and the test (probe) pulse (E_t) is used to measure its dynamics. The dynamic nonlinear optical response may be measured in the transmitted or reflected directions of the pump and test pulses, i.e., $\pm k_p$ and $\pm k_t$, respectively. In a pump–probe experiment one often investigates the change of the test absorption induced by the pump pulse, by measuring the absorption in the direction k_t with and without E_p. One may also measure the dynamic nonlinear optical response in scattering directions like $2k_p - k_t$ and $2k_t - k_p$ as indicated by the dashed lines. This is what is done in four-wave-mixing and photon-echo experiments.

The physical origin of the coherent oscillations is the pump-induced perturbation of the free-induction decay of the polarization generated by the probe pulse. This perturbation is delayed by the time τ at which the pump arrives. The temporal shift of the induced polarization changes in the time domain leads to oscillations in the spectral domain, since the Fourier transformation translates a delay in one domain into a phase factor in the conjugate domain.

Photon Echo

In the nonlinear optical regime one may (partially) reverse the destructive interference of a coherent, inhomogeneously broadened polarization. For example, in four-wave mixing, which is often performed with two incident pulses, one measures the emitted field in a background-free scattering direction, see **Fig. 5**. The first short laser pulse excites all transitions at $t=0$. As a result of the inhomogeneous broadening the polarization decays due to destructive interference, see **Fig. 3**. The second pulse arriving at $t=\tau>0$ is able to conjugate the phases ($e^{i(\varphi)} \rightarrow e^{-i(\varphi)}$) of the individual polarizations of the inhomogeneously broadened system. The subsequent unperturbed dynamical evolution of the polarizations leads to a measurable macroscopic signal at $t=2\tau$. This photon

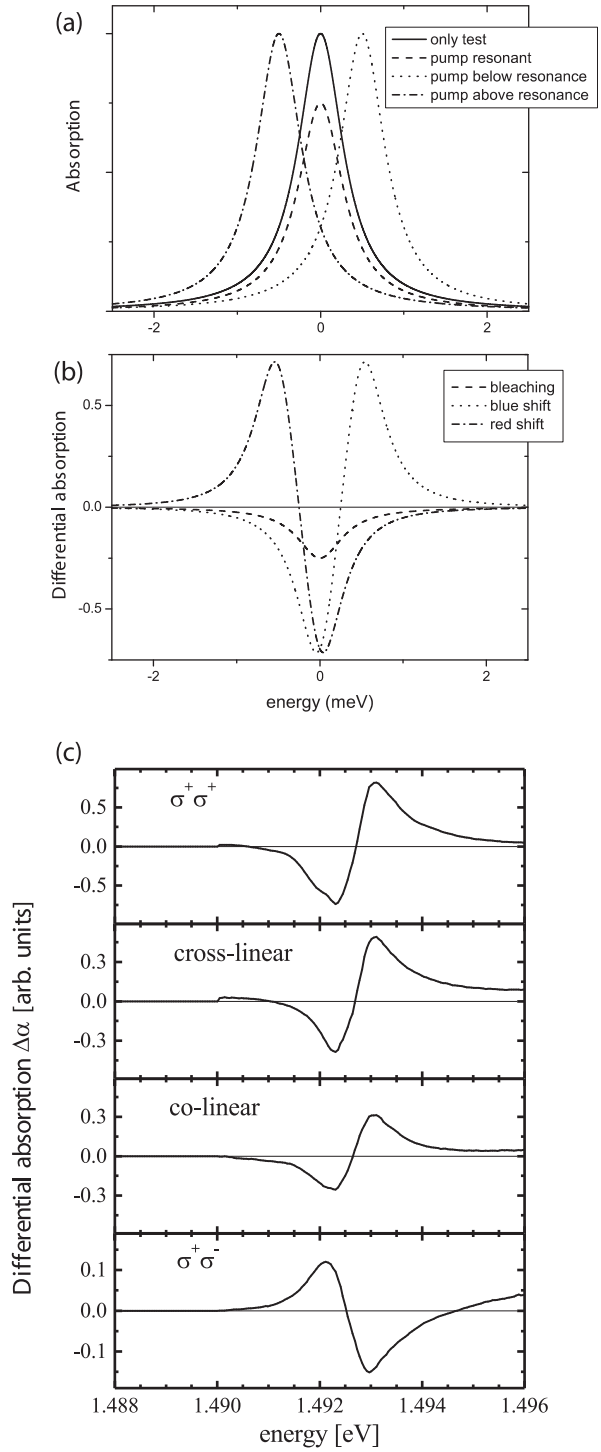


Fig. 8 (a) Absorption spectra of a two-level system. The solid line shows the linear optical absorption spectrum as measured by a weak test pulse. It corresponds to a Lorentzian line that is centered at the transition frequency $\Delta\omega$, which has been set to zero. The width of the line is inversely proportional to the dephasing time T_2. In the presence of a pump pulse which resonantly excited the two-level system, the absorption monitored by the test pulse is reduced in amplitude, i.e., bleached, since the pump puts the system in an excited state (dashed line). In the optical Stark effect the frequency of the pump is non-resonant with the transition frequency. If the pump is tuned below (above) the transition frequency, the absorption is shifted to higher (lower) frequencies, i.e., shifted to the blue (red) part of the spectrum, see dotted (dashed-dotted) line. (b) The differential absorption obtained by taking the difference between the absorption in the presence of the pump and the absorption without pump. The dashed line shows the purely negative bleaching obtained using a resonant pump. The dotted and the dashed-dotted lines correspond to the dispersive shape of the blue and red shift obtained when pumping below and above the resonance, respectively. (c) Measured differential absorption spectra of a semiconductor quantum well which is pumped off-resonantly 4.5 meV below the 1s heavy-hole exciton resonance. The four lines correspond to different polarization directions of the pump and probe pulses as indicated. Part (c) is taken from Fig. 1 of Sieh, C., Meier, T., Jahnke, F., *et al.*, 1999. Phys. Rev. Lett. 82, 3112.

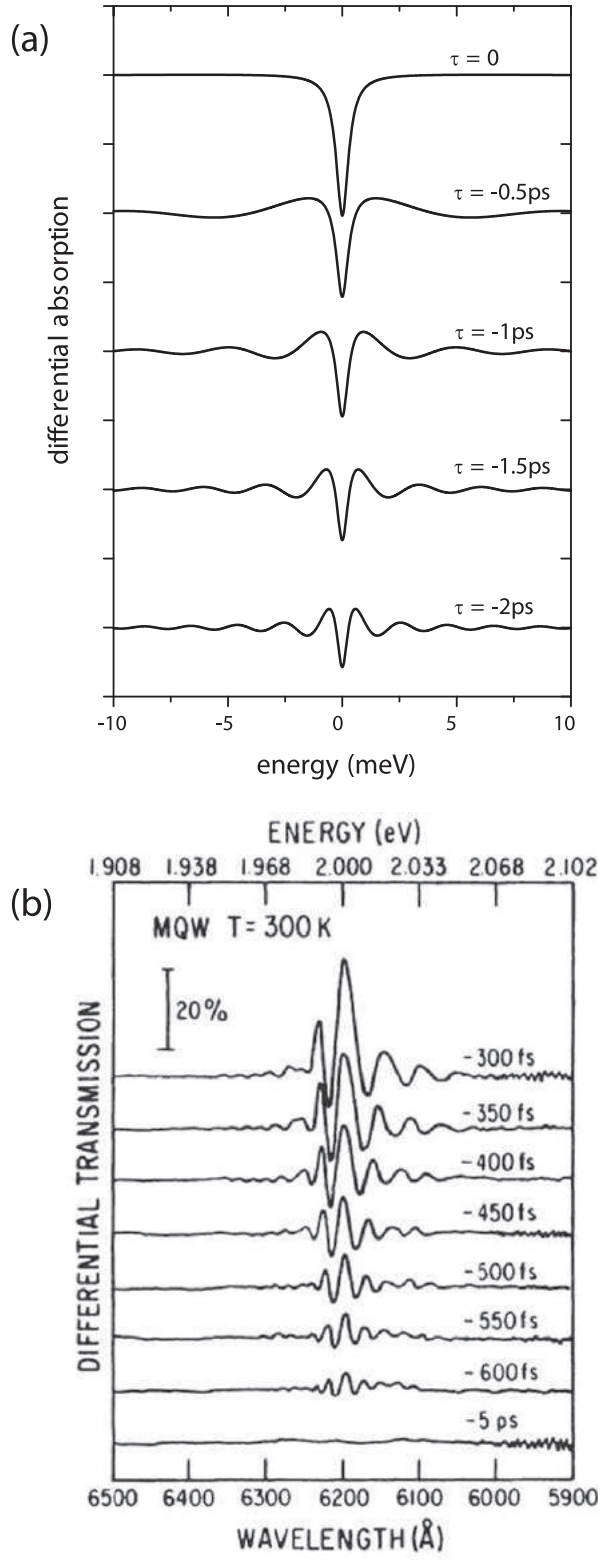


Fig. 9 (a) Differential absorption spectra are shown for resonant pumping of a two-level system for various time delays τ between the pump and the test pulse. When the pump precedes the test, i.e., $\tau<0$, the differential absorption exhibits spectral oscillations with a period which is inversely proportional to the time delay. When τ approaches zero, these spectral oscillations vanish and the differential absorption develops into purely negative bleaching. (b) Measured differential transmission spectra of a multi quantum well for different negative time delays as indicated. Part (b) is taken from **Fig. 3** of Sokoloff, J.K., Joffre, M., Fluegel, B., *et al.* 1988. Phys. Rev. B 38, 7615.

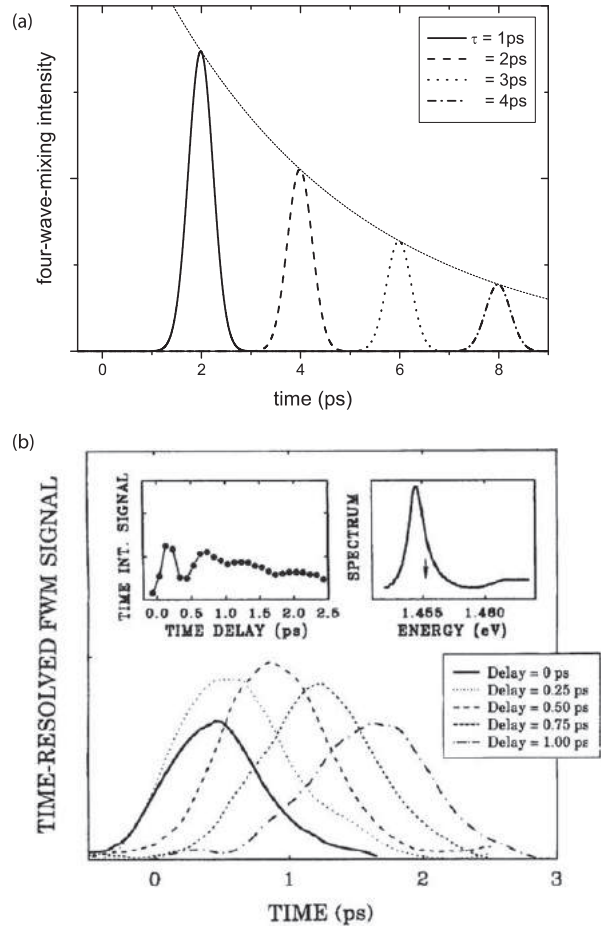


Fig. 10 (a) The intensity dynamics as measured in a four-wave-mixing experiment on an ensemble of inhomogeneously broadened two-level systems for various time delays τ of the two incident pulses. The first pulse excites the system at $t=0$. In the time period $0<t<\tau$ the individual polarizations of the inhomogeneously broadened system oscillate with their respective frequencies and due to destructive interference the total polarization decays, cp. **Fig. 3**. The second pulse arriving at $t=\tau$ leads to a partial rephasing of the individual polarizations. Due to phase conjugation all individual polarizations are in phase at $t=2\tau$ which results in a macroscopic measurable signal, the photon echo. Due to dephasing processes the magnitude of the photon echo decays with increasing τ, see solid, dashed, dotted and dashed-dotted lines. Describing the dephasing via a T_2 time corresponds to an exponential decay as indicated by the thin dotted line. By measuring the decay of the echo with increasing τ, one thus gets experimental information on the dephasing of the optical polarization. (b) Time-resolved four-wave-mixing signal measured on an inhomogeneously broadened exciton resonance of a semiconductor quantum well structure for different time delays as indicated. The origin of the time axis starts at the arrival of the second pulse, i.e., for a non-interacting system the maximum of the echo is expected at $t=\tau$. The insets show the corresponding time-integrated four-wave-mixing signal and the linear absorption. Part (b) is taken from **Fig. 1** of Jahn ke, F., Koch, M., Meier, T., *et al.* 1994. Phys. Rev. B 50, 8114.

echo occurs since at this point in time all individual polarizations are in phase and add up constructively, see **Fig. 10(a)**. Since this rephasing process leading to the echo is only possible as long as the individual polarizations remain coherent, one can analyze the loss of coherence (dephasing) by measuring the decay of the photon echo with increasing time delay.

Time-resolved four-wave-mixing signals measured for different time delays on an inhomogeneously broadened exciton resonance of a semiconductor quantum well structure are presented in **Fig. 10(b)**. Compared to **Fig. 10(a)**, the origin of the time axis starts at the arrival of the second pulse, i.e., for a non-interacting system the maximum of the echo is expected at $t=\tau$. Due to the inhomogeneous broadening the maximum of the signal is shifting to longer times with increasing delay. Further details of the experimental results, in particular, the exact position of the maximum and the width of the signal, cannot be explained on the basis of non-interacting two-level systems, but require the analysis of many-body effects in the presence of inhomogeneous broadening.

Multidimensional Fourier Transform Spectroscopy

Exciting a semiconductor with N short laser pulses from different directions may lead to a nonlinear signal that depends on real time and $N-1$ time delays between the incident pulses. The idea of multidimensional Fourier transform spectroscopy is to Fourier transform the time-domain signal with respect to some or all of the time and time delay arguments to the frequency domain.

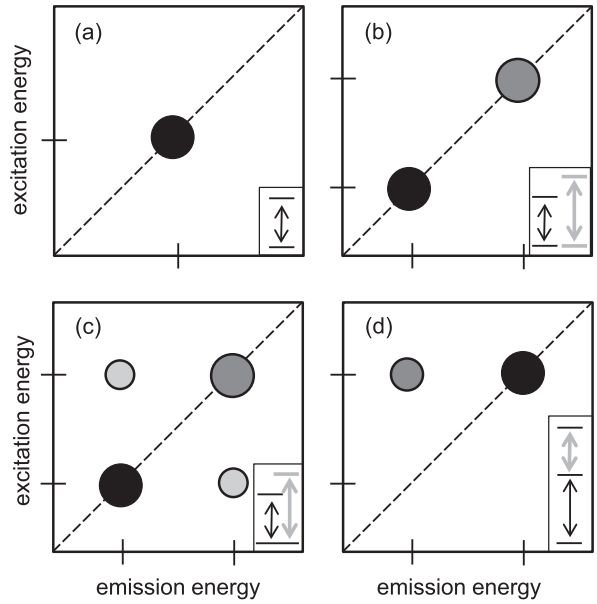


Fig. 11 Schematical illustration of two-dimensional Fourier-transform spectroscopy signals for a few-level systems. (a) For the case of a single two-level system the absorption and emission take place at the same frequency which leads to a single peak on the diagonal of the two-dimensional signal. (b) Two uncoupled two-level systems absorb and emit at two frequencies and therefore correspond to two peaks on the diagonal. (c) Similarly to case (b), a three-level system can also absorb and emit at two frequencies. However, for a three-level system the coupling among the transitions via the common ground state leads to the appearance of two additional off-diagonal peaks. (d) When the excitation to an energetically higher state is only possible if the system is already excited, this transition frequency appears as an additional off-diagonal peak in the emission. Such a situation can be realized in systems that contain, e.g., exciton and biexciton resonances.

The so-obtained multi-dimensional frequency-domain signal depends on up to N frequencies and contains correlations among the frequency components which arise due to the dynamics in the respective time delays. Such multi-dimensional signals have been shown to provide a transparent way to unravel couplings and interaction processes that determine the complex nonlinear dynamics of several investigated systems.

Fourier transform spectroscopy is frequently performed in four-wave-mixing set-ups where the system is excited by two or three short laser pulses. Fourier transforming with respect to time and time delays provides signals that depend on two or three frequency arguments corresponding to two- or three-dimensional spectroscopy, respectively. A detailed analysis of such signals and theoretical modelling is able to provide valuable information on the nature and the coupling among the optical excitations, their homogeneous and inhomogeneous broadening, many-body effects, etc.

In the most simple example of two-dimensional Fourier transform spectroscopy, one may interpret the two frequencies arguments as excitation and emission frequencies. Optical transitions that can be excited directly from the ground state and contribute to the nonlinear emission appear as peaks on the diagonal corresponding to identical excitation and emission frequencies, see **Fig. 11(a)** and **(b)** for the cases of a single and two two-level systems, respectively. Consequently, inhomogeneous broadening, cp. Section 3.3, leads to a broadening of the signal along the diagonal. A coupling of optical transitions, e.g., by sharing a common state as in a three-level system, leads to off-diagonal peaks appearing in Fourier transform spectroscopy, see **Fig. 11(c)**. Also transitions that are only possible in an already excited system lead to off-diagonal peaks that describe emission at frequencies where the system has no linear absorption, see **Fig. 11(d)**.

See also: Applications in Semiconductors. Excitons

Further Reading

Allen, L., Eberly, J.H., 1975. Optical Resonance and Two-Level Atoms. New York: Wiley.
Baumert, T., Helbing, J., Gerber, G., *et al.*, 1997. Coherent control with femtosecond laser pulses. Adv. Chem. Phys. 101, 47–82.
Bergmann, K., Theuer, H., Shore, B.W., 1998. Coherent population transfer among quantum states of atoms and molecules. Rev. Mod. Phys. 70, 1003–1025.
Bloembergen, N., 1965. Nonlinear Optics. New York: Benjamin.
Brewer, R.G., 1977. Coherent optical spectroscopy. In: Balian, *et al.* (Eds.), Course 4 in Les Houches Session XXVII, Frontiers in Laser Spectroscopy 1. Amsterdam: North-Holland.
Cohen-Tannoudji, C., Dupont-Roc, J., Grynberg, G., 1989. Photons and Atoms. New York: Wiley.
Ernst, R.R., Bodenhausen, G., Wokaun, A., 1987. Principles of Nuclear Magnetic Resonance in One and Two Dimensions. New York: Oxford Science.

Haug, H., Koch, S.W., 2009. Quantum Theory of the Optical and Electronic Properties of Semiconductors, 5th ed Singapore: World Scientific.
Koch, S.W., Peyghambarian, N., Lindberg, M., 1988. Transient and steady-state optical nonlinearities in semiconductors. J. Phys. C 21, 5229–5250.
Macomber, J.D., 1976. The Dynamics of Spectroscopic Transitions. New York: Wiley.
Meier, T., Duc, H.T., Vu, Q.T., *et al.*, 2008. Ultrafast dynamics of optically-induced charge and spin currents in semiconductors. Adv. Solid State Phys. 46, 199–210.
Mukamel, S., 1995. Principles of Nonlinear Optical Spectroscopy. New York: Oxford.
Peyghambarian, N., Koch, S.W., Mysyrowicz, A., 1993. Introduction to Semiconductor Optics. Englewood Cliffs, NJ: Prentice-Hall.
Pötz, W., Schroeder, W.A. (Eds.), 1999. Coherent Control in Atoms, Molecules, and Semiconductors. Dordrecht: Kluwer.
Schäfer, W., Wegener, M., 2002. Semiconductor Optics and Transport Phenomena. Berlin: Springer.
Shah, J., 1999. Ultrafast Spectroscopy of Semiconductors and Semiconductor Nanostructures, 2nd ed Berlin: Springer.
Stone, K.W., Gundogdu, K., Turner, D.B., *et al.*, 2009. Two-quantum 2D FT electronic spectroscopy of biexcitons in GaAs quantum wells. Science 324, 1169–1173.
Yeazell, J.A., Uzer, T. (Eds.), 2000. The Physics and Chemistry of Wave Packets. New York: Wiley.
Zhang, T., Kuznetsova, I., Meier, T., *et al.*, 2007. Polarization-dependent optical two-dimensional fourier transform spectroscopy of semiconductors. Proc. Natl Acad. Sci. USA 104, 14227–14232.

Nonlinear Optics in Disordered Media: Anderson Localization

Arash Mafi, University of New Mexico, Albuquerque, NM, United States

Introduction

Is Anderson localization preserved, enhanced, weakened, or destroyed in the presence of nonlinearity? In this short article, we will learn that there may not be a straightforward answer to this question. A survey of the literature published on this subject reveals that the answer to this rather basic question can best be described as inconclusive at this point. Of course, it should not be a surprise that the potentially chaotic behavior as a result of the nonlinear dynamics is largely behind the present state of debate in the literature on the interplay between disorder and nonlinearity in the context of Anderson localization.

This article presents a survey of some highlights in the literature over the past few years with the hope that an eventual conclusive answer can be found to such a fundamental question in a not so distant future. The question can be cast in a limited form dealing only with the optical Kerr effect, which is of greater interest to the photonics community, or in a more general context to explore and differentiate the impact of various forms of nonlinearity on localization. The interplay between disorder and nonlinearity has been explored over the years in various contexts and different disciplines. Some inevitable overlap remains in the existing literature belonging to different fields of study, primarily due to similarities among the dynamical equations of different physical systems. For example, the Gross–Pitaevskii equation describing the Bose-Einstein condensate is similar to the nonlinear Schrödinger equation describing the propagation of light in a nonlinear optical medium.

Anderson localization in general, and even Anderson localization of light in particular are broad research topics. In the optical domain, the most conclusive results on localization have so far been obtained in the context of transverse Anderson localization of light, where light is transversely localized in a waveguide-like geometry with a disordered transverse profile and propagates freely in a longitudinally uniform medium. Anderson localization is more readily observed in this transverse form as explained in the next section. Moreover, the transversely localized and longitudinally propagating platform allows for a relatively long-distance interaction of light with the optical medium, resulting in an appreciable nonlinear phase shift. As such, we will primarily focus on results that are more relevant to the impact of nonlinearity on transverse Anderson localization. For more details, we refer the interested reader to an excellent review article on this subject by Fishman, Krivolapov, and Soffer and references therein.

Anderson Localization of Light

Anderson localization is the absence of diffusive wave transport in highly disordered scattering media. Its origin dates back to a theoretical study conducted by P. W. Anderson in 1958 who investigated the behavior of spin diffusion and electronic conduction in random lattices. It was soon realized that because the novel localization phenomenon is due to the wave nature of the quantum mechanical electrons scattering in a disordered lattice, it can also be observed in other coherent wave systems, including classical ones.

The fact that Anderson localization was deemed possible in non-electronic systems was encouraging, given that the observation of disorder-induced localization was shown to be inhibited by thermal fluctuations and nonlinear effects in electronic systems. Subsequently, localization was studied in various classical systems including in acoustics, elastics, electromagnetics, and optics. It was also recently investigated in various quantum optical systems, such as atomic lattices and propagating photons.

General studies of Anderson localization have revealed that waves in one-dimensional (1D) and two-dimensional (2D) unbounded disordered systems are always localized. However, in order for a three-dimensional (3D) random wave system to localize, the scattering strength needs to be larger than a threshold value. This statement is often cast in the form of $kl^* \sim 1$, where k is the effective wavevector in the medium, and l^* is the wave scattering transport length. This is referred to as the Ioffe-Regel condition and shows that in order to observe Anderson localization, the disorder must be strong enough such that the wave scattering transport length becomes on the order of the wavelength. The Ioffe-Regel condition is notoriously difficult to satisfy in 3D disordered optical media. For the optical field to localize in 3D, very large refractive index contrasts are required that are not generally available in low-loss optical materials. The fact that Anderson localization is hard to achieve in 3D optical systems may be a blessing in disguise; otherwise, no sunlight would reach the earth on highly cloudy days.

Transverse Anderson Localization of Light

Unlike 3D lightwave systems in which the observation of the localization is prohibitively difficult, observation of Anderson localization in a quasi-2D optical system is readily possible, as was first shown by Abdullaev *et al.* and De Raedt *et al.* In particular, De Raedt *et al.* showed that 2D Anderson localization can be observed in a dielectric with transversely random and longitudinally uniform refractive index profile. An optical field that is launched in the longitudinal direction tends to remain localized in the transverse plane as it propagates in the longitudinal direction in a transversely random dielectric medium. This behavior was dubbed transverse Anderson localization of light.

 doi:10.1016/B978-0-12-803581-8.09446-7

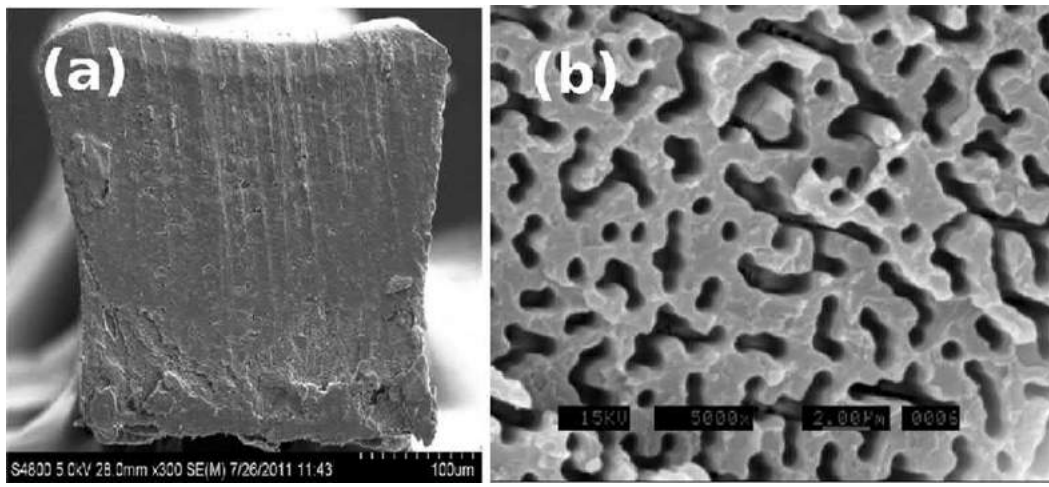


Fig. 1 (a) Cross section view of the polymer Anderson localization optical fiber developed by Karbasi *et al.* with a nearly square profile and an approximate side width of 250 μm; (b) zoomed-in scanning electron microscope image of a 24 μm wide region on the tip of the fiber exposed to a solvent to differentiate between PMMA and PS polymer components, where feature sizes are around 0.9 μm and darker regions are PMMA. Adapted with permission from Karbasi, S., Hawkins, T., Ballato, J., Koch, K.W., Mafi, A., 2012a. Transverse Anderson localization in a disordered glass optical fiber. Opt. Mater. Express 2, 1496–1503. doi:10.1364/OME.2.001496; Karbasi, S., Mirr, C.R., Yarandi, P.G., *et al.* 2012c. Observation of transverse Anderson localization in an optical fiber. Opt. Lett. 37, 2304–2306. doi:10.1364/OL.37.002304.

In their pioneering work, Schwartz *et al.* wrote the transversely disordered and longitudinally invariant refractive index profiles in a photorefractive crystal using a laser beam. They used another probe beam to investigate the transverse localization behavior. Their experiment was quite interesting as it allowed them to vary the disorder level by controlling the laser illumination of the photorefractive crystal in a controlled fashion to observe the onset of the transverse localization and the change in the localization radius as a function of the disorder level. The transverse localization of the beam and the free longitudinal propagation due to the longitudinal invariance of the dielectric medium strongly resembled the optical waveguides; therefore, applications of Anderson localization for light propagation in an optical fiber-like medium seemed an appealing extension of these ideas.

In 2012, Karbasi *et al.* reported the first observation of transverse Anderson localization in an optical fiber. In order to obtain large refractive index fluctuations required for a strong transverse localization, they randomly mixed 40,000 pieces of polymethyl methacrylate (PMMA) fiber with refractive index of 1.49, and 40,000 pieces of polystyrene (PS) fiber with refractive index of 1.59, and redrew the random stack to a 250-μm-wide optical fiber with random index fluctuations equal to 0.1 and an approximate transverse random feature size equal to 0.9 μm (see **Fig. 1**). The large index contrast of 0.1 ensured that both the beam localization radii and the sample-to-sample fluctuations over the ensemble of localized beam radii are sufficiently small to ensure that disordered fiber operates as a genuine optical fiber – they observed that the localized beam radii launched at different transverse positions over the facet of Anderson localizing fiber were all small and were nearly identical. Later, Karbasi *et al.* reported the first observation of Anderson localization in a silica optical fiber as well. The reported glass-air disordered fiber was made from "satin quartz", which is a porous artisan glass. After drawing the preform, the airholes (bubbles) in the glass were stretched to form the hollow air-rods required for transverse Anderson localization. The large draw ratio sufficiently preserved the longitudinal invariance, without significant disturbance over typical lengths used in the experiments.

Transverse Anderson Localization and Kerr Nonlinearity in Theory and Experiment

Concrete experimental results on the interplay between disorder and nonlinearity are rather sparse. Even in the very few available published work in the literature, one does not get a clear sense on the extent to which Anderson localization is affected by nonlinearity. This is most likely due to the large parameter space and the possibility of many configurations under which the studies can be performed. In this article, we review the results of two pioneering works by Schwartz *et al.* and Lahini *et al.*, involving both experiment and theory, and then examine further theoretical considerations on this issue. The interaction between disorder and other forms of nonlinearity (non-Kerr) will be discussed in the next section.

Earlier, we mentioned the numerical and experimental work of Schwartz *et al.* that resulted in the observation of transverse Anderson localization of light. The authors also investigated the transverse Anderson localization of light in the presence of Kerr nonlinearity, both numerically and experimentally. The defining equation for the nonlinear propagation of light is the nonlinear Schrödinger equation (NLSE) which includes a Kerr nonlinearity term:

$$i\frac{\partial A}{\partial z}+\frac{1}{2n_0k_0}\left[\nabla_T^2A+k_0^2\left(n^2-n_0^2\right)A\right]+k_0n_2|A|^2A=0 \tag{1}$$

where A is slow-varying amplitude of the optical field, k_0 is the vacuum wavevector, z is the propagation distance, T represents the transverse coordinates (x and y here), $n(x,y)$ is the transversely disordered refractive index profile, n_0 is the mean value of the refractive index around which index fluctuation occur, and n_2 is the nonlinear index, which is positive for self-focusing and negative for self-defocusing nonlinearity.

As a case study, the authors considered a disordered lattice where the maximum contribution of the nonlinear term to the index change "$\max(|n_2| \times |A|^2)$" was assumed to be a maximum of 15% of the index contrast of the underlying periodic waveguide. They also varied the disorder level from 0 to 30%, where the disorder level was defined as the magnitude of random index fluctuations relative to the index contrast of the underlying periodic waveguide. Using numerical simulations, they observed that over this range, the self-defocusing nonlinearity (n_2) results in a moderate (nearly negligible) widening of the average beam profile. They reported that after a short propagation distance, the beam broadens and the remaining propagation is essentially as if the medium were linear, primarily because the intensity is much lower and the nonlinear effects do not play a role anymore. In contrast, the self-focusing nonlinearity ($n_2 > 0$) resulted in a substantial reduction of the average localized beam diameter. The enhancement of localization due to the self-focusing nonlinearity was particularly noticeable when the disorder level was less than 15%. They observed that at high disorder levels, where localization takes place, the difference between linear and nonlinear propagation is reduced, and the behavior is dominated primarily by disorder.

The experiments were performed at different nonlinear strengths. Consider α as the ratio between the peak intensity of the probe beam and the maximum intensity of the lattice-forming beams in the photorefractive crystal. The experiments were carried out for values of $\alpha = 1$, 2, and 3. The statistical analysis of the localized beam radius clearly confirmed the expected reduction in the average beam radius due to the self-focusing nonlinearity. They observed that self-focusing enhances localization, altering the intensity profile from diffusive-like to exponentially decaying. They increased the nonlinearity (increased α) and observed that the intensity profile narrowed down accordingly. At $\alpha = 3$, the output beam profile resembles the input profile, suggesting the formation of a soliton.

Similar results were reported by Lahini *et al.* using disordered 1D waveguide lattices, as shown in **Fig. 2**. Their experiment consisted of a 1D lattice of $N = 99$ coupled optical waveguides patterned on an AlGaAs substrate. The nonlinear Schrödinger equation governing the propagation of light in this coupled lattice system can be expressed as

$$i\frac{\partial U_n}{\partial z} = \beta_n U_n + C(U_{n+1} + U_{n-1}) + \gamma |U_n|^2 U_n \tag{2}$$

where n is the index labeling lattice sites (waveguides), U_n is the optical field amplitude at site n, β_n is the propagation constant associated with the nth waveguide, C is the tunneling rate between adjacent sites, and z is the longitudinal space coordinate. γ characterizes the nonlinear Kerr effect. Disorder was introduced to the lattice by randomly changing the width of each waveguide, resulting in the randomization of the propagation constants over a range of $\beta_0 \pm \Delta$, where β_0 is the propagation constant for a mean value of the waveguide width. Δ/C is defined as the disorder strength, with the implicit assumption that the coupling coefficient between adjacent waveguides is nearly identical.

Light was injected into one or a few waveguides at the input, and light intensity distribution was measured at the output. In the linear regime, Lahini *et al.* identified that highly localized eigenmodes exist near the top edge of the propagation constant band with a flat-phase profile and near the bottom edge of the band with a highly varying-phase profile, having phase flips between adjacent sites. The modes near the middle of the band were not as localized and also showed some phase variations. In the weak nonlinear regime, Lahini *et al.* observed that nonlinearity enhances localization in flat-phased modes and induces delocalization in the staggered modes. This behavior is explained as follows: the presence of the weak nonlinearity perturbatively shifts (increases) the value of the propagation constant of each localized mode. For the flat-phased modes, the nonlinearity shifts the modes outside the original linear spectrum. However, for the staggered, which belong to the bottom edge of the propagation constant band, a perturbative increase in the value of the propagation constant shifts it further inside the original linear spectrum. Therefore, the propagation constant of a staggered mode can cross and resonantly couple with other modes of the lattice, resulting in delocalization.

The effect of nonlinear perturbations on localized eigenmodes was studied by Lahini *et al.* experimentally via exciting a pure localized mode and increasing the input beam power. The intensities were kept far below the self-focusing threshold in the

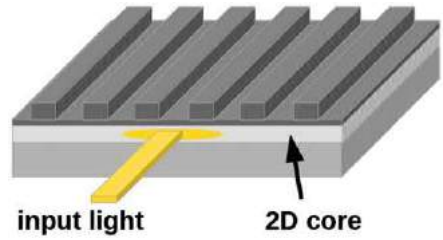


Fig. 2 Schematic view of the disordered lattice waveguide used in the experiment by Lahini *et al.* Reproduced from Lahini, Y., Avidan, A., Pozzi, F., *et al.* 2008. Anderson localization and nonlinearity in one-dimensional disordered photonic lattices. Phys. Rev. Lett. 100, 013906. doi:10.1103/PhysRevLett.100.013906.

periodic lattice. They observed that some of the localized modes exhibited significant response to nonlinearity. The experiments showed that weak positive nonlinearity tends to further localize flat phased localized modes, but tends to delocalize staggered modes, consistent with the theoretical prediction above. The results obtained by Lahini *et al.* do not contradict the general observations of Schwartz *et al.*; rather, they argue that the enhanced localization due to nonlinearity by Schwartz *et al.* is only applicable to certain optical modes.

We would like to point out an early work by Pertsch *et al.* in 2004 who experimentally investigated light propagation in a disordered two-dimensional array of mutually coupled optical fibers. In the linear case they observed that light either spreads in a diffusive manner or localizes at a few sites. The absence of Anderson localization was most likely due to the fact that the disorder was not sufficiently high to localize the field whiting the transverse dimensions of the fiber. However, they reported that for high excitation power diffusive spreading is arrested by the focusing nonlinearity and a discrete soliton is formed.

Further Theoretical Considerations

In 2008, Pikovsky and Shepelyansky presented a somewhat different account of the interaction between disorder and nonlinearity. In their analysis, they considered a discrete nonlinear Schrödinger equation with a third-order diagonal Kerr nonlinear term, isomorphic to Eq. (2), to study the temporal evolution of quantum wavefunction on coupled lattices in the form of

$$i\frac{\partial\psi_n}{\partial t} = E_n\psi_n + \beta|\psi_n|^2\psi_n + V(\psi_{n+1} + \psi_{n-1}) \tag{3}$$

where β characterizes nonlinearity, V is the hopping matrix element and is deterministic and identical for all terms. The disorder is introduced through the diagonal terms, where they are assumed to be randomly and uniformly distributed in the range $-W/2 < E_n < W/2$, and

$$\sum_n |\psi_n|^2 = 1 \tag{4}$$

is assumed. For optical waveguides, this equation describes the propagation of light in a coupled waveguide lattice, where the coupling strengths among the waveguides are identical but the propagation constants of individual waveguides are randomized by manipulating the geometry or the refractive index of each individual waveguide. Therefore, this conforms well with the problem studied by Lahini *et al.* described earlier. The normalization condition means that the total optical power remains unchanged as the light propagates through the coupled waveguide lattice. The authors verified that in the absence of nonlinearity ($\beta = 0$) and in the presence of weak disorder, all eigenstates are exponentially localized as expected from Anderson localization theory with the localization length $l \approx 96(V/W)^2$ at the center of the energy band.

Pikovsky and Shepelyansky considered the spreading of a field that is initially localized at the central lattice point with $|\psi_n(0)|^2 = \delta_{n,0}$. Of course, this problem is fully studied in the limit of no disorder with $W = 0$ and a general value of nonlinearity $\beta \neq 0$, e.g., in the context of nonlinear light propagation in discrete lattices, where diffraction and nonlinearity can give rise to interesting physics including diffraction-free propagation and self-localized states or discrete solitons. The opposite limit of no nonlinearity with $\beta = 0$ in the presence of disorder $W \neq 0$ is also well studies in the context of Anderson localization. The intermediate regime is the focus of the work presented by Pikovsky and Shepelyansky.

Intuitively speaking, one may think that in a nonlinear disordered coupled waveguide system, the dynamics of the beam is initially influenced by nonlinearity; and as the beam spreads, the effect of nonlinearity becomes weaker and the disorder dynamics takes over. Therefore, one should always expect Anderson localization after a sufficiently long time. The analysis by Pikovsky and Shepelyansky is presented in view of earlier work by Sanchez-Palencia *et al.* who similarly argued for the localization of the field at large time scales, consistent with the intuitive view discussed above. However, Sanchez-Palencia *et al.* argued that the high-momentum cutoff of the Fourier transform of the correlation function for 1D speckle potentials can change localization from exponential to algebraic. Per argument presented in by Pikovsky and Shepelyansky, at first glance, one may think that Kerr nonlinear effects may be favored strongly by localized field configurations because if the field spreads over $\tilde{N}$ lattice sites, the conservation of power implies that the Kerr nonlinearity should scale as $|\psi_n|^2 \sim 1/\tilde{N}$. However, Pikovsky and Shepelyansky argue that the nonlinear frequency shift $\beta|\psi_n|^2 \sim \beta/\tilde{N}$ should be compared with the characteristic distance between the frequencies of the exponentially localized modes, which scales as $1/\tilde{N}$; therefore, the effect of nonlinearity persists and does not quantitatively depend on the width of the field distribution and is omnipresent.

Using a theoretical argument, they suggested that in the presence of non-vanishing nonlinearity, the beam spreading characterized by

$$\sigma^2 = \sum_n (n - \langle n\rangle)^2|\psi_n|^2 \tag{5}$$

should follow a subdiffusive behavior where $\sigma^2 \propto t^\alpha$ with $\alpha = 2/5$. They verified their theoretical argument via numerical integration of Eq. (3) and monitoring the results, up to $t = 10^8$. For the numerical simulation they used the boundary condition $\psi_n(t=0) = \delta_{n,0}$, where $n = 0$ represents the middle waveguide, and the integration is performed by the operator splitting method. In a sample

set of simulations they chose the nonlinearity strength to be $\beta=V$ for two cases: case 1 with $W=2$; and case 2 with $W=2$. As expected, the initial expansion was ballistic for either case, but after some time t_0, the expansion became subdiffusive in either case. They fit the subdiffusive expansion to $\sigma \propto t^a$ over the range $t_0<t<10^8$ to obtain an estimate for the subdiffusive exponent. In case 1, for different instances of randomness, they obtained $0.32 \leq \alpha \leq 0.39$; and for case 2 they reported $0.28 \leq \alpha \leq 0.41$. Upon averaging over 8 independent realizations, they reported a fit of the form $57.5 \times t^{0.344}$ for case 1 and $8.7 \times t^{0.306}$ for case 2 over the subdiffusive ranges.

Pikovsky and Shepelyansky analyzed the probability distribution $w_n=|\psi_n|^2$ over lattice sites n; in the absence of nonlinearity they observed an exponential decay with an exponent consistent with the Anderson localization. In the presence of nonlinearity, however, they observed a rather flat distribution in the vicinity of $n=0$ where the width of the flat distribution depended on the value of β. As expected, the (low-intensity) tails of the distribution always followed a decay exponent consistent with the linear theory of Anderson localization. Another important observation reported by Pikovsky and Shepelyansky was the presence of a certain critical strength β_c for nonlinearity, beyond which the reported delocalization behavior happens. The numerical solutions suggested a value of $\beta_c \approx 0.1$ for this threshold value. However, Pikovsky and Shepelyansky argued that the threshold-like behavior might not be perfect and a slow spreading may persist all the way down to very small values of nonlinearity albeit at an extremely slow rate that could not be detected in the numerical integration scheme.

In another study published by Kopidakis *et al.* on the spreading of an initially localized wave packet, the authors confirmed that while there are many initial conditions such that the second moment of the norm in Eq. (5) and energy density distributions diverge with time in agreement with Pikovsky and Shepelyansky, the participation number of a wave packet does not diverge simultaneously. The participation number is defined as

$$P = \frac{\left(\sum_n |\psi_n|^2\right)^2}{\sum_n |\psi_n|^4} \tag{6}$$

which is an alternative measure of the wave spreading to Eq. (5). Kopidakis *et al.* prove this result analytically for norm-conserving models that satisfy Eq. (4) with strong enough nonlinearity. They showed that initially localized wave packets with a large enough amplitude cannot spread to arbitrarily small amplitudes. The consequence is that a part of the initial energy must remain well focused at all times.

A later report by Fishman *et al.* provide a somewhat different account of the impact of nonlinearity on Anderson localization in disordered lattices. Fishman *et al.* present a thorough survey of the many subtleties involved regarding the interaction of nonlinearity and disorder. The conclusion is that the situation can best be described as inconclusive at this point. For example, when using the numerical simulations, they caution that Eq. (3) is chaotic with an exponential sensitivity to numerical errors. Because of the possibly chaotic motion due to the presence of nonlinearity, Fishman *et al.* argue that the numerical solutions of Eq. (3) are not actual solutions. In order to draw conclusions similar to those by Pikovsky and Shepelyansky, one must assume that the numerical solutions are statistically similar to the correct solutions with no real theoretical support for this assumption. For long-time (or long-distance propagation in waveguides), it is not clear that reducing the t step size can control the cumulative numerical error, given that the limit of zero t step may be singular. Details of these arguments can be found in the review article by Fishman *et al.*

Other Forms of Nonlinearity

Other forms of nonlinearity besides Kerr can also interact with the disorder-induced localization. Recently Leonetti *et al.* explored the impact of thermally induced nonlinearity on Anderson localization for a beam of light propagating in the polymer Anderson localizing optical fiber of **Fig. 1**. They reported the observation of a self-focusing action occurring in the disordered fiber with the binary index distribution which was triggered by a defocusing thermal nonlinearity. The larger light absorption strength in PMMA than PS results in an inhomogeneous temperature distribution. The higher temperature in PMMA translates into a decrease of its refractive index. The result is an increased refractive index mismatch and stronger localization. In effect, they demonstrated that transversely localized modes shrink when the pump intensity is increased despite the fact that $n_2<0$ for the polymers, which seems quite counter-intuitive in the first glance.

In a subsequent publication, the authors provided further evidence of this behavior by analyzing the direct relation between the optical intensity and the localization length. In their experiments, they probed the behavior of light by using both a broadband (a femtosecond Ti: Sapphire laser at 800 nm wavelength with 80 nm bandwidth) and a monochromatic laser (solid-state continuous Nd: YAG laser at 1064 nm). They observed that the broadband light beam injected in the fiber is dispersed at the output into a series of peaks corresponding to localized modes. This output spectrum is modified when the input intensity in the system is increased, demonstrating that the nonlinearity also affects the modes structure. Importantly, they found that the spatial extension of the individual modes decreases with fluence due to the thermal nonlinearity. They repeated the experiment using a monochromatic continuous wave (CW) laser, activating only a single spatial mode, and observed similar self-focusing behavior with the optical power. By comparing their observations with standard homogeneous polymer fibers, they proved that the disorder has turned the defocusing medium into a focusing one.

Further Reading

Abdullaev, S., Abdullaev, F.K., 1980. On propagation of light in fiber bundles with random parameters. Radiofizika 23, 766–767.

Abrahams, E., 2010. 50 years of Anderson Localization. World Scientific.

Anderson, P.W., 1958. Absence of diffusion in certain random lattices. Phys. Rev. 109, 1492–1505. doi:10.1103/PhysRev.109.1492.

De Raedt, H., Lagendijk, A., de Vries, P., 1989. Transverse localization of light. Phys. Rev. Lett. 62, 47–50. doi:10.1103/PhysRevLett.62.47.

Fishman, S., Krivolapov, Y., Soffer, A., 2012. The nonlinear schrödinger equation with a random potential: results and puzzles. Nonlinearity 25, R53.

Ioffe, A.F., Regel, A.R., 1960. Non-crystalline, amorphous, and liquid electronic semiconductors. Prog. Semicond. 4, 237–291.

John, S., 1991. Localization of light. Phys. Today 44, 32–40.

Kopidakis, G., Komineas, S., Flach, S., Aubry, S., 2008. Absence of wave packet diffusion in disordered nonlinear systems. Phys. Rev. Lett. 100, 084103. doi:10.1103/PhysRevLett.100.084103.

Lagendijk, A., Tiggelen, B.V., Wiersma, D.S., 2009. Fifty years of Anderson localization. Phys. Today 62, 24–29.

Lahini, Y., Avidan, A., Pozzi, F., *et al.*, 2008. Anderson localization and nonlinearity in one-dimensional disordered photonic lattices. Phys. Rev. Lett. 100, 013906. doi:10.1103/PhysRevLett.100.013906.

Leonetti, M., Karbasi, S., Mafi, A., Conti, C., 2014a. Experimental observation of disorder induced self-focusing in optical fibers. Appl. Phys. Lett. 105, 171102. doi:10.1063/1.4900781.

Leonetti, M., Karbasi, S., Mafi, A., Conti, C., 2014b. Observation of migrating transverse Anderson localizations of light in nonlocal media. Phys. Rev. Lett. 112, 193902. doi:10.1103/PhysRevLett.112.193902.

Mafi, A., 2015. Transverse Anderson localization of light: A tutorial. Adv. Opt. Photon 7, 459–515. doi:10.1364/AOP.7.000459.

Pertsch, T., Peschel, U., Kobelke, J., *et al.*, 2004. Nonlinearity and disorder in fiber arrays. Phys. Rev. Lett. 93, 053901. doi:10.1103/PhysRevLett.93.053901.

Pikovsky, A.S., Shepelyansky, D.L., 2008. Destruction of anderso localization by a weak nonlinearity. Phys. Rev. Lett. 100, 094101. doi:10.1103/PhysRevLett.100.094101.

Sanchez-Palencia, L., Clement, D., Lugan, P., *et al.*, 2007. Anderson localization of expanding Bose–Einstein condensates in random potentials. Phys. Rev. Lett. 98, 210401. doi:10.1103/PhysRevLett.98.210401.

Schwartz, T., Bartal, G., Fishman, S., Segev, M., 2007. Transport and Anderson localization in disordered two-dimensional photonic lattices. Nature 446, 5255.

Segev, M., Silberberg, Y., Christodoulides, D.N., 2013. Anderson localization of light. Nat. Photon. 7, 197–204.

Second-Harmonic Generation in Two-Dimensional Materials

Myrta Grüning, Queen's University Belfast, Belfast, Northern Ireland, United Kingdom

Main Definitions and Background

Two-dimensional (2D) materials are structures in which atoms are arranged periodically in two of the three spatial directions and which have (sub-)nanometric dimension in the third direction. Furthermore, when more than one layer is present, the chemical bonding between atoms in the periodic directions are much stronger (covalent or ionic) than those between layers (van der Waals). 2D materials are synthesized mostly by mechanical exfoliation of three-dimensional van der Waals layered materials. Some of the more common materials, such as monolayers of graphene, h-BN, Gallium chalcogenides (e.g. GaTe, GaSe) and transition metal dichalcogenides (e.g. MoS_2, WSe_2) have been obtained as well by chemical vapour deposition (CVD) or by van der Waals epitaxy (VDWE) (see e.g. Butler *et al.* (2013)). These processes produce polycrystalline structures with crystal flakes that have one to few layers thickness and with dimensions of 10 μm to 70–80 μm. The flakes which have typically a triangular shape – reflecting the crystal symmetry – are oriented in different directions and separated by grain boundaries. A variety of spectroscopic techniques are in place to characterize the crystal orientation, layers number and stacking, grain boundaries of these polycrystalline structures (see e.g. Butler *et al.* (2013)).

Second-harmonic generation (SHG) is a nonlinear optical process in which a crystal irradiated with laser light of a given frequency ω generates radiation at the second-harmonic frequency 2ω (see Boyd (2008)). In the context of 2D materials, SHG is relevant as it is the basis of spectroscopic techniques to structurally characterize the samples obtained e.g. by exfoliation or CVD. In addition, there is a general interest in optical properties of 2D materials which are expected to be characterized by strong excitonic effects because of the geometrical confinement and poor dielectric screening. Finally technologically there is an interest in the possibility of employing 2D materials with strong SHG as on-chip optical frequency converters.

We recall here some key relations for the SHG which are needed in the discussion that follows. The electric polarization field $\mathbf{P}$ can be expanded in powers of the total electric field $\mathbf{E}$ as (we assume the dipole approximation, if not otherwise stated, see Shen (2003)):

$$\mathbf{P}(\omega) = \varepsilon_0 \left[\chi^{(1)}(\omega)\mathbf{E}(\omega) + \chi^{(2)}(\omega;\omega_1,\omega_2)\mathbf{E}(\omega_1)\mathbf{E}(\omega_2) + \cdots + \chi^{(n)}(\omega;\omega_1,\ldots,\omega_n)\mathbf{E}(\omega_1)\ldots\mathbf{E}(\omega_n) + \cdots \right] \tag{1}$$

where the coefficients $\chi^{(n)}$ are the (non)linear optical susceptibilities and ε_0 is the permittivity of vacuum. The SHG corresponds to the second order term when $\omega_1 = \omega_2$ and thus $\omega = \omega_1 + \omega_2 = 2\omega_1$. Since $\mathbf{P}$ and $\mathbf{E}$ are vector, $\chi^{(2)}$ is a third-rank tensor, $\chi^{(2)}_{ijk}$, with the first index i referring to the direction of the polarization field and the remaining indexes, jk to the direction of the electric field:

$$P_i^{(2)}(2\omega) = \varepsilon_0 \chi^{(2)}_{ijk} E_j(\omega) E_k(\omega) \tag{2}$$

In Eq. (2), $P_i^{(2)}$ is the second order contribution to the polarization field. Symmetry properties of the crystal are reflected in the structure of the second order susceptibility tensor. For example in centrosymmetric crystals, i.e. with an inversion center, all elements of the tensor are identically zero (this implies that the SHG from the electric dipole is identically zero. The measured SHG in centrosymmetric crystals is very small, but not identically zero, because of higher order multipole contributions).

The most commonly studied 2D crystals belongs to the D_{3h} symmetry point-group: the crystal structure is invariant under rotation of 120° around an axis perpendicular to the plane of the 2D crystal (an example is given in **Fig. 1**). The D_{3h} symmetry point-group does not comprise the inversion operation, so crystals belonging to this point-group have a nonzero SHG. The second order susceptibility tensor in this case has only the $yyy = -yxx = -xxy = -xyx$ components different from zero where we labeled with x and y the armchair and the zig-zag directions (see **Fig. 1**). Note that graphene belongs to the D_{6h} symmetry point-group which includes the inversion operation and therefore does not have a dipole-allowed SHG.

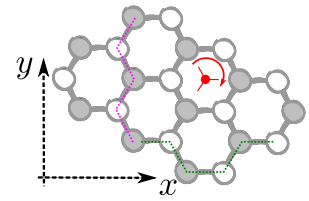


Fig. 1 Symmetry of 2D materials belonging to the D_{3h} symmetry point-group. As an example we consider a hexagonal structure with two atoms of different species. Only few units of the infinite periodic structure are drawn. The red dot represents the 3-fold rotation axis perpendicular to the plane of the 2D crystal. A rotation of 120° around this axis does not change the crystal structure. The three in-plane two-fold rotation axes are also represented as red lines with origin at the dot. The x axis is chosen parallel to one of the armchair directions (dotted green line) and the y parallel to one of the zig-zag directions (dotted magenta line).

 doi:10.1016/B978-0-12-803581-8.09447-9

SHG for Characterization of 2D Materials

SHG experiments on 2D materials are realized by shining laser light onto the sample on a substrate. The second-harmonic power (either reflected or transmitted) is then measured. Most of the SHG experiments on 2D materials are performed in the reflection geometry which is schematically shown in **Fig. 2**. The substrate is typically dielectric coated, e.g. SiO_2/Si. The substrate can be chosen to be transparent to both the fundamental and SH frequency so that the transmitted SH signal can be measured as well. Normally the sample can be both mechanically rotated around the out-of-plane direction and shifted in the two in-plane directions. Rotation allows to study dependence on the light polarization. Scans in the in-plane directions allows to take a SHG image of the sample. Because of the similarity of the problem, many of the concepts and techniques are reminiscent of concepts and techniques applied in SHG measurements of surfaces, films and interfaces.

Polarization Dependent Measurements

To illustrate the basic principle behind polarization dependent measurements we consider here a linearly polarized laser light with propagation direction perpendicular to the 2D crystal and the associated electric field forming an angle θ_0 with the armchair direction of the crystal which we assume to belong to the D_{3h} symmetry-point group. This is schematically depicted in **Fig. 3**. Because of the symmetry with respect the three-fold rotation axis, varying the field polarization of 120° leaves the SHG unchanged. That means that the parallel (along x') and perpendicular (along y') components of the SHG electric field $\mathbf{E}_{\mathrm{SH}}$ vary as $\sin(3\theta+\theta_0)$ and $\cos(3\theta+\theta_0)$ when the sample is rotated by θ. The components of the SH power then, which is proportional to $\mathbf{E}^2_{\mathrm{SH}}$, vary as $\sin^2(3\theta+\theta_0)$ and $\cos^2(3\theta+\theta_0)$. Plotting the reflected SH power as a function of the polarization angle on a polar grid results in a plot similar to that in **Fig. 3** (see e.g. Kumar *et al.* (2013)). A fit of the measured data against the expected behaviour for the SH power gives the SH intensity and the initial direction of the electric field with respect to the crystal armchair direction θ_0.

Then from polarization dependent measurements one can extract two important structural information on the 2D crystal: (1) whether it has a given/expected symmetry, (2) its orientation. Furthermore from the reflected power one can in principle deduce an absolute measure of the SHG, though this is not straightforward as discussed in Section "Absolute Measure of SHG."

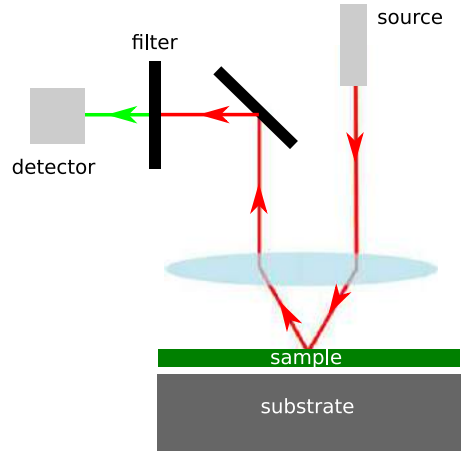


Fig. 2 Schematic set up of a SHG experiment in reflection geometry. The light pumped by the source is focus onto the sample by an objective which also collects the reflected light. Before being detected and analyzed from the detector, the fundamental and frequencies other than the SH frequency are blocked by a (set of) filter(s).

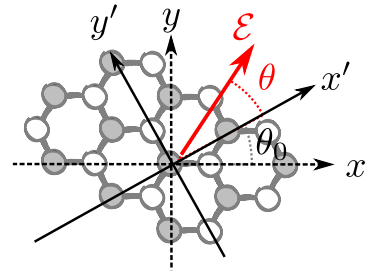


Fig. 3 Direction of electric field with respect to the 2D crystal sample. Initially, the electric field, polarized along the x axis, forms an angle θ_0 with the armchair direction of the crystal (x axis). Later the crystal is rotated, and forms an angle θ with the initial direction.

SHG Imaging

By scanning the SH power in the two in-plane directions of an sample illuminated with polarized (quasi-)monochromatic light, the SH can be rendered in a false-color image, where a scale of colors is associated to the (relative) SH intensity in either the x or y directions. If the substrate is chosen such that it has a negligible SHG, then high contrast images can be obtained. These images provide several information on the sample.

First, as the SH intensity does not change within a crystal, domains of the same color can be identified as crystal flakes. Second, because of the dependence on the electric field polarization discussed previously, crystals with a different orientation show a different SH intensity. Then those crystals can be distinguished from each other and their relative orientation can be estimated. However as it is evident from **Fig. 4**, crystals which form an angle of 60° with each other have the same color. Nevertheless it is still possible to distinguish them because the edges have a different electronic structure with respect to the rest of the crystal and, due to the sensitivity of the SHG to the electronic structure, have a different color (see e.g. Yin *et al.* (2014), Kumar *et al.* (2013)).

When compared with optical imaging, SHG imaging offers several advantages: stronger contrast with respect to the substrate, possibility of distinguish crystals with different orientation and of detecting edges. As discussed in the following section it also allows one to distinguish between different stacking geometries.

Dependence on Stacking and Number of Layers

The SHG is sensitive to the number of layers and to the stacking geometry. The most evident dependence on the number of layers occurs for 2D materials obtained from centrosymmetric 3D crystals (this is the case of several commonly studied crystals such as the transition metal dichalcogenides or the hexagonal-BN). For these 2D crystals a clear pattern is observed in the SHG of even and odd layers samples, with negligible SHG when the number of layers is even (Kim *et al.*, 2013; Li *et al.*, 2013). In fact, samples with an even number of layers have an inversion center and as discussed in Section "Main Definitions and Background" this is reflected in a zero dipole-allowed SHG. Multipole contributions to SHG are in fact several orders of magnitude smaller than the dipole one.

A simple interference model gives an alternative way to understand the even-odd dependence. Let's consider the SH electric field (that is the electric field driven by the SH polarization) of each crystal plane independently. Then for example the SH electric field parallel and perpendicular to the armchair direction in the ith layer are $E^{\parallel}_{2\omega} \propto \cos(3\theta_i)$ and $E^{\parallel}_{2\omega} \propto \sin(3\theta_i)$ where θ_i is the angle between the applied electric field and the armchair direction. If we consider two layers, the total SH electric field is then given by the sum of the fields. The intensity – related to the square of the total electric field – then is given by:

$$I_{1+2} = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos(3(\theta_1 - \theta_2)) \tag{3}$$

For identical layers $I_1 = I_2$, then when $\theta_1 - \theta_2 = 60^\circ$, I_{1+2} vanishes. The *AB* stacking, which gives a centrosymmetric crystal when the number of layers is even, corresponds indeed to an angle of 60° between the armchair directions of two successive layers (see **Fig. 5**). Instead, for crystals with an *AA* stacking $\theta_1 - \theta_2 = 0^\circ$ and I_{1+2} has its maximum value. The *AA* stacking occurs for example in GaSe and GaTe. Moreover it is possible to synthetize layered materials with a stacking different from that of the corresponding bulk layered material (Hsu *et al.*, 2014). Note that the interference model can be refined by considering the SH electric fields as a radiation fields which is equivalent to account for multipole contributions. Then also for an even number of layers in the *AB* stacking one obtains a small, but nonzero SHG.

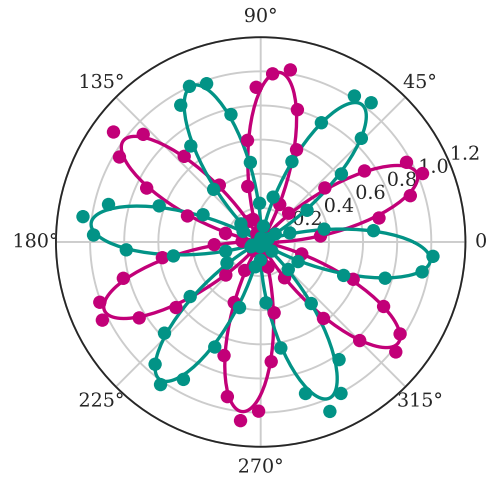


Fig. 4 Polar plot of the SH power versus the direction of the incident electric field, θ in **Fig. 3**. Reflecting the D_{3h} symmetry the parallel (magenta continuous line) and perpendicular (teal continuous line) components of the SH power are proportional to $\sin^2(3\theta + \theta_0)$ and $\cos^2(3\theta + \theta_0)$ respectively. By fitting the experimental data (dots) to the ideal behaviour the initial direction of the field θ_0 and absolute SH power are obtained. This plot is illustrative of typical plots obtained from experiments, but the data is from a real experiment. Dots have been obtained by adding a small random number to the sinusoidal curves.

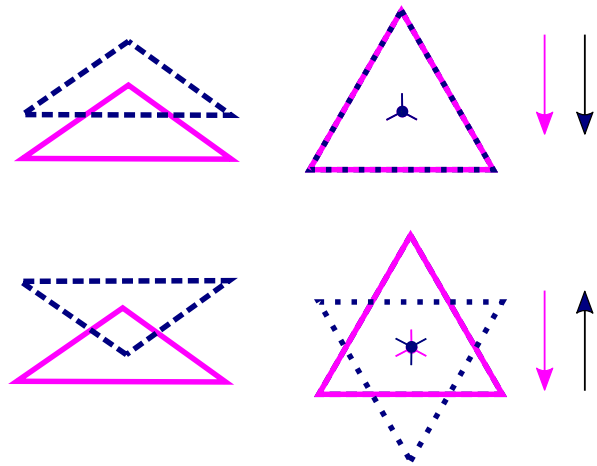


Fig. 5 Two possible stacking arrangements for two layers represented by equilateral triangles (magenta: bottom layer, blue: top layer). Equilateral triangles have been chosen because they are invariant under the same symmetry operations as 2D crystals belonging D_{3h} symmetry point-group. Top: angle between the two armchair directions of the layer is 0°, this is the so called *aa* stacking, side view on the left and top view on the right. When considering each layer independently, the radiated SH electric fields have the same direction, thus they sum up. Bottom: angle between the two armchair directions of the layer is 60°, this is the so called *AB* stacking (or Bernal stacking), side view on the left and top view on the right. When considering each layer independently, the radiated SH electric fields have opposite direction, thus they cancel. This is consistent with the system being centrosymmetric, i.e. invariant under the inversion operation (the inversion center is along the main rotation axis, at the midpoint between the two layers) and thus having no dipole allowed SHG.

Eq. (3) can be used to extract the stacking geometry of bilayers from polarization dependent measures. On the other hand by adding more layers (and knowing the stacking), one can predict the dependence of the SH intensity on the number of layers. As detailed above, from calculations based on the interference model (or modifications) one expects that in case of *AB* stacking samples with an even number of layers have negligible SHG, while samples with an odd number of layers have all similar SHG. Instead in the case of few layers non-centrosymmetric crystals, e.g. with *AA* stacking, the SH intensity is expected to grow quadratically. However the interference model– even when perfectioned by accounting for absorption and multiple reflections– captures the dependence on the number of layers only partially as it neglects the effects of interlayer interactions on the electronic structure. In particular for very few layers (typically <10), confinement effects are stronger and the screening less effective. These effects can change both the absorption and SHG spectral shape and onset and ultimately can be captured only by electronic structure calculations. Indeed in experimental studies is generally observed that a quadratic dependence hold for 10–100 layers, but for <10 layers the behaviour changes (see e.g. Tang *et al.* (2016)).

Absolute Measure of SHG

A measure of the absolute SHG intensity is not straightforward. From Eq. (1) the SH susceptibility is extracted from the change in the polarization field **P** of the sample as a function of the total electric field **E**.

In practice what is measured is the SH radiation intensity (or power) and what is known the intensity (or power) of the applied field. The relation between the two quantities is derived from the wave equation for the electric field driven by the second order polarization field and using Eq. (2) to introduce the SH susceptibility and the total electric field. Then the SH radiation intensity is calculated from the corresponding field as (Boyd (2008)). Because of the quadratic dependence of the polarization on the electric field (Eq. (2)), the SH radiation intensity (power) depends quadratically on the intensity (power) of the applied field. The coefficient of proportionality contains the square of the SH susceptibility and other factors depending on the sample i.e. the thickness, the presence of a substrate, etc. The model assumed to describe the sample influences the experimental estimate for the SH susceptibility as discussed in Section "Models for the Sample."

Furthermore, as it is the case for the SHG in bulk materials, the intensity of the applied field requires the detailed knowledge of the spatial and temporal dependence of the laser beam. Commonly the dependence on the intensity of the applied field is thus eliminated by measuring the SH power of the sample relatively to a reference material for which the SH susceptibility is accurately known (as e.g. quartz or beta barium borate).

Table 1 collects results from recent experimental measures of the SHG in 2D materials and reports parameters and characteristics that may influence the estimate, which are discuss below (Sections "Models for the Sample, Role of Substrate and Role of Excitonic Resonances"). The out-of-resonance values are usually comparable with those of nonlinear crystals, such as the beta barium borate and potassium titanyl phosphate, commonly used as frequency converters. At resonance, strong to very strong (one to three order of magnitude larger than conventional nonlinear crystal) is reported. In the subsection below aspects that may influence significantly the SHG experimental estimate are detailed.

Table 1 Measure of the SH for 2D materials [monolayer (ML), bilayer (BL), trilayer (TL) and few layers (FL)] from recent experiments. For each material beside the SH is reported how the material was synthesized, the substrate (thickness in parenthesis), excitation wavelength, whether excitonic resonances are excited, which model has been assumed to extract the SH from the measurements and the publication. Considering the number of parameters on which the estimates depend, those should be consider as order of magnitude estimates

	SH (pm/V)	*Synthesis*	*Substrate*	*Waveleght (nm)*	*Resonance*	*Model*	
MoS_2 ML	6	CVD	Si/SiO_2 (300 nm)	static limit	No	sheet	PRB 92, 159901 (2015)
	40	CVD	Si/SiO_2 (300 nm)	1100–2000	Yes (A,B excitons)	sheet	PRB 92, 159901 (2015)
	25–30	CVD	PET/fused silica	1360	Yes (A exciton)	sheet	APL 107, 13113 (2015)
	35–40	CVD	PET/fused silica	1240	Yes (B exciton)	sheet	APL 107, 13113 (2015)
	30–100	Exfoliated	amorphous quartz	680–1080	Yes (C exciton)	sheet	PRB 87, 201401 (2013)
	1×10^5	Exfoliated	Si/SiO_2 (90 nm)	810	Yes (C exciton)	bulk	PRB 87, 161403 (2013)
	5×10^3	CVD	Si/SiO_2 (90 nm)	810	Yes (C exciton)	bulk	PRB 87, 161403 (2013)
	150	CVD	sapphire	810	Yes (C exciton)	sheet	ACS Nano 8, 2951 (2014)
	160	Exfoliated	fused silica	810	Yes (C exciton)	sheet	Nano Lett 13, 3329 (2013)
	430	not rep.	not rep.	1600	No	bulk	JACS 137, 7994
MoS_2 TL	10–50	Exfoliated	amorphous quartz	680–1080	Yes (C exciton)	sheet	PRB 87, 201401 (2013)
$MoSe_2$ ML	50	CVD	Si/SiO_2	1200–1800	Yes (A,B excitons)	sheet	Ann. Phys. 528, 551 (2016)
WS_2 ML	9×10^3	CVD	suspended, Si/SiO_2	832	Yes (C exciton)	sheet	SciRep 4, 5530 (2014)
WSe_2 ML	10×10^3	Exfoliation	Si/SiO_2 (300 nm)	816	Yes (C exciton)	sheet	2D Mat 4, 045015 (2015)
GaSe ML	2.4×10^3	VDWE	fused silica	1210	No	bulk	JACS 137, 7994
	1.7×10^3	VDWE	fused silica	1350	No	bulk	JACS 137, 7994
	700	VDWE	fused silica	1600	No	bulk	JACS 137, 7994
GaSe BL	21–34	Exfoliated	Si/SiO_2 (90 nm)	1030–1460	No	sheet	PRB 94, 125302 (2016)
	60	Exfoliated	glass	900–1080	No	bulk	Ang. Chem. 127, 1201 (2015)
GaSe TL	5–9	Exfoliated	Si/SiO_2 (90 nm)	1030–1460	No	sheet	PRB 94, 125302 (2016)
	93	Exfoliated	glass	900–1080	No	bulk	Ang. Chem. 19, 1201 (2015)
GaTe FL	1×15	Exfoliated	Si/SiO_2	1560	No	?	APL 108, 073103 (2016)
h-BN ML	10	Exfoliated	fused silica	810	No	sheet	Nano Lett 13, 3329 (2013)

Models for the Sample

For 2D materials two models are commonly used (see e.g. Clark *et al.* (2014)):

- One model assumes that the 2D material sample behaves as bulk and uses the same formula for the SH radiation electric field with the only difference that the phase mismatch is neglected since the thickness of the material is much smaller than the coherence length. With this assumption the relation between the intensity I^{SH} of the SH radiation and the intensity I of the applied electric field at frequency ω to extract the magnitude of the SH susceptibility $|\chi^{(2)}|$ reads:

$$I^{SH}(2\omega) = \frac{\omega^2 |\chi^{(2)}|^2 \Delta h^2 I^2(\omega)}{2 n_{2D}^2(\omega) n_{2D}(2\omega) \varepsilon_0 c^3} \tag{4}$$

  where Δh is the sample thickness, n_{2D} is the refractive index of the 2D material and c is the speed of light.
- Another model instead considers the 2D material on a dielectric substrate with refractive index n_{sub} as a dipole sheet screened by the substrate and thus considers the electric field driven by the SH sheet polarization. With this assumption the relation between the applied and measured intensities reads:

$$I^{SH}(2\omega) = \frac{512 \omega^2 \pi^2 |\chi^{(2)}_{surf}|^2 I^2(\omega)}{(n_{sub}(\omega) + 1)^6 \varepsilon_0 c^3} \tag{5}$$

  Here $|\chi^{(2)}_{surf}|$ is the surface second order nonlinear susceptibility (Shen, 2003) which is relate to the usual (volume) second order nonlinear susceptibility by:

$$\chi^{(2)}_{surf} = \int dz \chi^{(2)}$$

  where the integral runs across the 2D material along the normal to the surface. If we neglect the dependence on z of the nonlinear susceptibility, $\chi^{(2)}_{surf} = \chi^{(2)} \Delta h$.

One important difference between the two models is that at the denominator of the sheet model (Eq. (5)) the refractive index of the dielectric substrate appears instead of that of the material (Eq. (4)). As the latter is typically much smaller than the former it has been argued that a difference up to three order of magnitude in the estimate of the SH susceptibility may arise.

Role of Substrate

Characteristics of the substrate plays a role in the measured intensity. Typically a dielectric coated substrate is chosen for which the bulk SHG is negligible. However even if the bulk SHG of the substrate is negligible, surface SHG can still be noticeable. In addition atoms may be trapped at the interface between the sample and the substrate. Those defects can modify the SHG of the sample. In particular charge-transfer defects can induce strong local fields and locally enhance the SHG signal.

Furthermore it as been reported that the measured SHG changes when changing the thickness of dielectric coating. This effect has been explained as an enhancement of the measured SHG due to the interference of multiple reflections from the substrate (Miyauchi *et al.*, 2016), similar to the enhancement observed in a Fabry-Perot resonator discussed in Section "Optical Devices."

Finally, while for measurements of the SHG of the sample, substrate effects are a disturbance, in applications one may want to actually enhance the SHG from the sample by choosing a appropriate substrate. In this context few experiments have explored effects of metal substrates as it is expected that strong local fields from surface plasmons could significantly enhance the SH intensity (Zeng *et al.*, 2015).

Role of Excitonic Resonances

Excitations are commonly interpreted in terms of electron-hole pairs. Pictorially, considering the electronic structure of a finite-gap material, an electron-hole pair is formed when an electron is excited from the valence to the conduction energy band. Within this picture the excited system then has an extra electron in the conduction, and a missing electron (or hole) in the valence. The electron and the hole can be considered as particles with opposite charge interacting through the Coulomb interaction and forming a bound state, the exciton. The physics of these bound states resembles that of the hydrogen atom in which the radius and the energies are modified by the electric screening from the surrounding electrons. The largest the screening, the largest the radius of the exciton (i.e. the more delocalized) and the smallest the binding energy. In the case of 2D materials one needs to account for the geometrical confinement that modifies the equations for the radius and binding energy: the binding energy of the ground state exciton is 4 times larger than that of the bulk counterpart and the radius smaller by a factor 2 (Haug and Koch, 1990). Then typically excitons in 2D materials are expected to be more localized and more bound than in their bulk counterpart. Because of the resemblance with the hydrogen atoms excitons are commonly addressed as 1*s*, 2*p*, etc. even though it should be noted that the symmetry properties of these excitations are different from those of the hydrogenic atomic orbitals and depends on the crystal symmetry.

When the laser frequency ω is such that 2ω matches the energy of one of these excitations the SHG is expected to be enhanced. Indeed experimentally enhancements up to three order of magnitude have been observed for transition metal dichalcogenides (Wang *et al.*, 2015) resulting in very strong SHG at excitonic resonances as can be seen in **Table 1**.

Electronic Structure Calculations

The SHG can be computed from the electronic structure of the material as (Shen, 2003):

$$\chi^{(2)}_{ijk}(\omega;\omega_1,\omega_2) = \int \mathrm{d}\mathbf{k} \sum_{v,c,c'} f_v(\mathbf{k}) \left\{ \frac{\langle v\mathbf{k}|\xi_i|c\mathbf{k}\rangle\langle c\mathbf{k}|\xi_j|c'\mathbf{k}\rangle\langle c'\mathbf{k}|\xi_k|v\mathbf{k}\rangle}{[\omega-\omega_{cv}(\mathbf{k})][\omega_1-\omega_{c'v}(\mathbf{k})]} + \frac{\langle v\mathbf{k}|\xi_i|c\mathbf{k}\rangle\langle c\mathbf{k}|\xi_k|c'\mathbf{k}\rangle\langle c'\mathbf{k}|\xi_j|v\mathbf{k}\rangle}{[\omega-\omega_{cv}(\mathbf{k})][\omega_2-\omega_{c'v}(\mathbf{k})]} + \frac{\langle v\mathbf{k}|\xi_k|c\mathbf{k}\rangle\langle c\mathbf{k}|\xi_j|c'\mathbf{k}\rangle\langle c'\mathbf{k}|\xi_i|v\mathbf{k}\rangle}{[\omega+\omega_{c'v}(\mathbf{k})][\omega_2+\omega_{cv}(\mathbf{k})]} \right.$$
$$\left. + \frac{\langle v\mathbf{k}|\xi_j|c\mathbf{k}\rangle\langle c\mathbf{k}|\xi_k|c'\mathbf{k}\rangle\langle c'\mathbf{k}|\xi_i|v\mathbf{k}\rangle}{[\omega+\omega_{c'v}(\mathbf{k})][\omega_1+\omega_{cv}(\mathbf{k})]} + \frac{\langle v\mathbf{k}|\xi_j|c\mathbf{k}\rangle\langle c\mathbf{k}|\xi_i|c'\mathbf{k}\rangle\langle c'\mathbf{k}|\xi_k|v\mathbf{k}\rangle}{[\omega_1-\omega_{cv}(\mathbf{k})][\omega_2+\omega_{c'v}(\mathbf{k})]} + \frac{\langle v\mathbf{k}|\xi_k|c\mathbf{k}\rangle\langle c\mathbf{k}|\xi_i|c'\mathbf{k}\rangle\langle c'\mathbf{k}|\xi_j|v\mathbf{k}\rangle}{[\omega_1+\omega_{c'v}(\mathbf{k})][\omega_2-\omega_{cv}(\mathbf{k})]} \right\} \quad (6)$$

The main quantities in Eq. (6) are the single particles wavefunctions $|a\mathbf{k}\rangle$ where $a=v$ corresponds to valence and $a=c$ to conduction states, and $\mathbf{k}$ is the crystal momentum. ω_{cv} is the difference between the energies of the conduction and valence states. ξ_l is the dipole operator in the l direction, defined consistently with periodic boundary conditions.

Single particle wavefunctions and energies can be obtained from any electronic structure approach. In practice most of nowadays calculations are based on single-particle energies and wavefunctions from a density-functional theory calculations. The main advantages of this approach is computational ease, availability of computer codes while not relying on ad-hoc or experimental parameters. On the other hand, regarding the latter point it has to be noted that due to the underestimation of band gap inherent to this approach, often in calculations the knowledge of the material experimental band-gap is used (as well the band gap obtained from other approaches, such as the GW approximation can be used).

The role of electronic structure calculations of SHG is mainly (1) to connect features in the SHG spectrum to the electronic structure of the material and (2) to provide estimates for the absolute SHG. In the context of 2D materials connecting spectral features to the electronic structure can be important for instance to understand and interpret changes of the SHG with the number

of layers. On the other hand, an estimate of SHG is relevant considering the experimental difficulties in measuring the SHG. Comparison with the theory can provide a more firmly ground for the assumptions made in extracting the experimental SHG.

Eq. (6) is derived considering non-interacting electrons, in what is commonly addressed as independent particle model. More precisely, electron correlation is included in the ground-state calculations from which the electronic structure is extracted, but neglected in the response of the system to external perturbation.

Descriptions beyond the independent particle model include local fields and excitonic effects. The former effect is due to the microscopic electric fields from crystal inhomogeneity which locally counteract the applied electric field. The latter effect is due to the electron–hole long range interaction as discussed in Section "Role of Excitonic Resonances." Modifications of Eq. (6) so to introduce these effects are nontrivial.

At present very few calculations of the SHG of 2D material exists including local fields and excitonic effects (see e.g. Grüning and Attaccalite (2014), Trolle *et al.* (2014)) . For transition metal dichalcogenides available calculations confirm the enhancement of the SHG at excitonic resonances and the strong SHG of these materials. As a side result of these works, calculations at the independent particle level are typically underestimating the SHG at excitonic resonances by a factor 2–3. Considering the experimental uncertainties in the determination of the SHG, calculations at the independent particle level can then be considered in general sufficient to provide a reasonable estimate of the SHG.

Optical Devices

As seen in Section "Absolute Measure of SHG" transition metal dichalcogenides have a significant SHG in the frequency range of Ti: Sapphire laser and Gallium chalcogenides have a significant SHG for wavelengths near 1560 nm relevant for telecommunications. This strong SHG can potentially be exploited for on-chip optical devices (e.g. frequency converter). One problem in the realization of such devices is that in spite of the intrinsically strong SHG the length scale available for light-matter interaction is subnanometric, so that the resulting signal is weak. One proposed solution is to embed the 2D material in a resonant optical micro Fabry-Perot cavity. Prototypes of devices built on this principle showed enhancement of the SHG from one to three order of magnitude (Day *et al.*, 2016; Yi *et al.*, 2016).

Furthermore, a prototype of optical data storage has been proposed (Zeng *et al.*, 2015) which exploits the dependence of the SHG intensity on the number of layers in MoS_2. In particular for a coated metallic substrate the SHG is found to increase with number of layers up to a certain thickness (17 nm for gold) and then to decrease again. Then, by choosing the number of layers exhibiting the strongest SHG, information can be stored by locally reducing the number of layers and can be read out by detecting the SHG intensity.

References

Boyd, R.W., 2008. Nonlinear Optics. Academic Press.

Butler, S.Z., Hollen, S.M., Cao, L., *et al.*, 2013. Progress, challenges, and opportunities in two-dimensional materials beyond graphene. ACS Nano 7 (4), 2898–2926.

Clark, D.J., Senthilkumar, V., Le, C.T., *et al.*, 2014. Strong optical nonlinearity of cvd-grown MoS_2 monolayer as probed by wavelength-dependent second-harmonic generation. Phys. Rev. B 90, 121409.

Day, J.K., Chung, M.-H., Lee, Y.-H., Menon, V.M., 2016. Microcavity enhanced second harmonic generation in 2d MoS_2. Opt. Mater. Express 6 (7), 2360–2365.

Grüning, M., Attaccalite, C., 2014. Second harmonic generation in *h*-bn and MoS_2 monolayers: role of electron-hole interaction. Phys. Rev. B 89, 081102.

Haug, H., Koch, S.W., 1990. Quantum Theory of the Optical and Electronic Properties of Semiconductors. World Scientific.

Hsu, W.-T., Zhao, Z.-A., Li, L.-J., *et al.*, 2014. Second harmonic generation from artificially stacked transition metal dichalcogenide twisted bilayers. ACS Nano 8 (3), 2951–2958.

Kim, C.-J., Brown, L., Graham, M.W., *et al.*, 2013. Stacking order dependent second harmonic generation and topological defects in *h*-bn bilayers. Nano Letters 13 (11), 5660–5665.

Kumar, N., Najmaei, S., Cui, Q., *et al.*, 2013. Second harmonic microscopy of monolayer MoS_2. Phys. Rev. B 87, 161403.

Li, Y., Rao, Y., Mak, K.F., *et al.*, 2013. Probing symmetry properties of few-layer MoS_2 and *h*-bn by optical second-harmonic generation. Nano Letters 13 (7), 3329–3333.

Miyauchi, Y., Morishita, R., Tanaka, M., *et al.*, 2016. Influence of the oxide thickness of a SiO_2/Si(001) substrate on the optical second harmonic intensity of few-layer $MoSe_2$. Japan. J. Appl. Phys. 55 (8), 085801.

Shen, Y.R., 2003. The Principles of Nonlinear Optics. Wiley Inter-science.

Tang, Y., Mandal, K.C., McGuire, J.A., Lai, C.W., 2016. Layer-and frequency-dependent second harmonic generation in reflection from gase atomic crystals. Phys. Rev. B 94, 125302.

Trolle, M.L., Seifert, G., Pedersen, T.G., 2014. Theory of excitonic second-harmonic generation in monolayer MoS_2. Phys. Rev. B 89, 235410.

Wang, G., Marie, X., Gerber, I., *et al.*, 2015. Giant enhancement of the optical second-harmonic emission of wse_2 monolayers by laser excitation at exciton resonances. Phys. Rev. Lett. 114, 097403.

Yi, F., Ren, M., Reed, J.C., *et al.*, 2016. Optomechanical enhancement of doubly resonant 2d optical nonlinearity. Nano Letters 16 (3), 1631–1636.

Yin, X., Ye, Z., Chenet, D.A., *et al.*, 2014. Edge nonlinear optics on a MoS atomic monolayer. Science (New York, NY) 344 (6183), 488–490.

Zeng, J., Yuan, M., Yuan, W., *et al.*, 2015. Enhanced second harmonic generation of MoS_2 layers on a thin gold film. Nanoscale 7, 13547–13553.

Parity-Time Symmetry in Optics

Mercedeh Khajavikhan, University of Central Florida, Orlando, FL, United States
Mohammad-Ali Miri and Andrea Alu, University of Texas at Austin, Austin, TX, United States
Demetrios N Christodoulides, University of Central Florida, Orlando, FL, United States

Introduction

Refractive index, gain and loss play a pivotal role in light propagation. Modulating the refractive index as a means to mold the flow of light dates back to antiquity when rock crystals like the "Loupe of Sargon" were used to produce fire by focusing sun rays. Since then, index crafting has reached such level of sophistication that artificial structures as diverse as optical fibers, photonic crystals, and meta-materials became possible. Optical amplification or gain was realized much later, followed by the discovery of the laser, and has enabled numerous applications in many areas of science and technology. However, this is not the case for attenuation. Loss, unlike the other two processes, is still perceived to be a foe, an undesirable attribute that should be avoided or compensated at all costs. It is perhaps for this particular reason that up until recently the simultaneous use of index, gain, and loss as a viable route to achieve new optical behavior has been generally overlooked.

At a first glance, deliberately intermixing gain with loss may appear counterintuitive and perhaps pointless. Recently, however, there has been a growing evidence within the field of non-Hermitian optics, indicating that the presence of *exceptional points* (phase transition points) arising in judiciously designed parity-time symmetric systems can be utilized to build structures with customized properties and functionalities- previously thought to be unattainable. In what follows, we first provide a mathematical description of parity-time symmetry both in the context of quantum mechanics and optics, and then continue by overviewing some of the current developments concerning the use of PT-symmetry for laser mode management and sensing, while also looking into their anomalous scattering properties for designing invisibility cloaks.

Spontaneous Symmetry Breaking and Exceptional Points

The field of parity-time (PT) symmetry began in 1998 following the discovery by Bender and Boettcher, who first indicated that a wide class of non-Hermitian Hamiltonians can exhibit entirely real spectra, provided that they commute with the parity-time ($\hat{P}\hat{T}$) operator (Bender and Boettcher, 1998). While the use of non-Hermitian operators has been suggested in the past as a means to incorporate loss in quantum mechanics, Bender and Boettcher's work has been largely perceived as counterintuitive, since it went against the commonly held view that real eigenvalues are only associated with Hermitian observables. Starting from the aforementioned premise, one can directly show that a necessary (but not sufficient) condition for PT symmetry to hold is that the complex potential involved in such a Hamiltonian should satisfy $V(x)=V^*(-x)$. In other words, the real part of the complex potential must be an even function of position, while its imaginary component should be anti-symmetric. In such pseudo-Hermitian configurations, the eigenfunctions are no longer orthogonal, i.e., $\langle m|n\rangle \neq \delta_{mn}$, and hence the vector space is skewed. Even more intriguing is the possibility for a sharp symmetry-breaking transition – once a non-Hermiticity parameter exceeds a certain critical value. In this latter regime, the Hamiltonian and the $\hat{P}\hat{T}$ operator no longer display the same set of eigenfunctions (even though they commute) and as a result, the eigenvalues of the system become partially, or entirely, complex. This broken PT-symmetry phase is associated with a so-called exceptional point (EP) or a non-Hermitian degeneracy.

While the implications of these mathematical theories in quantum mechanics are still a matter of debate (since the realization of complex potentials in quantum mechanics is inherently out-of-reach), later it has been recognized that photonics can provide a fertile ground where PT symmetry concepts can be experimentally investigated (Makris *et al.*, 2008). The transition from quantum mechanics to optics can be formally justified by considering the isomorphism between the Schrödinger equation and the paraxial wave equation (see **Table 1**). In this regard, the complex refractive index profile plays the role of an optical potential: $k_0(n_R(x)+in_I(x))$. In optical systems, PT symmetry demands that the spatial distribution of the refractive index is an even function ($n_R(x)=n_R(-x)$), whereas the imaginary component (representing gain or loss) is an odd function of position ($n_I(x)=-n_I(-x)$). One can also show that in high contrast settings, where the electrodynamic problem must be treated fully vectorially, the condition for PT symmetry is expressed through the complex permittivity, e.g., $\varepsilon(\mathbf{r})=\varepsilon^*(-\mathbf{r})$. Nonetheless, since in most systems the imaginary part of the complex refractive index function is considerably smaller than its real counterpart, the PT symmetry condition in these two scenarios is almost equivalent.

To elucidate some the physics of exceptional points, here we provide a simple example of a two-level non-Hermitian PT symmetric system. In optics, this may be realized by two coupled resonators or waveguides, one exhibiting gain, while the other one experiencing an equal amount of loss. In the case of two cavities, energy exchange occurs in time, while for waveguides it takes place in space. Schematics of these configurations are depicted in **Fig. 1**.

Here we consider the case of two identical evanescently coupled cavities as depicted in **Fig. 1(b)**. In order to simplify our analysis, we assume both resonators are single moded. For this system, the field evolution dynamics are described by the following

Table 1 Isomorphic relations between the Schrödinger equation in quantum mechanics and the paraxial wave equation in optics

	Quantum mechanics	*Optics*
Governing Equation	Schrödinger equation $i\hbar\frac{\partial\psi}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2} + V(x)\psi$	Paraxial wave equation $-i\frac{\partial E}{\partial z} = \frac{1}{2k}\frac{\partial^2 E}{\partial x^2} + k_0 n(x)E$
Wave function	$\psi(x,t)$	$E(x,z)$
Eigenvalues	Energy states	Propagation constants
Potential energy	Potential $V(x) = V_R(x) + iV_I(x)$	Refractive index $K_0 n(x) = K_0 n_R(x) + iK_0 n_I(x)$
PT-symmetry necessary condition	$V_R(x) = V_R(-x)$ $V_I(x) = -V_I(-x)$	$n_R(x) = n_R(-x)$ $n_I(x) = -n_I(-x)$

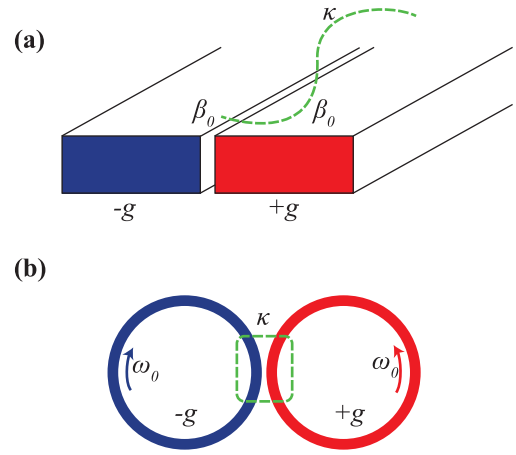


Fig. 1 Two realizations of two-level parity-time symmetric optical systems, (a) a pair of coupled waveguides, one subjected to gain and the other to loss, trade energy along the direction of propagation at a rate of κ, (b) a pair of cavities, one experiencing gain and the other loss, exchange energy in time at a rate of κ.

set of equations:

$$\begin{cases} i\dfrac{da}{dt} - i\dfrac{g}{2}a + \kappa b = 0 \\ i\dfrac{db}{dt} + i\dfrac{g}{2}b + \kappa a = 0 \end{cases} \tag{1}$$

Here a and b represent the modal field amplitudes in the gain and loss cavities, respectively, g stands for the gain/loss coefficient, and κ is the coupling strength. This system supports two supermodes with distinct characteristics that depend entirely on the ratio between gain/loss (g) and coupling (κ). When $g/2\kappa<1$, the two eigenvalues are real and are given by $\lambda_{1,2}=\pm\cos(\theta)$ where $\theta=\sin^{-1}(g/2\kappa)$ and the corresponding eigenvectors are $|1\rangle=[1e^{i\theta}]^T$ and $|2\rangle=[1-e^{-i\theta}]^T$. Note that these eigenvectors are not orthogonal in spite of the fact that the spectrum is real. In this case, none of the two modes experiences a net gain or loss, instead they oscillate, i.e., they remain neutral.

The spectral response of this PT-symmetric two-level system drastically changes as soon the gain-loss contrast exceeds the coupling strength ($g/2\kappa>1$). In this regime, the eigenvalues are expressed by $\lambda_{1,2}=\pm\sinh(\theta)$ where $\theta=\cosh^{-1}(g/2\kappa)$, and their corresponding non-orthogonal eigenvectors turn out to be $|1\rangle=[1\ ie^{\theta}]^T$ and $|2\rangle=[1\ ie^{-\theta}]^T$, indicating that the symmetry of each mode is broken in the sense that one of them resides mostly in the gain cavity while the other one in the lossy resonator. As a result, one of the eigenmodes enjoys amplification while the other experiences attenuation. Obviously, the supermode that mostly occupies the gain/loss region is amplified/attenuated. **Fig. 2** displays the complex component of the eigenvalues as a function of $g/2\kappa$. The transition point $g/2\kappa=1$ between these two different regimes is called the *exceptional point*. At this crossing, something quite remarkable happens. The dimensionality of the vector space is abruptly reduced from 2 to 1. This is evident from the fact that the two eigenvectors become entirely identical $[1\ i]^T$ and neither oscillate nor exponentially varying. Instead the modal amplitudes depend algebraically on time ($\sim C_0 + C_1 t$).

Initial theoretical and experimental work carried out by several groups indicates that light propagating through non-Hermitian or PT-synthetic media can exhibit surprisingly unusual behavior. Among the numerous properties that such systems can display are: absorption enhanced transmittivity (Guo *et al.*, 2009; Rüter *et al.*, 2010), unidirectional invisibility (Lin *et al.*, 2011),

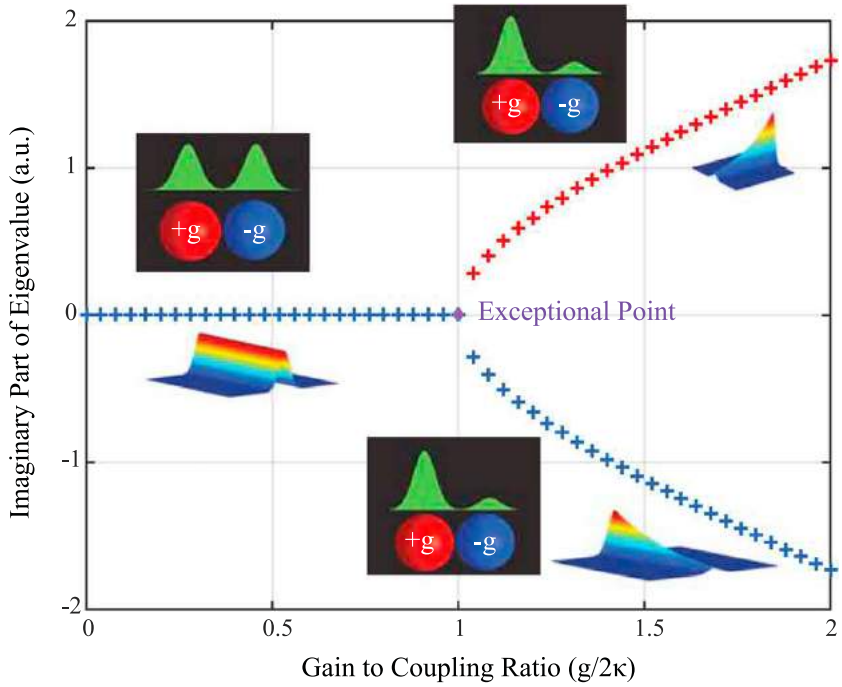


Fig. 2 The imaginary part of the eigenvalue vs. the change in the imaginary part of refractive index (gain-loss contrast) is depicted for a PT two-level system. At the exceptional point the symmetry of the modes breaks and two supermodes emerge, one amplifying and the other one attenuating.

unidirectional light transport (Peng *et al.*, 2014), topological chirality (Doppler *et al.*, 2016; Xu *et al.*, 2016), single mode lasing (Hodaei *et al.*, 2014; Feng *et al.*, 2014), and enhanced sensitivity at the exceptional point (Wiersig, 2014). In most cases, one may design the system close to the exceptional point such that one desired attribute can be extracted whereas other undesired properties remain below the threshold breaking. In what follows we provide an overview of the current developments in this field. In particular, we explain how the selective breaking of parity-time symmetry in multimode laser cavities can result in single mode operation. We will then look into the sensitivity of a system that is biased at an exceptional point and the anomalous scattering properties of PT-symmetric arrangements.

Mode Management in PT-Symmetric Lasers

Since the invention of diode lasers in 1962, there has been an ongoing effort to design optical cavities that funnel the maximum output power into a single spatiotemporal mode. Ultimately, a cavity becomes intrinsically single-moded only when its dimension is comparable to the wavelength of light. Most of the time, this criterion severely limits the amount of coherent emission that can be generated by a laser. On the other hand, when the active region is enlarged beyond this limit, most semiconductor lasers, because of their inhomogeneous lineshape, become susceptible to multimode operation something that eventually leads to output power fluctuations, instabilities, and lower beam quality. This explains why developing a compact, single mode laser has been the subject of numerous research activities over the past few decades. Such efforts have since resulted in several ingenious laser designs. For example, single transverse mode operation can be achieved using broad area slab coupled waveguide lasers (SCOWL). On the hand, distributed feedback (DFB), distributed Bragg reflector (DBR), and vertical cavity surface emitting lasers (VCSEL) have been designed to operate in a single longitudinal mode.

In recent years, the stringent requirements in photonic integrated circuits, have set new challenges for on-chip light sources. Not only such lasers must occupy a small size and exhibit a low threshold, they should also couple easily to the rest of the photonic network, preferably through a bus waveguide. In addition, cavity designs that rely on multi-step growth processes must be avoided. One family of lasers that can address these challenges are micro-ring lasers. Microring resonators have small footprint and as lasers they expected to show low thresholds because of their high quality factors. However, even a microring with a radius on the order of few tens of micrometer, tends to support several longitudinal modes across the gain bandwidth of most III-V active gain materials. Up until recently, the only systematic way to enforce single mode operation in such arrangements was through the incorporation of dispersive vertical gratings. Such a technique, however, seriously deteriorates the quality factor of the ring; hence increasing the threshold and reducing the slope efficiency. In what follows we will describe how selective PT symmetry breaking can be used to enforce single mode operation in microring lasers. This technique was first proposed by Miri *et al.* (2012) in the context of broad area lasers and amplifiers. In 2014, the use of PT symmetry for controlling the longitudinal mode contents of microring resonators

has been proposed by two independent groups. In one approach, the ring is patterned by a parity-time-symmetric grating structure (Feng *et al.*, 2014). In the other work, the active ring is paired with a lossy cavity to establish parity-time symmetry (Hodaei *et al.*, 2014). In what follows we will focus on the latter technique because of its simplicity in terms of avoiding additional gratings.

Single Longitudinal Mode Lasing

PT symmetry breaking can be elegantly exploited to establish single longitudinal mode operation in inherently multi-moded microring lasers. One way to do so, is by utilizing a coupled arrangement of two structurally identical ring resonators; one experiencing gain, and the other one an equal amount of loss. In general, when two fully identical resonators (both subjected to gain, loss, or neither) are placed next to one another, the degeneracy between their respective modes will be broken. The frequency splitting $\Delta\omega$ of the resulting supermode doublets $\omega_n^{(1,2)}$ is directly proportional to the coupling coefficient, i.e., $\Delta\omega = 2\kappa$ (**Fig. 3(b)**). On the other hand, in a PT-symmetric arrangement (like that depicted in **Fig. 3(c)**), the relationship between eigenfrequencies and coupling is by nature different, as indicated in the following expression:

$$\omega_n^{(1,2)} = \omega_n \pm \sqrt{\kappa_n^2 - g_n^2} \tag{2}$$

Consequently, any pair of modes, whose gain/loss contrast (g_n) remains below the coupling coefficient (κ_n), undergoes bounded neutral oscillations. However, as soon as g_n exceeds κ_n, a conjugate pair of lasing/decaying modes emerges. Consequently, the judicious placement of this PT threshold will allow a complete suppression of all non-broken mode pairs in favor of a single amplified mode associated with the aforementioned conjugate pair (**Fig. 3(c)**). As the imaginary parts of the eigenvalues diverge, degeneracy between their real parts is restored.

Even in the absence of PT-symmetry, any resonator with a spectrally non-uniform gain distribution $g(\omega)$ can in principle exhibit single-mode operation, provided that the loss exceeds the gain for all but one resonance. However, in this regime, the amplification cannot surpass the gain contrast $g_{\max} = g_0 - g_1$, where g_0 refers to the gain of the principal mode and g_1 to that of the next-strongest competing resonance. Obviously, this approach will impose severe constraints on the operating parameters, especially in the case of cavities having wide gain linewidths and/or closely spaced resonator modes, where $g_{\max}$ is very small. On the other hand, in a PT-symmetric setting, the coupling κ now plays the role of a virtual loss, and all undesirable modes must fall below its corresponding threshold. According to Eq. (2) we find that in this case the maximum achievable differential gain is:

$$g_{\max,\mathrm{PT}} = \sqrt{g_0^2 - g_1^2} = g_{\max} \cdot \sqrt{\frac{g_0/g_1 + 1}{g_0/g_1 - 1}} \tag{3}$$

Given that $g_0 \geq g_1$, a selective breaking of PT symmetry can therefore systematically increase the available amplification for single-mode operation. As a matter of fact, the square-root behavior of this enhancement (a direct outcome of PT symmetry breaking), as characterized by the enhancement factor $G = g_{\max,\mathrm{PT}}/g_{\max}$, is capable of providing substantially higher selectivity, especially when the initial contrast between adjacent modes is small ($g_1 \to g_0$). Consequently, this approach naturally exhibits a type of broadband self-adaptive single mode selection that is, in principle, applicable to any active resonator configuration. It should be stressed that this type of mode discrimination is resilient and remains valid even when PT symmetry does not exactly hold for all modes involved. On a fundamental level, the mechanism is related to the presence of an exceptional point in the system.

To experimentally verify these findings, active ring resonators based on InGaAsP quantum wells were defined using electron beam lithography and reactive ion etching (RIE). The gain and loss regions were determined by selective optical pumping at 1064 nm. The resonators were tested by placing them within a circularly symmetric Gaussian pump beam to implement three pumping configurations (single ring, evenly pumped double ring, and PT-symmetric ring arrangement). Accordingly, the effective pump powers were calculated from the geometric overlap between the active medium and the pump profile. **Fig. 4(a,b)** illustrate the behavior of a single ring (radius 10 μm, ring width 500 nm, height 210 nm) when exposed to an effective peak pump power of 2.5 mW (15 ns pulses with a repetition rate of 290 kHz). Under these conditions, at least four modes contribute significantly to lasing in the isolated ring. When two such rings are both placed at a distance of 200 nm from one another and exposed to the same pump power, one can clearly see the coupling-induced mode splitting (**Fig. 4(c,d)**), which occurs symmetrically around the resonance wavelengths of each ring in isolation. In this coupled regime, both structures are contributing equally. Once PT-symmetry is established, by withholding the pump from one of the resonators (**Fig. 4(e,f)**), lasing occurs exclusively in the active ring, where single-mode operation is now achieved. The presence of the lossy ring serves to suppress the unwanted longitudinal modes with a contrast exceeding 20 dB.

Comparing the light-light curves of these three lasing arrangements reveals that the slope efficiency in all scenarios remains virtually the same- indicating that the presence of lossy cavity in the PT-symmetric arrangement has no effect on the efficiency and output power. Moreover, the spectrally resolved light-light curves that compare the power in the fundamental mode of the single and the PT double ring resonators show that the PT system indeed offers superior performance, since the emission from the single ring includes contributions from several modes. This mechanism of mode selectivity using parity-time symmetry is robust in terms of fabrication inaccuracies and can accommodate active media with wide gain spectra. Moreover, as the occurrence of PT symmetry breaking is exclusively determined by the relation between net gain and coupling, the proposed arrangement is by nature self-adapting.

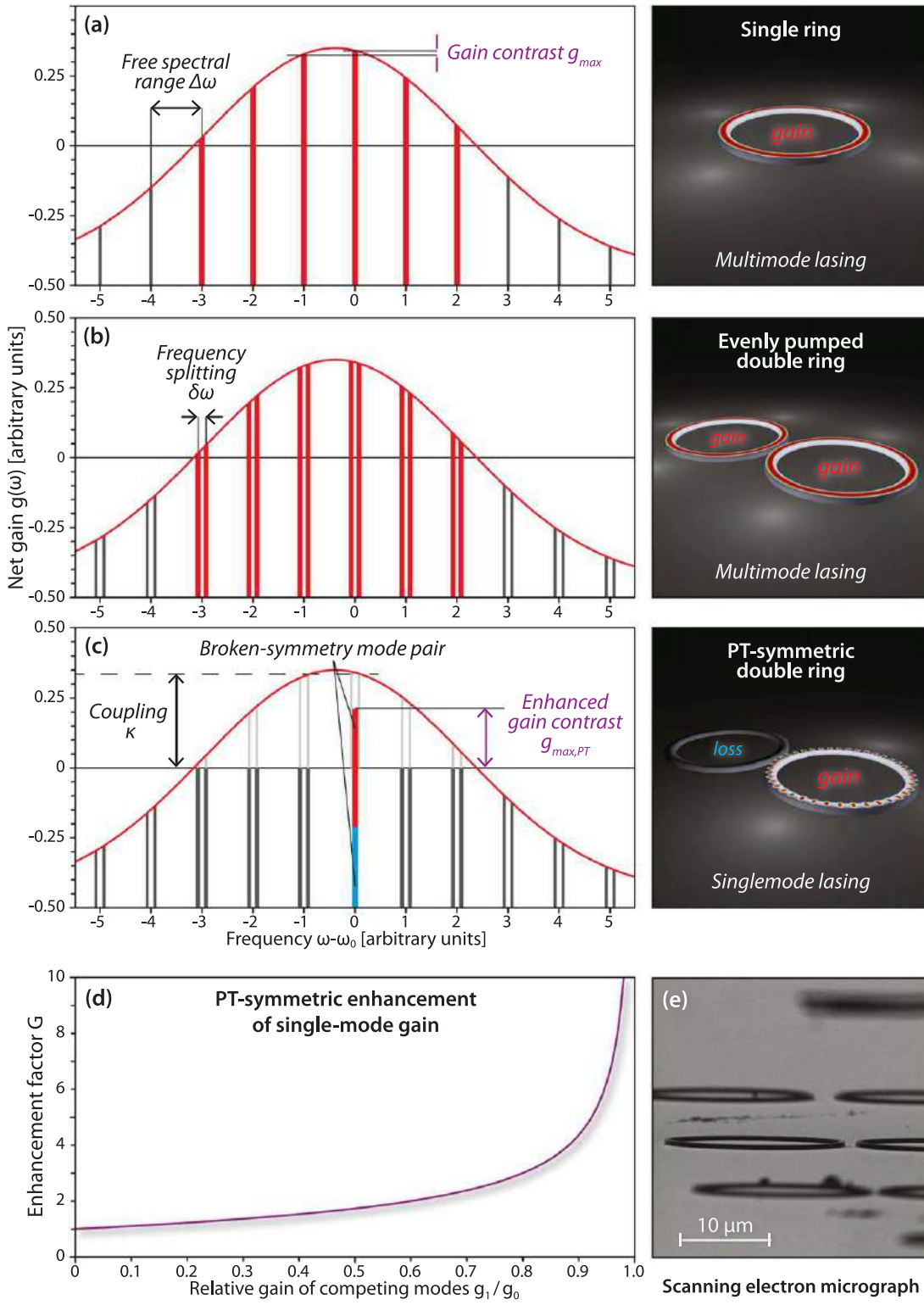


Fig. 3 Schematic principle of mode suppression in PT-symmetric microring lasers. (a), (b) In a coupled arrangement of two identical and evenly pumped rings, mode pairs emerge. (c), (d) PT-symmetry can be exploited to enforce stable single mode operation in otherwise multi-moded cavities. Reproduced from Hodaei, H., Miri, M.A., Heinrich, M., Christodoulides, D.N., Khajavikhan, M., 2014. Parity-time–symmetric microring lasers. Science 346, 975–978.

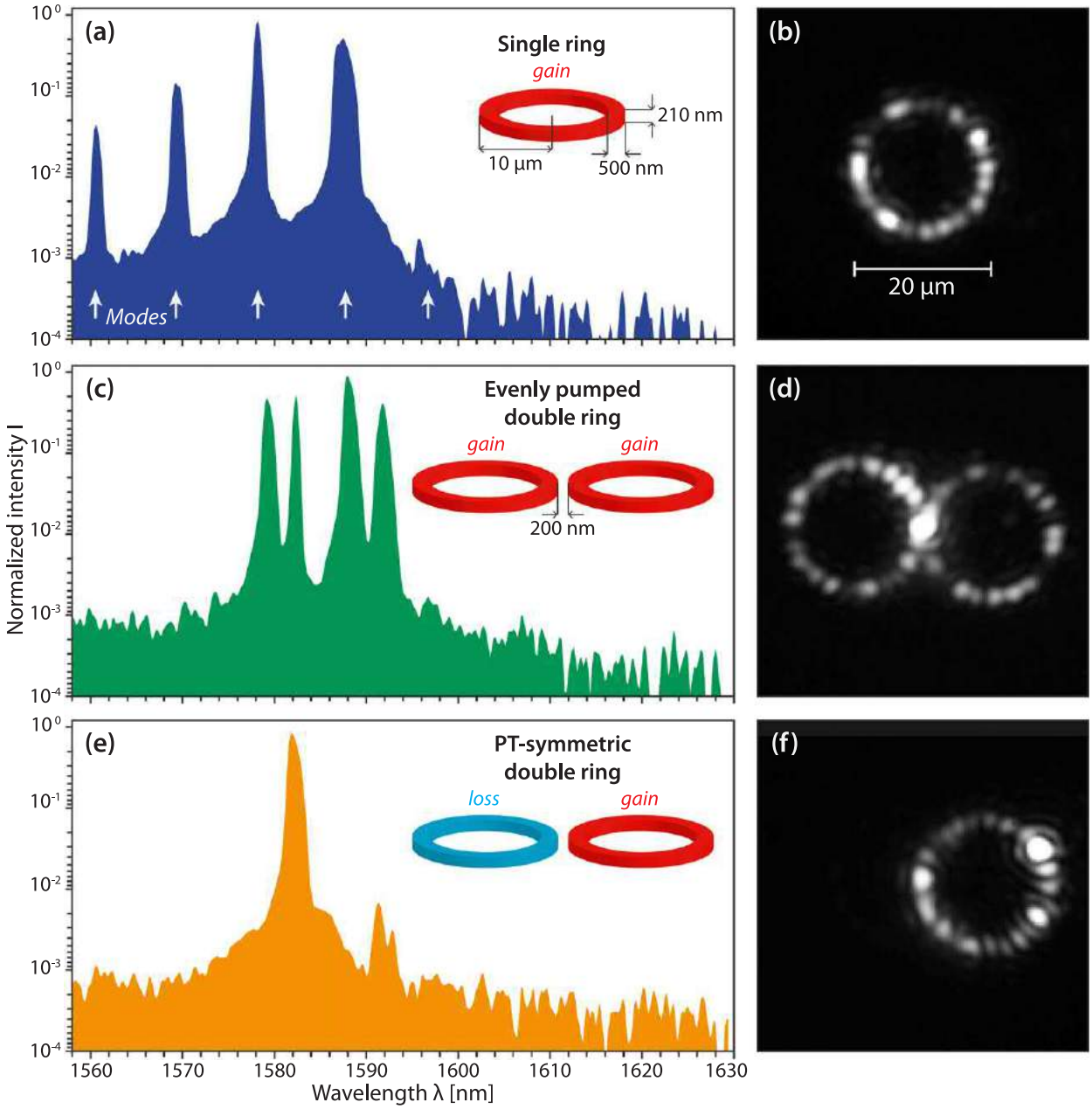


Fig. 4 Experimental observation of mode suppression by PT-symmetry breaking. (a) Emission spectrum of a single resonator. (b) Corresponding intensity pattern within the ring as observed from scattered light. (c) Spectrum obtained from an evenly pumped pair of such rings. (d) The intensity pattern shows that both resonators equally contribute. (e) Single-moded spectrum under PT-symmetric conditions. The mode suppression ratio exceeds 20 dB. (f) Lasing exclusively occurs in the active resonator. Reproduced from Hodaei, H., Miri, M.A., Heinrich, M., Christodoulides, D.N., Khajavikhan, M., 2014. Parity-time–symmetric microring lasers. Science 346, 975–978.

Laser Transverse Mode Filtering

PT-symmetry can also be utilized in promoting the fundamental transverse mode in broad-area and multi-moded microring lasers (Hodaei *et al.*, 2016). In fact, as shown in Eq. (2), the virtual threshold at the exceptional point $g/\kappa=1$ introduces an additional degree of freedom: the coupling constant κ between the active and the lossy cavity mediated by the evanescent overlap of their respective modes. As it is well known, higher-order spatial modes systematically exhibit stronger coupling coefficients due to their lower degree of confinement. Consequently, in a PT symmetric arrangement, as the gain increases (when $g>\kappa$), the fundamental mode is the first in line to break its symmetry, thus experiencing a net amplification. On the other hand, for this same gain level, the rest of the modes retain an unbroken symmetry and therefore remain entirely neutral.

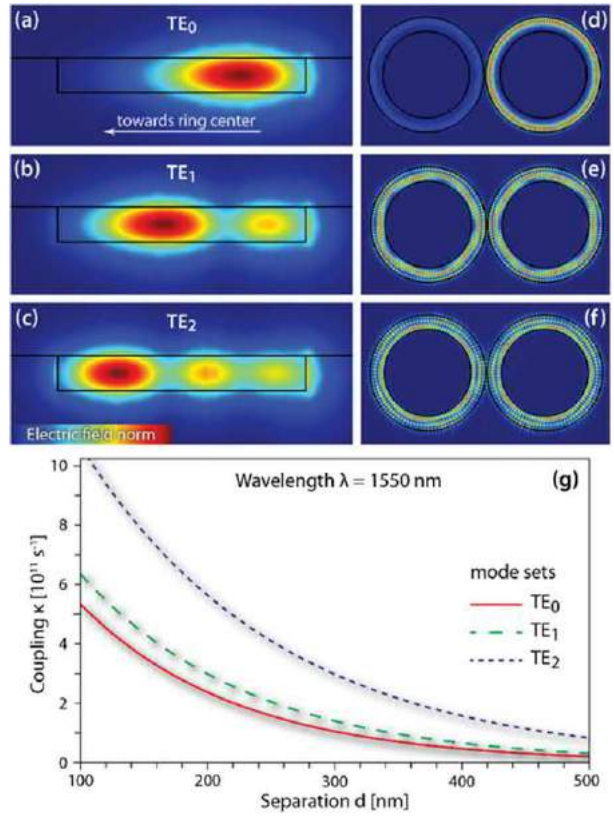


Fig. 5 (a)–(c) Intensity distributions in a microring resonator with a cross section 0.21 μm × 1.5 μm and a radius of R = 6 μm as obtained by finite element simulations for the first three transverse modes, (d)–(f) intensity distribution of these same modes within the PT symmetric ring resonators. While the TE_0 mode operates in the broken PT symmetry regime and lases, all other modes remain in their exact PT phase and therefore they stay neutral, and (g) Exponential decay of the temporal coupling coefficients κ with cavity separation *d*. Higher order modes exhibit a much larger coupling coefficient than their lower-order counterparts. Reproduced from Hodaei, H., Miri, M.A., Hassan, A.U., *et al.*, 2016. Single mode lasing in transversely multi-moded PT-symmetric microring resonators. Laser & Photonics Reviews 10 (3), 494–499.

Fig. 5(a)–(c) depicts the transverse intensity profiles of different spatial modes supported by a single microring resonator with a radius of $R = 6$ μm and waveguide dimensions of 0.21 μm × 1.5 μm. The curvature of the ring imposes a radial potential gradient, which deforms the mode fields into whispering-gallery-like distributions. Whereas the centroid of all modes shift towards the ring center, the exponential decay outside the ring still grows strongly with the mode order. As shown in **Fig. 5(d)–(f)**, for a certain coupling coefficient, set by the distance between the two rings, the transverse TE_0 field is the only mode to break its PT symmetry while all the higher order modes (TE_1, TE_2) are still in the exact PT phase and hence occupy both rings equally. This behavior is also evident in **Fig. 5(g)** where the coupling strength between different transverse modes is depicted as a function of the separation between the two rings. For a fixed distance, the coupling coefficient increases with the order of the transverse mode, since the effective indices of higher order modes lie closer to that of the surrounding medium, allowing for stronger evanescent interactions across the cladding region. This trend persists for all wavelengths, and in conjunction with the difference in confinement enables PT symmetry breaking to be employed as a mode-selective virtual loss.

Fig. 6(a) and **(b)** illustrate the transition from multimode behavior in a broad area microring laser to single mode operation in a twin-ring configuration as enabled by preferential PT symmetry breaking. These experiments were conducted in high-contrast active ring resonators based on quaternary InGaAsP (Indium-Gallium-Arsenide-Phosphide) multiple quantum wells embedded in SiO_2 (silicon dioxide). The gain bandwidth of the active medium spans the spectral region between 1260 and 1590 nm. Whereas the quantum wells are present in all wave-guiding sections, gain and loss is provided by selectively pumping the respective rings (pump wavelength: 1064 nm). **Fig. 6(a)** shows the spectrum obtained from a single microring laser. As expected from the differences in confinement between mode sets, only the TE_0 and TE_1 and TE_2 modes undergo lasing. In an isolated ring, a decrease of the overall pump power does not yield single mode operation. On the other hand, when the active ring of the PT-symmetric double-ring arrangement is supplied with the same pump power as in **Fig. 6(a)**, the TE_1 and TE_2 modes are readily eliminated (see **Fig. 6(b)**) as they fall below the PT breaking threshold and therefore experience zero net gain. The corresponding resonances are suppressed with a fidelity of over 30 dB down to the noise floor of the measurement.

The PT transverse mode filtering mechanism described above is fundamentally different from the standard spatial mode filtering techniques that are typically used in lasers. For example, intra-cavity apertures or spatial filters have long been used to

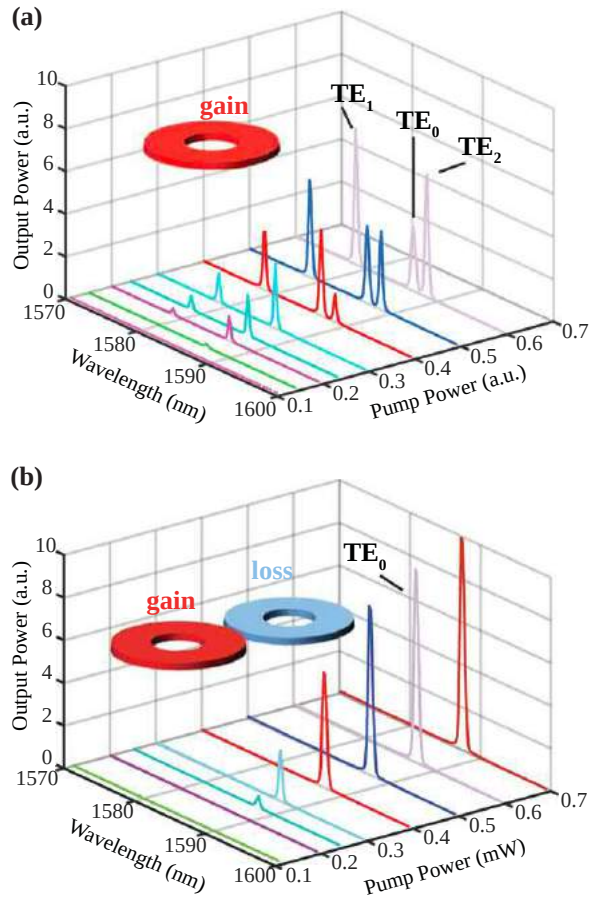


Fig. 6 (a) A single microring resonator lases in various TE_0, TE_1 and TE_2 resonances, (b) in contrast, a PT-symmetric double ring arrangement reliably suppresses all the TE_1 and TE_2 modes with a contrast exceeding 30 dB. Reproduced from Hodaei, H., Miri, M.A., Hassan, A.U., *et al.*, 2016. Single mode lasing in transversely multi-moded PT-symmetric microring resonators. Laser & Photonics Reviews 10 (3), 494–499.

limit the presence of higher order transverse modes by introducing differential loss between modes. However, in such schemes, all modes including the fundamental one, experience loss, though to a varying degree. Therefore, the resulting mode suppression obtained by such techniques is often incremental and comes at the expense of undesirable losses in the fundamental mode. In contrast, the abrupt onset of PT symmetry breaking at the exceptional point relocates the selected mode to the active region, whereas all higher order modes remain equally distributed between loss and gain regions. Consequently, this method offers a high degree of mode discrimination, without any detrimental impact on the overall lasing efficiency or output power.

Another advantage of PT-symmetric laser cavities is their robustness with respect to nonlinearity-induced instabilities. Conventional broad-area weakly-guiding single-mode lasers often are subject to severe filamentation even at moderate intensities. In contrast, the influence of such effects on the above system remains negligible. The observed robustness against nonlinear perturbations can be intuitively understood by keeping in mind that in the microring systems examined here, the index contrast between the core and cladding regions is significant and hence for all practical purposes is not affected by nonlinearity-induced index changes. For example, for circulating powers in the range of tens of mW, the resulting change in the refractive index due to self-focusing properties of the semiconductor gain material is estimated to be at most on the order of 10^{-3}. This level of nonlinearity will be an issue only for much more densely packed sets of transverse modes, e.g., when the width of the waveguides involved is substantially increased.

In conclusion, single spatial mode operation can be effectively enforced in parity-time symmetric broad area coupled cavities. Unlike most other schemes for laser transverse mode control, this method establishes mode selectivity through coupling to a lossy cavity without compromising the optical power extracted from the fundamental mode. Following the initial experiments, more recently, this approach has been tested in the context of two coupled waveguide lasers. This technique is versatile, scalable and can be applied to a wide range of broad area laser systems.

Sensing at Exceptional Points

Degeneracy occurs ubiquitously in nature. Within the context of eigenvalue problems, this property is encountered in a broad range of physical disciplines and is behind some of the most intriguing phenomena observed in both classical and quantum

physics. In terms of applications, through lifting an existing degeneracy, a small perturbation can lead to a detectable splitting between the corresponding eigenvalues. This principle has been utilized as a measurement tool in many and diverse settings. Similarly, in optics, breaking the degeneracy of resonant frequencies can be used as a means to detect small changes in the refractive index and/or loss. Along these lines, ultrahigh Q microcavities have been developed that nowadays are regarded as one of the most promising photonic arrangements for a variety of extreme sensing applications, ranging from single molecule detection to recording minute rotation rates.

The action of a perturbation on the degeneracy of a conservative (Hermitian) system is well understood. A point in parameter space at which such a degeneracy occurs is called a diabolic point (DP). At such points, a disturbance of strength ε leads to energy shifts and splittings that are at most on the order of ε. In non-Hermitian settings, however, another type of degeneracy is possible: that associated with an exceptional point. At this point, the dimensionality of the arrangement is abruptly reduced and as a result, the reaction of a system to perturbations around an EP is more drastic as compared to that at a DP. In fact, a perturbation of strength ε acting on a second-order EP (when only two eigenvalues and eigenvectors merge) so happens to result to an energy splitting that is proportional to $\varepsilon^{1/2}$. This implies that the sensitivity of a sensing set-up can be enhanced by several orders of magnitude (since $\varepsilon^{1/2} \gg \varepsilon$ for small perturbations) by exploiting the physics of EPs. Furthermore, one can show that the enhanced sensing of non-Hermitian systems is not limited to that offered by $\varepsilon^{1/2}$. By using even higher-order exceptional points (EP-Ns), it is possible to further boost the sensitivity up to $\varepsilon^{1/N}$. For example, an EP system of order N=5 can "amplify" a disturbance on the order of 10^{-10} up to 10^{-2}, an eight order improvement over that expected from a Hermitian sensor.

Exceptional points appear in various non-Hermitian settings, where there is coupling between resonances of the system while they are subjected to a spatially non-uniform gain and/or loss distribution. As of today, however, perhaps one of the most elegant ways to systematically generate them is through PT symmetry. In applications pertaining to sensing, a key advantage of PT-symmetric systems is their high sensitivity to external perturbations. Such a response not only facilitates the experimental observation of the square-root behavior as suggested above, but it can also provide a practical route to increase the sensitivity of microcavity sensors. **Fig. 7** depicts schematics of such non-Hermitian sensor arrangements comprising of coupled resonators. The cavities are identical in shape and size, but are subject to different levels of gain and/or loss. The contrast between the amplification levels experienced by the cavities can be introduced, for example, through preferential pumping. Although strictly speaking, such systems may not be invariant under the simultaneous action of parity (P) and time (T) operators, they can nevertheless be transformed into becoming PT symmetric once they are gauged through a constant gain/loss bias. **Fig. 7(d)** shows how the sensitivity in such arrangements can be enhanced depending on the order of the exceptional point involved.

The mathematical explanation behind the enhanced sensitivity expected from a PT symmetric coupled cavity configuration (when biased at an exceptional point) can be obtained, for example, by considering the modal behavior of the structure depicted in **Fig. 1(b)**. In general, each ring, when uncoupled, can support a number of longitudinal modes both in the clockwise and counterclockwise directions. Without loss of generality, here we limit our analysis to a single longitudinal mode in one direction. We also assume that the cross- section of the rings are designed so as to support only the fundamental TE mode. In order to analyze the sensitivity response, we assume that a small perturbation is applied to the ring resonator subjected to gain. In this respect, the interplay between the electric modal fields in the two identical rings can be effectively described through a set of

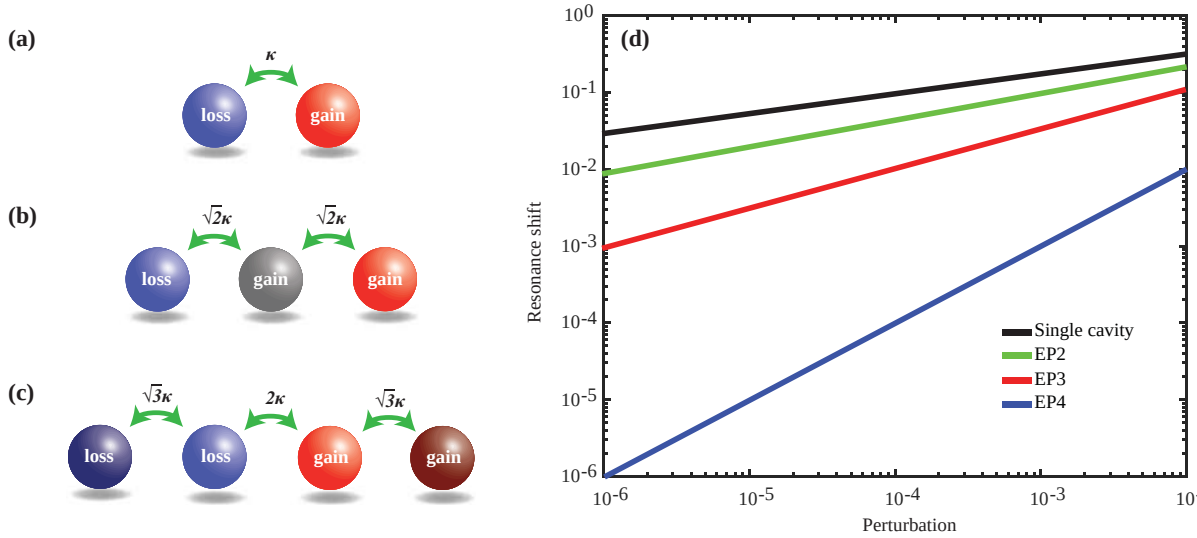


Fig. 7 A schematic of three photonic PT molecules exhibiting (a) 2nd , (b) 3rd, and (c) 4th order exceptional points. Red cavities exhibit gain, blue loss, and gray are neutral. (d) The dependence of sensitivity to perturbation strength.

time dependent coupled equations:

$$\begin{cases} \dfrac{da}{dt} = -i(\omega_0 + \epsilon)a + i\kappa b + g_1 a \\ \dfrac{db}{dt} = -i\omega_0 b + i\kappa a + g_2 b \end{cases} \tag{4}$$

Here, the cavities are treated as lumped resonators, a and b are modal amplitudes for a particular longitudinal resonance, κ and $g_{1,2}$ represent the coupling factor between the resonators and their associated gain/loss values, respectively, and ϵ is an intentionally induced perturbation. The two eigenvalues associated with this set of equations are given by the following expression:

$$\omega_{1,2} = \omega_0 + \epsilon/2 \pm \sqrt{\kappa^2 - (\Delta g/2 + i\epsilon/2)^2} \tag{5}$$

where $\Delta g = |g_1 - g_2|$ represents the gain-loss contrast between the resonators. If $\kappa = \Delta g/2$, the system is known to be at the exceptional point. Under this condition, Eq. (2) can be rewritten as:

$$\omega_{1,2} = \omega_0 + \epsilon/2 \pm i\sqrt{i\epsilon\kappa/2} \tag{6}$$

In the absence of any perturbation ($\epsilon = 0$), the two frequencies are expected to be exactly equal to ω_0. This is a large departure from a standard Hermitian system in which the two resonances are expected to split by 2κ. Furthermore, when operating at the exceptional point ($\kappa = \Delta g/2$), an induced perturbation ϵ leads to a $\Delta\omega_{PT} = \sqrt{2\kappa\epsilon}$ splitting in the real part of the eigen-frequencies. Therefore, by monitoring the splitting between the two resonances in the spectral domain, one can effectively determine the magnitude of the applied perturbation. Clearly, in this two-level system, the corresponding wavelengths diverge according to a square-root function of the perturbation ($\Delta\lambda \sim \epsilon^{1/2}$). Compared to a Hermitian or Hermitian-like single cavity configuration, the aforementioned square root behavior can substantially amplify the response of the system to a small perturbation (Chen *et al.*, 2017; Hodaei *et al.*, 2017). This approach has been recently suggested as a means to devise ultrasensitive microscale gyroscopes (Ren *et al.*, 2017).

Anomalous Scattering Properties of PT-Symmetric Arrangements

At a first sight, it may seem that balanced regions of gain and loss in a PT-symmetric system cannot affect the propagation of light in such settings, as the net amplification/attenuation in such systems is zero. However, as shown in several recent studies, the inter-mixing of gain and loss can significantly alter the dynamics of light in PT systems and can lead to a host of intriguing phenomena (Miri *et al.*, 2016). In order to shed light on anomalous scattering properties of PT-symmetric arrangements, here we focus on a PT-symmetric grating defined as follows (Lin *et al.*, 2011; Regensburger *et al.*, 2012):

$$n(z) = n_0 + n_R \cos(k_B z) + i n_I \sin(k_B z) \tag{7}$$

where, $k_B = 2\pi/\lambda_B$ and λ_B is the grating period. Clearly, this refractive index profile satisfies the necessary condition of PT symmetry, i.e., $n^*(-z) = n(z)$. Under weak refractive index and gain/loss modulations, i.e., $n_R, n_I \ll n_0$, this system is analytically solvable and its dynamics is governed by the coupling between the forward and backward propagating waves, as well as the gain/loss contrast. To show this, we consider a solution of the form $E(x,t) = E_f(x,t)e^{-i(\omega_0 t - k_0 n_0 z)} + E_b(x,t)e^{-i(\omega_0 t + k_0 n_0 z)}$, where E_f and E_b represent the slowly varying envelopes of the forward and backward propagating waves, ω_0 is a central frequency and $k_0 = \omega_0/c$ is the free space wavenumber. It is straightforward to show that this ansatz leads to the following simple coupled mode equations:

$$\left(\frac{\partial E_f}{\partial z} - \frac{1}{v}\frac{\partial E_f}{\partial t}\right) = +i(\kappa + g)e^{-i2\delta z}E_b \tag{8a}$$

$$\left(\frac{\partial E_b}{\partial z} + \frac{1}{v}\frac{\partial E_b}{\partial t}\right) = -i(\kappa - g)e^{+i2\delta z}E_f \tag{8b}$$

where, in these relations, $v = c/n_0$ represents the phase velocity in the background medium, $\delta = k_0 n_0 - k_B/2$ shows the detuning in the grating-assisted phase matching of the forward and backward propagating waves, and finally $\kappa = k_0 n_R/2$ and $g = k_0 n_I/2$ are two coupling coefficients associated with the real and imaginary parts of the complex grating. An interesting property of the coupled mode Eq. (8) is an asymmetry in the coupling between the forward and backward waves; according to these relations, the forward wave is coupled to the backward wave at rate $\kappa + g$, while the reverse process occurs at rate $\kappa - g$. Quite interestingly, for $\kappa = g$, the backward wave becomes completely independent from the forward. As a result, the complex grating will not have any influence on a wave propagating toward the left.

Under the ansatz of $(E_f, E_b) = (F_0 e^{-i\delta z}, B_0 e^{i\delta z})e^{i\delta v t}e^{i(Kz - \Omega t)}$, the band structure of the PT grating is found to be:

$$\Omega^2/v^2 = K^2 + \kappa^2 - g^2 \tag{9}$$

This dispersion diagram is shown in **Fig. 8** for different ratios of g/κ. According to this figure, for $\kappa > g$ there is a finite band gap $\Delta_\omega = 2v\sqrt{\kappa^2 - g^2}$. By increasing the gain/loss, the band gap reduces such that for $\kappa = g$ the dispersion relation becomes the same as a the background medium, i.e., $\Omega/v = \pm K$. Finally, for $\kappa < g$, the dispersion diagram flips and regions with complex eigen-frequencies, associated with amplification and attenuation, emerge.

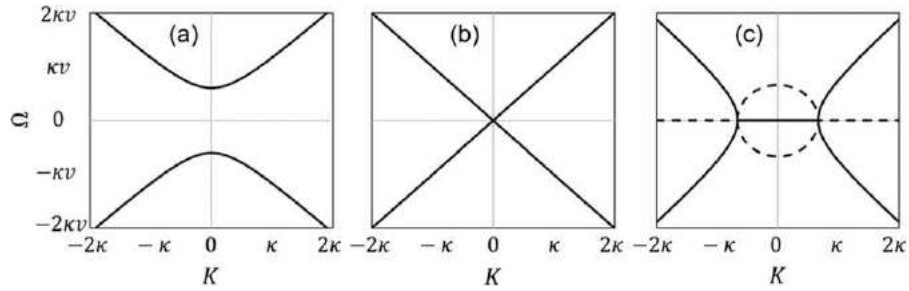


Fig. 8 The dispersion diagram of a PT-symmetric grating for (a) $g=0.8\kappa$, (b) $g=\kappa$, and (c) $g=1.2\kappa$. The dashed line in part (c) shows the imaginary part of Ω.

Such periodic PT-symmetric structures can exhibit surprising behavior such as unidirectional invisibility and unusual reflection characteristics. More specifically, light propagating in such a system can experience reduced or enhanced reflections (with a coefficient that can even exceed unity) depending on the direction of light propagation. This can be understood by considering that the left–right symmetry is now broken in this PT configuration and propagation is no longer the same when the sequence of the gain and loss regions is exchanged. Even more interesting is what happens at the PT threshold: light waves entering the structure from one side do not experience any reflection and can fully traverse the grating with complete transmission. Given that this occurs without acquiring any phase imprint from the PT grating, the periodic structure is essentially invisible. The situation is very different when light is incident from the other side.

See also: Nonlinear Optical Phase Conjugation

References

Bender, C.M., Boettcher, S., 1998. Real spectra in non-Hermitian Hamiltonians having PT symmetry. Physical Review Letters 80 (24), 5243.
Chen, W., Özdemir, S.K., Zhao, G., Wiersig, J., Yang, L., 2017. Exceptional points enhance sensing in an optical microcavity. Nature 548, 192–196.
Doppler, J., Mailybaev, A.A., Böhm, J., *et al.*, 2016. Dynamically encircling an exceptional point for asymmetric mode switching. Nature 537 (7618), 76–79.
Feng, L., Wong, Z.J., Ma, R.M., Wang, Y., Zhang, X., 2014. Single-mode laser by parity-time symmetry breaking. Science 346 (6212), 972–975.
Guo, A., Salamo, G.J., Duchesne, D., *et al.*, 2009. Observation of PT-symmetry breaking in complex optical potentials. Physical Review Letters 103 (9), 093902.
Hodaei, H., Hassan, A.U., Wittek, S., *et al.*, 2017. Enhanced sensitivity in photonic molecule systems using higher order non-Hermitian exceptional points. Nature 548, 187–191.
Hodaei, H., Miri, M.A., Hassan, A.U., *et al.*, 2016. Single mode lasing in transversely multi-moded PT-symmetric microring resonators. Laser & Photonics Reviews 10 (3), 494–499.
Hodaei, H., Miri, M.A., Heinrich, M., Christodoulides, D.N., Khajavikhan, M., 2014. Parity-time–symmetric microring lasers. Science 346, 975–978.
Lin, Z., Ramezani, H., Eichelkraut, T., *et al.*, 2011. Unidirectional invisibility induced by PT-symmetric periodic structures. Physical Review Letters 106 (21), 213901.
Makris, K.G., El-Ganainy, R., Christodoulides, D.N., Musslimani, Z.H., 2008. Beam dynamics in PT symmetric optical lattices. Physical Review Letters 100 (10), 103904.
Miri, M.A., Eftekhar, M.A., Facao, M., *et al.*, 2016. Scattering properties of PT-symmetric objects. Journal of Optics 18 (7), 075104.
Miri, M.A., LiKamWa, P., Christodoulides, D.N., 2012. Large area single-mode parity-time-symmetric laser amplifiers. Optics Letters 37 (5), 764–766.
Peng, B., Özdemir, Ş.K., Lei, F., *et al.*, 2014. Parity-time-symmetric whispering-gallery microcavities. Nature Physics 10 (5), 394–398.
Regensburger, A., Bersch, C., Miri, M.A., *et al.*, 2012. Parity-time synthetic photonic lattices. Nature 488 (7410), 167–171.
Ren, J., Hodaei, H., Harari, G., *et al.*, 2017. Ultrasensitive micro-scale parity-time-symmetric ring laser gyroscope. Optics Letters 42 (8), 1556–1559.
Rüter, C.E., Makris, K.G., El-Ganainy, R., *et al.*, 2010. Observation of parity-time symmetry in optics. Nature Physics 6 (3), 192–195.
Wiersig, J., 2014. Enhancing the sensitivity of frequency and energy splitting detection by using exceptional points: Application to microcavity sensors for single-particle detection. Physical Review Letters 112 (20), 203901.
Xu, H., Mason, D., Jiang, L., Harris, J.G.E., 2016. Topological energy transfer in an optomechanical system with exceptional points. Nature 537 (7618), 80–83.

Laser-Induced Damage in Optical Materials

Wolfgang Rudolph and Luke A Emmert, University of New Mexico, Albuquerque, NM, United States

Introduction

Shortly after the invention of the laser, in the early 1960s, it became apparent that the ability of optical materials to withstand high laser intensity and fluences was of utmost importance for advances in laser development and nonlinear optics. Despite considerable progress in the past decades (Wood, 2003; Ristau, 2015) the physics and materials science of the interaction of high-intensity lasers with optical materials is still a very active and challenging research area. Laser-induced damage (LID) and laser ablation rest on similar physical foundations and describe the phenomena of irreversible material modification. Laser induced damage of surfaces and films is often associated with the removal of material.

The defining parameter is the laser-induced damage threshold (LIDT). Its value is often published as a fluence (J cm^{-2}) for pulsed laser systems and as an intensity (W cm^{-1}) for continuous wave (CW) lasers. The CW LIDT is valid for exposures long enough to reach steady-state temperatures. Typical time scales range from microseconds to seconds depending on the spot size. Solutions to the thermal diffusion equation reveal that the maximum temperature reached is proportional to the incident power divided by the beam radius (Ristau, 2015, see Chapter 2) for uniform bulk and surface absorption. Pulse LID is the subject of this article. LIDT values depend on material type (metals, semiconductor, insulators), topology (bulk, surface, thin film, multilayer), preparation (surface polishing, film deposition method, annealing, etc.), as well as laser parameters (wavelength, pulse duration, spot size, repetition rate, etc.). They also depend on how the damage test is performed. For example, if each test pulse illuminates a new sample site (1-on-1 test) or if several (S) pulses illuminate the same sample spot (S-on-1) (ISO 11254-2, 2011). Being transparent dielectrics have the highest LIDTs of all materials for lasers from the near infrared to the near ultraviolet spectral region.

Fig. 1 illustrates the material response of a dielectric sample as a function of the incident fluence. Often the LIDT behavior of a material is characterized by the probability that damage is observed when a pulse of certain fluence is incident, $P(F)$. There is a fluence range that separates catastrophic damage (for instance a surface crater) from reversible material modifications such as refractive index changes. The width of this transition region ΔF is narrow for fs pulses and becomes increasingly broader for ps, ns, and longer pulses. Different physical origins of damage are responsible – deterministic damage initiation for short pulses governed by intrinsic material properties and probabilistic, extrinsic-defect controlled damage for longer pulses. It should be noted that the physical mechanisms responsible for optical damage also enable laser machining and material processing.

The LIDT F_{th} is defined as the minimum fluence for damage to occur, i.e., $P(F \leq F_{th}) = 0$ (Porteus and Seitel, 1984). An optic will remain undamaged provided the incident fluence $F < F_{th}$. Because of the statistical nature of laser-induced damage there are practical difficulties in determining F_{th} so a damage fluence F_d, where $P(F = F_d) = 0.5$, is often used in the scientific literature. For short pulse LID, where the transition region ΔF is narrow, F_{th} and F_d are nearly the same. LIDT can be different for 1-on-1 and S-on-1 tests. Typically, the multiple pulse LIDT is smaller. There are also cases where the damage threshold increases when the sample is illuminated with a laser. This laser conditioning is used in high-power laser systems by gradually ramping up the laser power. Annealing of defects is a possible mechanism enabling laser conditioning (Bercegol, 1998).

Table 1 lists LIDTs for representative materials and pulse durations. Scaling laws (for dielectrics in particular) are known to extrapolate to other materials and pulse durations (Ristau, 2015, see chapter 5).

To discuss issues related to LIDT we refer to the schematic diagram of the morphology of an optical surface (or film) shown in **Fig. 2**.

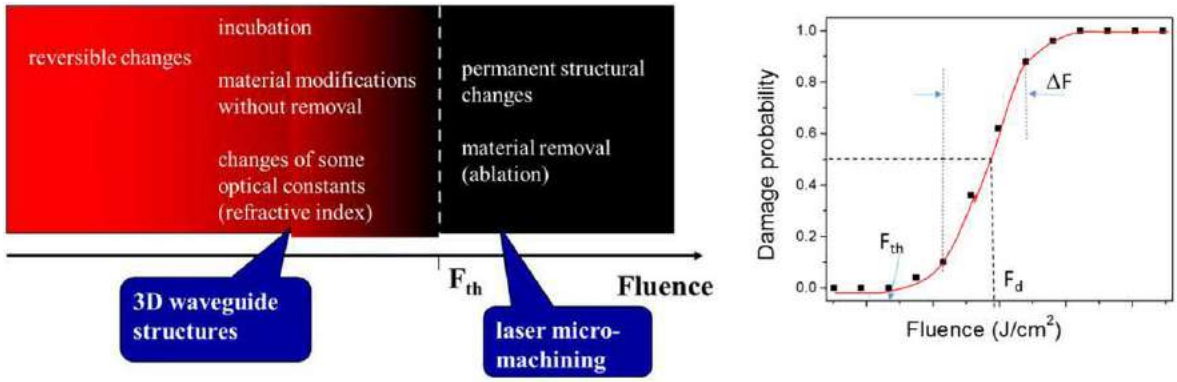


Fig. 1 Response of a dielectric material as a function of incident pulse fluence. The graph on the right shows a typical damage probability as a function of incident pulse fluence measured with femtosecond pulses that probe intrinsic material properties (F_{th}, F_d, and ΔF are defined in the text.).

 doi:10.1016/B978-0-12-803581-8.09451-0

Table 1 Representative 1-on-1 LIDTs for selected optical components, pulse durations and wavelengths

Optics	*Pulse duration*	*Wavelength*	*LIDT (J cm^{-2})*	*References*
Fused silica	100 fs	800 nm	~4	Lenzner *et al.* (1998)
(bulk)	10 ns	1064 nm	5000	Smith and Do (2008)
Dielectric HR mirror	100 fs	800 nm	0.2–1.0	Stolz *et al.* (2009)
	10 ns	1064 nm	5–140	Stolz *et al.* (2008)
Dielectric polarizer	10 ns	1064 nm	18	Stolz and Runkel (2012)
Metal (Al) film	100 fs	800 nm	0.3	Sun *et al.* (2015)

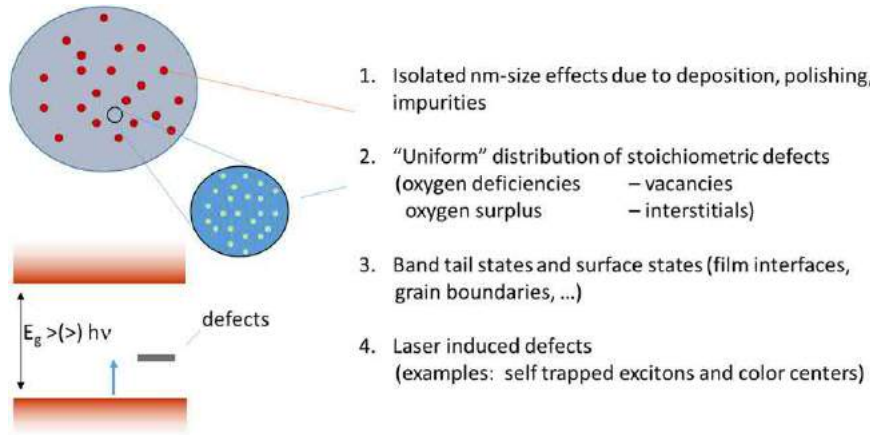


Fig. 2 Schematic diagram of the morphology of an optical wide-gap material. In high-quality materials, the density of defects of type 1 is low. Defects of type 2 and 3 are considered intrinsic. They determine macroscopic material properties probed by an incident laser. Additional defects (type 4) can be created during sub-threshold laser irradiation, for example, by a train of pulses. Defects can introduce states within the bandgap of the host material, which makes optical excitation more efficient, lowering the LIDT.

A laser excitation spot will typically contain several defects of type 1. Controlled by such statistically distributed defects, long-pulse LIDTs are stochastic and depend on spot size (DeShazer *et al.*, 1973). For sufficiently small spot sizes LIDTs representing intrinsic material properties are more likely to be observed. When measuring LIDTs of bulk materials care has to be taken to avoid or properly take into account propagation effects like self-focusing and dispersive effects in particular for short pulses. Characterization of damage controlling defects remains a challenge. Nondestructive techniques often lack sensitivity when the defect density is low.

Because of their importance for high power optical components dielectrics have received a lot of attention and are the focus of this encyclopedia entry. The physical processes leading to visible damage depend on the material and illumination conditions. In all cases a sufficiently large energy density has to be deposited first, which then triggers the material response that leads to the damage signature. This is illustrated in **Fig. 3**.

Fundamental Processes of Laser-Induced Damage

The intrinsic laser-induced damage mechanism of wide bandgap dielectrics is rather well understood (Emmert and Rudolph, 2015 and references therein). From the femtosecond to the nanosecond pulse regime, LIDT is reached when the excitation pulse has created an electron plasma of sufficient energy density in the conduction band. The processes leading to this electron plasma involve multiphoton ionization (interplay of multiphoton absorption and tunneling) of valence band electrons and impact ionization. The relative importance of these two processes has long been a question (Jones *et al.*, 1989) and remains an active research topic even today (Rajeev *et al.*, 2009; Karras *et al.*, 2011; Mouskeftaras *et al.*, 2013).

Damage will occur if the density of energy deposited into the conduction band plasma reaches some critical value. Its exact value is not *a priory* clear and depends, among other things, on the pulse duration. If damage were purely thermal, the critical energy would be related to the evaporation enthalpy. This, however, is rarely the case. A large enough density of conduction band electrons (electron plasma) can destabilize the lattice – similar to the excitation of a molecule to a non-binding state.

Dielectric breakdown used in modeling LID provides a useful proxy for the critical energy as a damage criterion (Bloembergen, 1974). The electron plasma can linearly absorb light through "free" carrier absorption also known as inverse Bremsstrahlung. With increasing absorption, the skin depth in which energy is deposited becomes smaller amplifying the rate of the energy deposition even more. Dielectric breakdown occurs at the plasma density (i.e., critical electron density) at which the plasma frequency

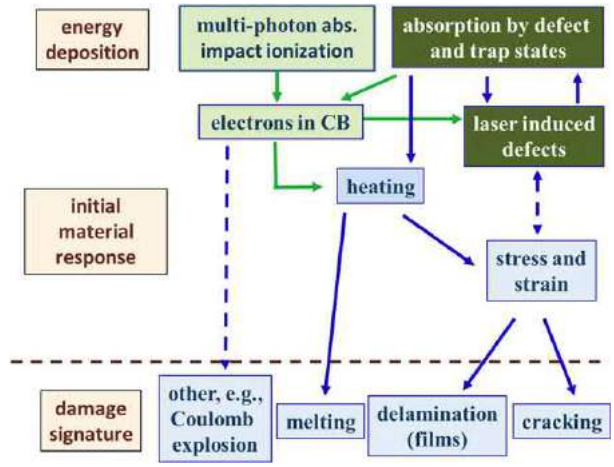


Fig. 3 Laser-induced damage in optical materials. Energy deposition is followed by an initial material response leading to visible, irreversible changes.

coincides with the laser frequency. The success of this criterion also stems from the fact that because of the exponential increase in plasma density during the incident pulse, the predicted threshold pulse fluence depends only logarithmically on the numerical value of the critical electron density.

Rate equations that model the electron density in the conduction have been quite successful in explaining the main features of intrinsic LIDT in particular in the short pulse regime. The principal structure of such a rate equation is of the form

$$\frac{d}{dt}n_e = K(I) + A(I, n_e) - L(n_e) \tag{1}$$

where n_e is the conduction band electron density and I is the laser intensity. The first term, $K(I)$, is the multiphoton ionization rate. It represents a combination of multiphoton absorption and tunneling described by Keldysh (1965). The second term, $A(I,n_e)$, describes impact (avalanche) ionization. This is the process where a highly excited electron in the conduction band (CB) transfers its energy to a valence band electron through collision, promoting it to the CB. This collision results in two electrons near the bottom of the CB. Repeating this process leads to exponential growth of n_e (avalanche ionization). This term is often approximated by aIn_e (from the flux-doubling model, which assumes that impact ionization takes place as soon as a CB electron has acquired sufficient energy (Stuart *et al.*, 1996)), where a is the avalanche ionization coefficient. This coefficient is usually taken as constant, but attempts have been made to measure an intensity dependence (Rajeev *et al.*, 2009; Karras *et al.*, 2011). The last term in Eq. (1) accounts for relaxation of electrons out of the conduction band. This term can represent either uni- or bi-molecular recombination with holes or existing trap states.

Using the concept of dielectric breakdown when a critical electron density n_{crit} is reached for LID, this rate equation predicts the observed subpicosecond pulse duration and bandgap scaling (Mero *et al.*, 2005) for subps pulses. Refinements to the rate equations are (1) multiple rate equations/multiple levels in the conduction band (Rethfeld, 2004); (2) effect of plasma on the local field in a film (Gallais *et al.*, 2010); and (3) estimation of energy deposition (Gallais *et al.*, 2015). Measurements and solutions to Eq. (1) for dielectric films and surfaces suggest that in the subps regime the LIDT scales approximately linearly with the material bandgap and as a power law with respect to pulse duration, τ^{κ} with $\kappa \approx 0.3 \ldots 0.4$, for 1-on-1 damage (Mero *et al.*, 2005).

Multiple pulse, S-on-1, LIDTs need to take into account accumulating laser-induced material changes. This material *incubation* is apparent in the dependence of LIDT on the number of pulses in the test that illuminate each sample site. As the number of pulses increases, the damage threshold decreases. Possible underlying mechanisms are that (1) during the pulse train laser-induced-defects and other material modifications increase the absorption and/or (2) the critical energy for damage decreases due, for example, to defect caused weakening of bonds. The latter can be expressed as a decrease in a critical enthalpy G, for example, evaporation enthalpy that leads to damage. In a simple model, the threshold fluence for pulse number S is determined by the absorption (coefficient) of the sample and the critical enthalpy after the previous pulse, $\alpha(S\text{-}1)$ and $G(S\text{-}1)$, respectively (Sun *et al.*, 2015):

$$F_{th}(S) \approx \frac{G(S-1)}{\alpha(S-1)} \tag{2}$$

Fig. 4 shows a typical incubation curve $F_{th}(S)$.

To explain incubation on a microscopic level we refer to the energy level diagram in **Fig. 5**. At fluences just below the LIDT a subcritical electron density is generated in the conduction band. Between pulses most of these electrons relax to the valence band, but some remain trapped in midgap native and laser-induced states. The electrons can be re-excited by subsequent laser pulses and enhance the generation of the conduction band plasma. The process repeats with each pulse until the midgap states become

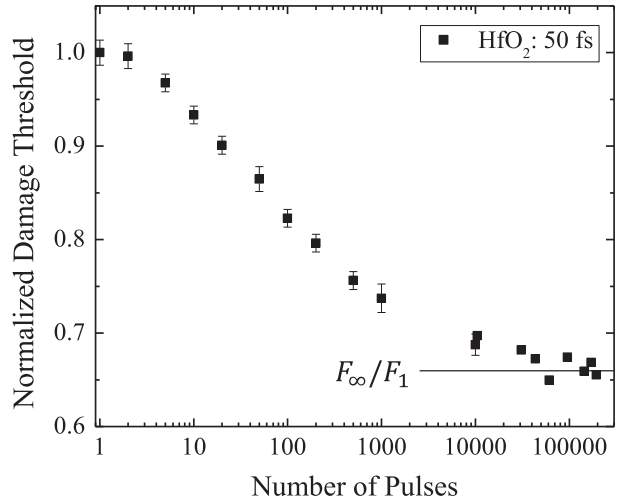


Fig. 4 Measured LIDT of a HfO_2 film as a function of the number of 50-fs pulses illuminating one and the same sample spot, normalized to the single-pulse LIDT F_1. The first (S-1) pulses modify (incubate) the material leading to a reduction of F_{th} for the last pulse (number S) that triggers catastrophic damage. The optic can be used as long as the laser fluence $F < F_\infty$, i.e., the multiple pulse LIDT. Reproduced from Nguyen, D.N., Emmert, L., Mero, M., *et al.*, 2008. The effect of annealing on the subpicosecond breakdown behavior of hafnia films. In: Proc. of SPIE, vol. 7132, p. 71320N.

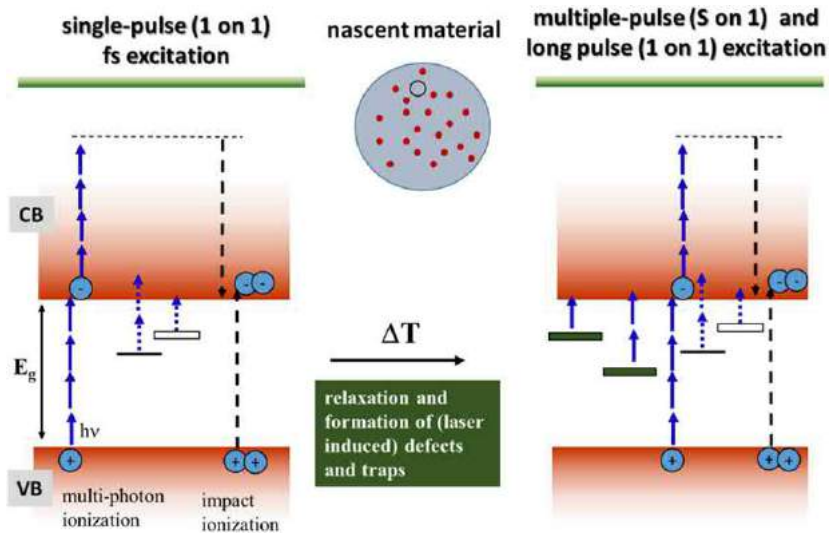


Fig. 5 Energy diagram of optical excitation and relaxation of dielectric materials. It is assumed here that the laser probes intrinsic material properties. The bandgap E_g exceeds the single photon energy. Optical excitation of electrons from valence band (VB) to conduction band (CB) is therefore possible only through a multi-photon process. The left group of processes is important for 1-on-1 events and short pulses; the right group plays a role for long pulses and pulse trains. Relaxation out of the CB can occur on a subps time scale. There is a certain probability that defects like self-trapped excitons and color centers are created (laser induced defects) during a long pulse or train of pulses.

saturated or damage occurs. This phenomenon can be modeled by adding additional rate equations (Emmert *et al.*, 2010) to include the various electronic states and transitions (see **Fig. 5**). In the case of fused silica, the microscopic states responsible for incubation are well known. Electrons in the conduction band combine with valance band holes to form self-trapped excitons (STEs). In most cases these relax through a non-radiative mechanism, but sometimes the STE will relax into an E′ center and a non-bridging oxygen. The generation of these defects has been exploited to create a variety of optical structures in bulk glass (Beresna *et al.*, 2014). In many oxide films, the trap states have not yet been identified, but oxygen vacancies are predicted (Gavartin *et al.*, 2006).

Defect-Induced Damage

At pulse durations on the order of picoseconds and longer, a second class of extrinsic defects often controls the optical performance limits even in high-quality materials (Papernov, 2015). Defects in dielectric oxides can be, for example, clusters of partially

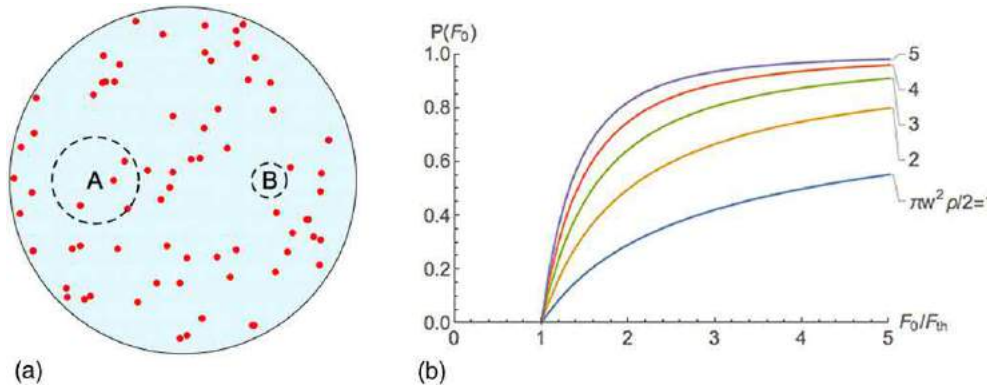


Fig. 6 (a) Schematic of an optical sample with randomly distributed extrinsic defects (dots) and incident laser beams of spot size A and B (dashed circles). The large spot (A) is likely to contain one or more defects and measure LIDT controlled by defects. A smaller beam (B) will often probe the intrinsic material that has a higher damage threshold. (b) $P(F_0)$ as a function of F_0/F_{th} for different values of the product $\pi w^2\rho/2$. The defect density and/or beam radius increases from bottom to top.

oxidized material with metal-like properties, impurities, or residues from surface polishing, cf. defects of type 1 in **Fig. 2**. Metal-like defects, for example, lack a bandgap and thus single photon absorption of long pulse and CW lasers can be efficient. One damage mechanism here can be the creation of a localized, hot, high-pressure plasma that eventually explodes, ejecting material. The fact that damage is controlled by defects is the main cause for the wide range of reported LIDT values.

As pointed out previously, the presence of such defects leads to an excitation spot size dependent LIDT. A large beam spot is more likely to illuminate defects that determine the damage threshold. A sufficiently small spot is likely to probe the intrinsic, defect-free material. This behavior is illustrated in **Fig. 6(a)**. In general, a material is described by a defect distribution $\overline{\rho}(F)$. The product $\overline{\rho}(F)dF$ gives the density (areal or volume) of defects whose damage threshold is in the range from F to $F+dF$. This distribution also depends on the wavelength of the excitation laser (Jensen *et al.*, 2009).

A consequence of defect-initiated damage is that LIDT is probabilistic rather than deterministic. For fluences below the intrinsic LIDT, the probability of damage is equivalent to the probability that a defect is located in a region of the beam where the fluence is sufficiently large to initiate damage. Given a defect distribution and a beam profile, the probability of damage can be predicted. An example that is often used is an ensemble of identical defects described by a damage threshold F_{th} and density ρ illuminated by a Gaussian beam of radius w ($1/e^2$ intensity) and peak fluence F_0. In this case, the damage probability is of the form

$$P(F_0) = 1 - (F_{th}/F_0)^{-\pi w^2\rho/2} \tag{3}$$

Note that the product in the exponent ($\pi w^2\rho$) is the average number of defects within a circle of radius w. The shape of the function $P(F_0)$ for different values of $\pi w^2\rho$ is illustrated in **Fig. 6(b)**.

A variety of nano- and microstructures can be responsible for defect-initiated damage. Subsurface cracks can result from the cutting and polishing of optical surfaces. Cracks have sharp edges that can lead to local field enhancements or act as local absorbers due to surface states. An important defect in multilayers, which are typical for optical mirrors and beam splitters, are nodules (Tench *et al.*, 1994). These are particles embedded in the multilayer stack during deposition. Large nodules can disrupt the multilayer leading to mechanically weak interfaces and local intensification of the electric field. Laser conditioning has proved particularly successful in mediating the performance limiting effect of nodules (Bercegol, 1998).

Finally, defect-initiated damage even occurs in material that appears free of structural defects. This observation is attributed to nanoabsorbers that may result from local substoichimetric or metal-like regions. These nanometer-scale absorbers trigger an expansion of the absorption volume at intensities above the LIDT which leads to micron-scale craters (Lange *et al.*, 1987; Papernov and Schmid, 2002; Demos *et al.*, 2013).

Measuring LIDT

The LIDT of materials and optical components is measured by destructive damage tests. A variety of tests have been developed to address both engineering and scientific goals. On the one hand LIDT measurements are used to determine the operational limits of commercial products and/or for qualifying an optic for use in a system. In this case a witness sample is often used. But also, as a scientific tool, damage testing is used for characterizing materials. Examples of scientific interest include pulse duration scaling discussed earlier, bandgap scaling, incubation, and the distribution and characterization of defects. LIDT tests reveal material imperfections that are often not detected by other (non-destructive) diagnostic measurements.

Traditional damage tests are based on the evaluation of whether damage occurred following illumination of the sample with a train of pulses. This evaluation can be done either immediately by observation of damage with a scatter probe or an in-situ microscope or later using post mortem microscopy. Alternatively, measuring the ablation crater diameter as a function of fluence

has also been applied successfully for LIDT tests. Extrapolation to the fluence where the crater diameter is zero gives the LIDT for 1-on-1 (Liu, 1982) and S-on-1 experiments (Sun *et al.*, 2015). The following are few common examples of LIDT measurements methods.

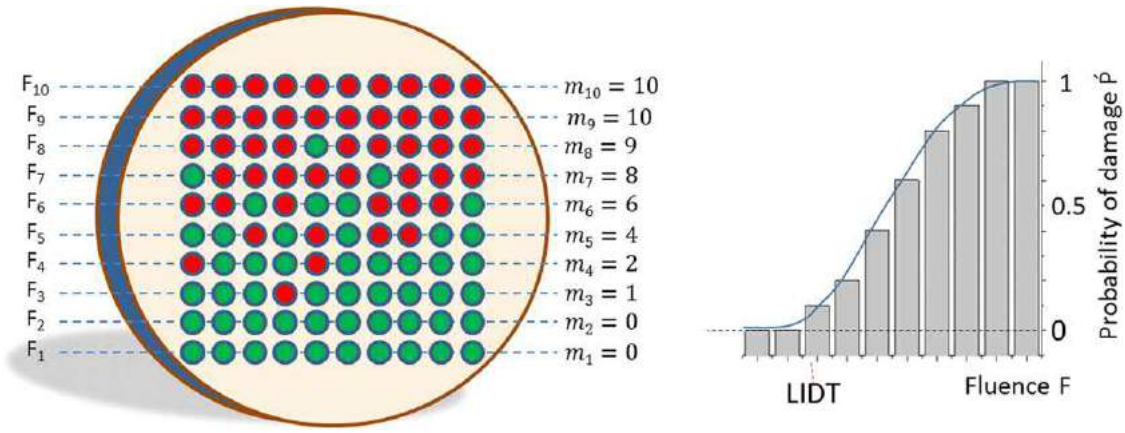


Fig. 7 Illustration of the ISO standard for LIDT tests. The damage probability versus fluence is measured. A tested optic (left) is divided into a number of tests sites and tested at various fluences F_i. The result of each test is either damage (red) or no damage (green). For each fluence F_i, the number of damage events m_i is counted. The probability is estimated by dividing the number of damage events by the number of tests at the given fluence. The probability vs. fluence (right) is plotted and LIDT is determined by extrapolating the curve to zero probability.

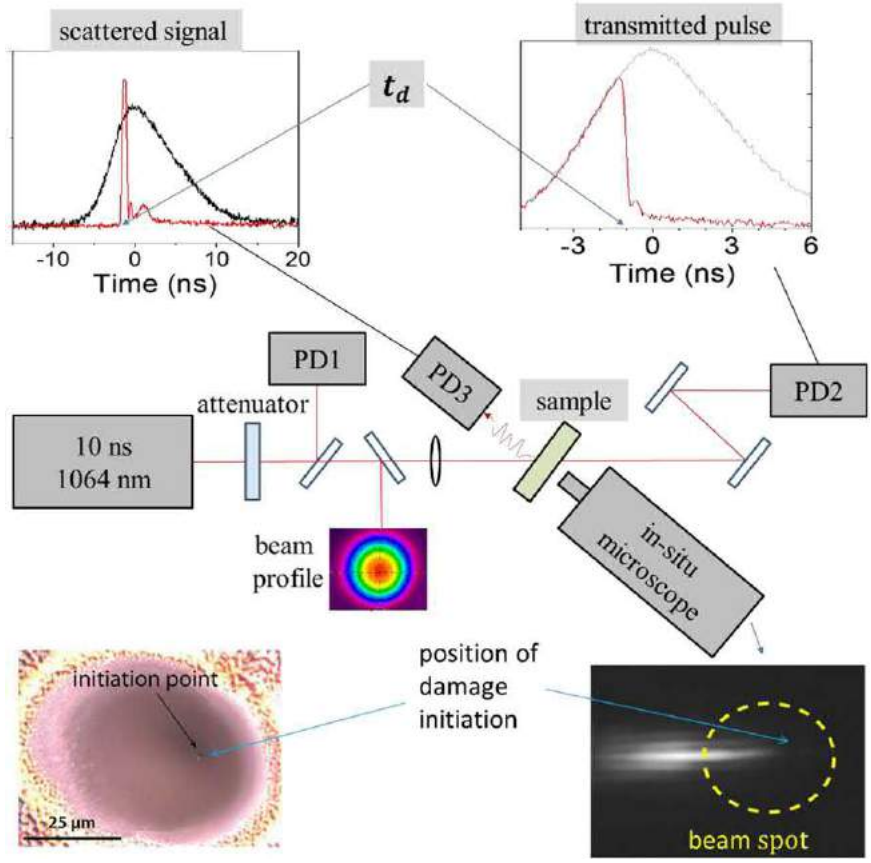


Fig. 8 Experimental setup for a nanosecond (or longer) pulse STEREO-LID test. The onset of damage and the location of damage initiation within the excitation spot are monitored for each event. Defect controlled damage leads to the ejection of a material jet as seen in the inset. Reproduced from Xu, Y., Emmert, L.A., Rudolph, W., 2015. Spatio-TEmporally REsolved Optical Laser Induced Damage (STEREO LID) technique for material characterization. Opt. Express 23, 21607–21614.

An early and intuitive methodology is the R-on-1 test. In this test, each site is tested repeatedly with increasing fluence until damage occurs. While still used occasionally, the R-on-1 method does not account for laser conditioning and incubation discussed previously and therefore the test depends on the details of how the fluence is increased.

The ISO LIDT standard (ISO 11254-2, 2011) is based on a measurement of damage probability (Porteus and Seitel, 1984). The method is illustrated in **Fig. 7**. The damage probability is plotted as a function of the incident fluence and the damage threshold F_{th} is determined by extrapolating the curve to zero probability. Strictly speaking, this extrapolation requires knowledge of the defect density distribution function. Since this distribution is rarely known, it is done typically with a linear fit. It is important that a site is tested just once even if no damage occurred, because the material might have changed. The shape of the damage probability curve provides qualitative information about the optic. For small beam sizes, a narrow fluence range between the 0% and 100% damage probability is an indication of intrinsic damage.

The ISO protocol provides a balance between practical aspects of component testing and obtaining material data for scientific studies. It is poor at quantifying sparse fluence limiting defects. This issue, however, is especially important for large area optics. For such applications, raster scan methods using large spots have been developed (Borden *et al.*, 2005). At the National Ignition Facility a 1-mm test beam is scanned over a 1-cm^2 area. The tests are overlapping in order to illuminate the entire square. The scans are repeated with increasing fluence, like the R-on-1 method, but this is considered to mimic the laser conditioning that would occur for the optic under realistic operation conditions. Finally, to reduce the effect of outliers, the LIDT is defined as the fluence at which 10 or more damage events are observed.

The previous methods all share the common feature that they test for damage or no damage and therefore set an upper bound on the true damage threshold. A method that measures the damage fluence/intensity during a pulse is Spatio-TEmporally REsolved Optical Laser-Induced Damage (STEREO-LID (Xu *et al.*, 2015a)) so named because it measures both the time during the pulse and the location within the beam spot where damage is initiated. Using the spatial/temporal intensity profile of the pulse, the critical fluence/intensity for each test event can be determined (**Fig. 8**).

Using STEREO-LID an optic is tested at multiple sites, like the ISO test, but all at an incident fluence sufficient to ensure a damage event. Each test event (one pulse at one sample site) determines where within the beam profile the damage initiating defect was located and when during the pulse "damage" occurred.

From these data one can derive the following: (1) the LIDT of performance-limiting defects: These are simply the lowest LIDT observed in the data set; (2) the defect distribution function $\overline{\rho}(F)$: this function can be retrieved without *a priori* assumptions by relating it to a histogram of the damage fluences of the data set (Xu *et al.*, 2015b): (3) the damage probability $P(F)$: This is done by a straightforward simulation of the ISO test using $\overline{\rho}(F)$. It should be noted that the same number of test sites provide not only new information about the sample compared to the ISO test, but also produce the ISO data with much greater accuracy.

See also: Ultrafast and Intense-Field Nonlinear Optics

References

Bercegol, H., 1998. What is laser conditioning? A review focused on dielectric multilayers. In: Proc. of SPIE, vol. 3578, pp. 421–426.

Beresna, M., Gecevicius, M., Kazansky, P.G., 2014. Ultrafast laser direct writing and nanostructuring in transparent materials. Adv. Opt. Photon 6. doi:10.1364/AOP.6.000293.

Bloembergen, N., 1974. Laser-induced electric breakdown in solids. IEEE J. Quant. Electron. 10, 375–386.

Borden, M.R., Folta, J.A., Stolz, C.J., *et al.*, 2005. Improved method for laser damage testing coated optics. In: Proc. of SPIE, vol. 5991, p. 59912A.

Demos, S.G., Negres, R.A., Raman, R.N., Rubenchik, A.M., Feit, M.D., 2013. Material response during nanosecond laser induced breakdown inside of the exit surface. Laser Photonics Rev. 7, 444–452.

DeShazer, L.G., Newnam, B.E., Leung, K.M., 1973. Role of coating defects in laser-induced damage to dielectric thin films. Appl. Phys. Lett. 23, 607–609.

Emmert, L.A., Mero, M., Rudolph, W., 2010. Modeling the effect of native and laser-induced states on the dielectric breakdown of wide band gap optical materials by multiple subpicosecond laser pulses. J. Appl. Phys. 108, 043523.

Emmert, L.A., Rudolph, W., 2015. Femtosecond laser-induced damage in dielectric materials. In: Ristau, D. (Ed.), Laser-Induced Damage in Optical Materials. Boca Raton, FL: CRC Press/Taylor & Francis Group, pp. 127–151.

Gallais, L., Douti, D.-B., Commandré, M., *et al.*, 2015. Wavelength dependence of femtosecond laser-induced damage threshold of optical materials. J. Appl. Phys. 117, 223103.

Gallais, L., Mangote, B., Commandré, M., *et al.*, 2010. Transient interference implications on the subpicosecond laser damage of multidielectrics. Appl. Phys. Lett. 97, 051112.

Gavartin, J.L., Ramo, D.M., Shluger, A.L., Bersuker, G., Lee, B.H., 2006. Negative oxygen vacancies in HfO_2 as charge traps in high-k stacks. Appl. Phys. Lett. 89, 082908.

ISO 11254-2:2011. Laser and laser related equipment – Test methods for laser-induced damage threshold – Part 2: Threshold determination. International Standard, 2011. International Organization for Standardization.

Jensen, L., Schrameyer, S., Jupe, M., Blaschke, H., Ristau, D., 2009. Spot-size dependence of the LIDT from the NIR to the UV. In: Proc. of SPIE, vol. 7504, p. 75041E.

Jones, S., Braunlich, P., Casper, R., Shen, X., Kelly, P., 1989. Recent progress on laser-induced modifications and intrinsic bulk damage of wide-gap optical materials. Opt. Eng. 28, 1039–1068.

Karras, C., Sun, Z., Nguyen, D.N., Emmert, L.A., Rudolph, W., 2011. The impact ionization coefficient in dielectric materials revisited. In: Proc. of SPIE, vol. 8190, p. 819028.

Keldysh, L., 1965. Ionization in the field of a strong electromagnetic wave. Sov. Phys. JETP 20, 1307–1314.

Lange, M.R., McIver, J.K., Guenther, A.H., 1987. Anomalous absorption in optical coatings. Nat. Bur. Stand. (U.S.) Spec. Publ. 746, 515–528.

Lenzner, M., Krüger, J., Sartania, S., *et al.*, 1998. Femtosecond optical breakdown in dielectrics. Phys. Rev. Lett. 80, 4076–4079.

Liu, J.M., 1982. Simple technique for measurements of pulsed Gaussian-beam spot sizes. Opt. Lett. 7, 196–198.

Mero, M., Liu, J., Rudolph, W., Ristau, D., Starke, K., 2005. Scaling laws of femtosecond laser pulse induced breakdown in oxide films. Phys. Rev. B 71, 115109.

Mouskeftaras, A., Guizard, S., Fedorov, N., Klimentov, S., 2013. Mechanisms of femtosecond laser ablation of dielectrics revealed by double pump–probe experiment. Appl. Phys. A 110, 709–715.
Papernov, S., 2015. Defect-induced damage. In: Ristau, D. (Ed.), Laser-Induced Damage in Optical Materials. Boca Raton, FL: CRC Press/Taylor & Francis Group, pp. 25–73.
Papernov, S., Schmid, A.W., 2002. Correlations between embedded single gold nanoparticles in SiO_2 thin film and nanoscale crater formation induced by pulsed-laser radiation. J. Appl. Phys. 92, 5720–5728.
Porteus, J.O., Seitel, S.C., 1984. Absolute onset of optical surface damage using distributed defect ensembles. Appl. Opt. 23, 3796–3805.
Rajeev, P.P., Gertsvolf, M., Corkum, P.B., Rayner, D.M., 2009. Field dependent avalanche ionization rates in dielectrics. Phys. Rev. Lett. 102, 083001.
Rethfeld, B., 2004. Unified model for the free-electron avalanche in laser-irradiated dielectrics. Phys. Rev. Lett. 92, 187401.
Ristau, D. (Ed.), 2015. Laser-Induced Damage in Optical Materials. Boca Raton, FL: CRC Press/Taylor & Francis Group.
Smith, A., Do, B., 2008. Bulk and surface laser damage of silica by picosecond and nanosecond pulses at 1064 nm. Appl. Opt. 47, 4812–4832.
Stolz, C.J., Ristau, D., Turowski, M., Blaschke, H., 2009.Thin film femtosecond laser damage competition. In: Proc. of SPIE, vol. 7504, p. 75040S.
Stolz, C.J., Runkel, J., 2012. Brewster angle polarizing beam splitter laser damage competition: p polarization. In: Proc. of SPIE, vol. 8530, p. 85300M.
Stolz, C.J., Thomas, M.D., Griffin, A.J., 2008. BDS thin film damage competition. In: Proc. of SPIE, vol. 7132, p. 71320C.
Stuart, B., Feit, M.D., Herman, S., *et al.*, 1996. Nanosecond-to-femtosecond laser-induced breakdown in dielectrics. Phys. Rev. B 53, 1749–1761.
Sun, Z., Lenzner, M., Rudolph, W., 2015. Generic incubation law for laser damage and ablation thresholds. J. Appl. Phys. 117, 073102.
Tench, R.J., Chow, R., Kozlowski, M.R., 1994. Characterization of defect geometries in multilayer optical coatings. J. Vac. Sci. Technol. A 12, 2808–2813.
Wood, R.M., 2003. Laser Damage in Optical Materials. Bristol: IOP Publishing/Taylor & Francis Group.
Xu, Y., Emmert, L.A., Rudolph, W., 2015a. Spatio-TEmporally REsolved Optical Laser Induced Damage (STEREO LID) technique for material characterization. Opt. Express 23, 21607–21614.
Xu, Y., Emmert, L.A., Rudolph, W., 2015b. Determination of defect densities from spatiotemporally resolved optical-laser induced damage measurements. Appl. Opt. 54, 6813–6817.

Four-Wave Mixing

L Canioni and L Sarger, University of Bordeaux, Talence, France

Introduction

Light waves propagate along straight lines in a vacuum and in materials. However, their amplitude and phase can only be affected by the intrinsic properties of the sample, such as absorption and refraction index. Experimental work from the early 1960s, has deviated from this simple approach, mainly through propagating intense light beams in crystals and glasses. Among many puzzling new effects, frequency conversion and beam coupling have triggered intense studies and opened a whole new field, theoretically and experimentally. These various phenomena have been classified according to the order of their optical nonlinearity. The second order mainly occurs in frequency generation and has already been discussed in previous articles of this encyclopedia. Therefore, this article will be devoted to the higher order, in which self focusing, coupling, and mixing of light beams will be explained. We will illustrate some simple ideas with a convenient model before describing a few applications.

The framework of electromagnetic wave interaction with matter is embedded in the Lorentz force. The oscillating electric field drives the components, ions, and electrons, but due to the large mass differences, only the electron movement is strongly perturbed. In dielectric media, these carriers are connected with static bonds such that the applied electric field modulates the spatial position of equilibrium of the charges. As a result, a dipolar moment is created and oscillates quasi-adiabatically at very high frequencies (between 10^{14} and 10^{17} Hz) for an applied optical wave. These collective oscillations of the induced dipoles compose the so-called matter polarization, proportional to the driving electric field amplitude. Magnetic field interaction is usually much weaker and so can be neglected in nonmagnetic media.

As an electromagnetic wave, lasers are a coherent source of very high intensity, i.e., large electric fields. For example, in standard pulsed lasers focused on a target, the light intensity can easily reach 10 Tw/cm^2, corresponding to an electric field of the order of magnitude of the internal bonding field (10^{10} V/m). At this level of interaction, dipolar moments are no longer strictly proportional to the applied electric field, as depicted in Fig. 1. In the weak field regime, the simple perturbation approach displaces the carriers in a power series of the applied electric field. The polarization is thus developed in series, in respect to this electric field.

Formalism

Relation Between Field and Polarization: The Response Function

Although numerous works have been published with a frequency description of the material polarization, we will discuss here the study of nonlinearity in the time domain in order to follow the physical reality. In most processes, both spaces are equal and can be selected at will, as long as the entire requirements are fully understood; the electromagnetic field is classically described.

In the time domain, the polarization can be developed as

$$P(r,t) = P_{\text{Linear}}(r,t) + P_{\text{Nonlinear}}(r,t) \tag{1}$$

where the linear part of the polarization can be written as

$$P_{\text{Linear}}(r,t) = \varepsilon_0 \int_{R^3} \int_{-\infty}^{\infty} R^{(1)}(r_1,t_1) E(r-r_1, t-t_1) \mathrm{d}t_1 \mathrm{d}r_1 \tag{2}$$

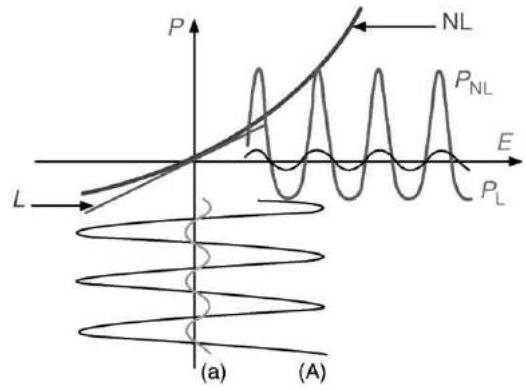


Fig. 1 Polarization for various field amplitude: (a) Weak field case, linear response. (A) Strong field, distortion and nonlinearity.

 doi:10.1016/B978-0-12-809283-5.00755-2

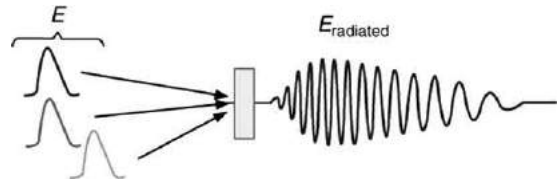


Fig. 2 Three incoming beams E_1, E_2, E_3 create a polarization (nonlinear). The sample responds while radiating a fourth field E_4.

This general expression is rather complex but shows that the polarization, due to collective movement of carriers in a macroscopic volume (thousands of dipoles) centered at position r, always depends on the applied electric field. This expression is also nonlocal as dipoles–dipoles interactions play an important role.

In general, for the nonlinear (NL) part of the polarization, an approximation of the local interaction is described for simplicity. For a second-order interaction, one can write:

$$P_{\mathrm{NL}}^{(2)}(r,t) = \varepsilon_0 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} R^{(2)}(t_1,t_2)E(r,t-t_2)E(r,t-t_2-t_1)\mathrm{d}t_1\mathrm{d}t_2 \tag{3}$$

For the other series expansion – at third order, one obtains

$$P_{\mathrm{NL}}^{(3)}(r,t) = \varepsilon_0 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} R^{(3)}(t_1,t_2,t_3)E(r,t-t_3)E(r,t-t_3-t_2)E(r,t-t_3-t_2-t_1)\mathrm{d}t_1\mathrm{d}t_2\mathrm{d}t_3 \tag{4}$$

This third-order polarization already combines three electric fields. This nonlinear polarization also radiates a fourth wave and is therefore the adopted framework for describing four-wave mixing experiments (**Fig. 2**).

As the physical properties of materials depend only on time intervals between the different pulse interactions, **Eq. (4)** can be rearranged as

$$P_{\mathrm{NL}}^{(3)}(r,t) = \varepsilon_0 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} R^{(3)}(t-\tau_1,t-\tau_2,t-\tau_3)E(r,\tau_1)E(r,\tau_2)E(r,\tau_3)\mathrm{d}\tau_1\mathrm{d}\tau_2\mathrm{d}\tau_3 \tag{5}$$

If the sample response time is much shorter than pulse duration, the so-called adiabatic limit (as for electronic cloud deformation excited distant from resonance) the polarization is written as

$$P_{\mathrm{NL}}^{(3)}(r,t) = \varepsilon_0 \sigma^{(3)} E(r,t)E(r,t)E(r,t) \tag{6}$$

We will see later how this expression is useful for third-harmonic generation and optical Kerr effects, etc. This expression is accurate for materials excited at a distance from resonances and care must be taken to reserve such a model for corresponding processes. In this case, the frequency approach is quite simple too and both spaces, time or Fourier, can be chosen at will.

For very long response time, such as orientational or structural molecular reorganization, one can use the Born–Oppenheimer approximation and present the nonlinear polarization as

$$P_{\mathrm{NL}}^{(3)}(r,t) = \varepsilon_0 E(r,t)R'(t) \tag{7}$$

Similar is linear polarization, where the response function takes into account the changes of the molecular structure under the light intensity stress:

$$R'(t) = \int_{-\infty}^{+\infty} d^{(3)}(t-t_1)E(r,t_1)E(r,t_1)\mathrm{d}t_1 \tag{8}$$

The material response will then be mostly driven by the field amplitude rather than the frequency.

Properties of Nonlinear Susceptibilities

The third-order nonlinear susceptibility $\chi^{(3)}$ is the Fourier transform of the response function R and has proved to be a most practical tool in optical nonlinear theory. The symmetry properties of the material lead to simplification on the tensor of rank 4 (fourth rank tensor), involved in four-wave mixing:

$$P_i^{(3)}(\omega_m) = \varepsilon_0 \sum_{n_1,\ldots,n_p} \sum_{i_1,\ldots,i_p} \chi_{i,i_1,\ldots,i_p}^{(3)}\left(\omega_m;\omega_{n_1},\ldots,\omega_{n_p}\right)E_{i_1}(\omega_{n_1})\ldots E_{i_p}\left(\omega_{n_p}\right) \tag{9}$$

The nonlinear susceptibility tensor must account for all the spatial and symmetrical order of the sample. This will eventually reduce the number of independent elements from a possible 81.

Wave Equation in Nonlinear Regime

From Maxwell's equations, with a sample without free carriers and current, one can deduce the wave equation in the time domain, as:

$$\Delta E - \frac{1}{c^2}\frac{\partial^2 E}{\partial t^2} - \frac{1}{c^2}\frac{\partial^2 P_{\mathrm{L}}}{\partial t^2} = \frac{1}{c^2}\frac{\partial^2 P_{\mathrm{NL}}}{\partial t^2} \tag{10}$$

In the frequency domain, this equation reduces to

$$\Delta\tilde{E}(r,\omega) + k(\omega)^2\tilde{E}(r,\omega) = -\mu_0\omega^2\tilde{P}_{\mathrm{NL}}(r,\omega) \tag{11}$$

In order to solve these (coupled) wave equations in relation to four-wave mixing, we must develop the electric field and the polarization in a slowly varying envelope approximation, using the following conventions:

$$E_i(r,t) = \frac{1}{2}\left(A_i(r,t)\mathrm{e}^{\mathrm{i}(\omega_i t - k(\omega_i)z)} + \overline{A}_i(r,t)\mathrm{e}^{-\mathrm{i}(\omega_i t - k(\omega_i)z)}\right) \tag{12}$$

$$P_{\mathrm{NL}}(r,t) = \frac{1}{2}\left(p_{\mathrm{nl}}(r,t)\mathrm{e}^{\mathrm{i}(\omega t - k(\omega)z)} + \overline{p}_{\mathrm{nl}}(r,t)\mathrm{e}^{-\mathrm{i}(\omega t - k(\omega)z)}\right) \tag{13}$$

Within the framework of small perturbations, diffraction can be neglected and the wave equation can be easily reduced to one dimension:

$$\frac{\partial A}{\partial z} = \frac{\mathrm{i}\omega}{2n(\omega)\varepsilon\mathrm{o}c}p^{\mathrm{NL}}(z,\omega)\exp\left(-\mathrm{i}k(\omega)z\right) \tag{14}$$

where ω represents the sum or difference permutations over all the frequencies ω_{i}, thus reflecting the law of energy conservation.

In the case of four-wave mixing, one has to solve four coupled equations for the four interacting waves. Using the polarization Eq. (13), one gets:

$$\frac{\partial A_j}{\partial z} = \frac{\mathrm{i}\omega_j}{2n(\omega_j)c}\chi^3 A_1^{(*)}A_2^{(*)}A_3^{(*)}\exp\left(\sum_{i=1,2,3}\exp\left(\pm\mathrm{i}k_i(\omega_i)z\right)\right)\exp\left(-\mathrm{i}k(\omega_j)z\right) \tag{15}$$

This set of generic equations respectively represent indifferently, either the three incoming beams ($j = 1, 2, 3$), or the radiated field E_4 at ω. The algebra used here associates at each complex conjugate field (*) with the counter propagating wave ($-k$). The amplitude variation of the j field at frequency ω_j is due to the coupling of the three other fields within the sample. The accumulated field is nevertheless always strongly dependent of the phase term and will vanish except when this phase term is zero. This is known as the phase matching condition which has to be fulfilled to ensure an efficient energy conversion. A more general approach of this condition can be found in the corresponding article of this encyclopedia.

Example and Selected Applications

When inserting the full material polarization (linear and nonlinear) into the propagation equation, even restricted at third order, a whole set of processes must be considered. All of them are described by the generic equation already presented, but only a few very important processes – with increasing complexity – will be described here.

Optical Kerr Effect

We consider one single beam incident at frequency ω on a sample. The non-linear polarization is given by

$$P_{\mathrm{NL}}^{(3)}(r,t) = \frac{\varepsilon_0}{8}\chi^{(3)}\left(E(r,t)^3 + E^*(r,t)^3 + 3E(r,t)^2E^*(r,t) + 3E(r,t)E^*(r,t)^2\right) \tag{16}$$

The polarization at frequency ω reduces to

$$P_{\mathrm{NL}}^{(3)}(r,t) = \frac{3\varepsilon_0}{4}\chi^{(3)}|E(r,t)|^2E(r,t) \tag{17}$$

This term has the same phase as the incident beam and, therefore, the phase matching is automatic. Using Eqs. (12) and (13), one can rewrite the wave equation for the optical Kerr effect in the thin sample approximation:

$$\frac{\partial A}{\partial z} = \frac{3\mathrm{i}\omega}{8nc}\chi^{(3)}|A|^2A \tag{18}$$

The solution of this equation is

$$A(z) = A(0)\exp\left(\frac{3\mathrm{i}}{8nc}\chi^{(3)}|A(0)|^2z\right) \tag{19}$$

Along the beam, the amplitude of the input pulse stays constant for this process but its phase is affected by a nonlinear shift proportional to power density I.

In literature, one can find the concept of n_2, the sample nonlinear index, in the following formula:

$$\psi_{\mathrm{NL}}(z) = \frac{\omega}{c} n_2 I z \tag{20}$$

In this way of describing the total phase along the beam, the index of refraction in the nonlinear regime can be expressed more conveniently as

$$n = n_0 + n_2 I \tag{21}$$

One of the most significant demonstrations of this effect is the self-focusing phenomenon. The usual Gaussian transverse structure of a laser beam presents high intensity at the center of the beam, therefore creating an index variation in the isotropic sample due to this optical Kerr effect. An induced lens can then be applied to this refractive index gradient, following the field distribution. As a result, the beam focuses itself along the propagation and can result in a complete breakdown.

Third Harmonic Generation

In the third harmonic generation, i.e., the third-order nonlinear process where the frequency of the generated wave is three times the frequency of the input wave at ω, only two waves are in the sample (at ω and 3ω). For the nonlinear wave, the master equation is then:

$$\frac{\partial A_{3\omega}}{\partial z} = \frac{3\mathrm{i}\omega}{8n(3\omega)c}\chi^{(3)}A_{\omega}^{3}\exp((3\mathrm{i}k(\omega) - \mathrm{i}k(3\omega))z) \tag{22}$$

Here, propagation effects have to be carefully considered as theses waves do travel at different velocities and will eventually be out of phase. Obviously this discrepancy of phase velocities precludes the co-propagation and strongly affects the global efficiency. The only way to ensure a coherent generation of the harmonic wave throughout the material is to fulfill the relation:

$$k(3\omega) = 3k(\omega) \tag{23}$$

This is the so-called phase matching condition. Although difficult because the usual dispersion of materials is very large for such a frequency difference, this can be satisfied using either geometrical arrangement or, if possible, using the birefringent properties of the material. The efficiency of this method is poor and a better approach is implemented when two $\chi^{(2)}$ processes are cascaded (doubling and sum frequency mixing).

Degenerate Four-Wave Mixing and Phase Conjugation

One of the most popular and intriguing interactions is depicted in Fig. 3. Where the sample is excited by three fields at the same frequency ω in this particular geometry: two-waves, named pump beams E_1 and E_2, are exactly counter-propagating with opposite wave vectors of $+k_1$ and $k_2 = -k_1$, while the third wave (E_3 arrives at the sample with a given wave vector of k_3.

Of possible couplings, one contribution E_4 corresponds to a re-emitted beam along $k_4 = -k_3$, i.e., opposite to the E_3 wave.

The third-order nonlinear polarization is given here by

$$P_{\mathrm{NL}}^{(3)}(r,t) = \frac{\varepsilon_0}{8}\chi^{(3)}\left(E_1(r,t) + E_1^*(r,t) + E_2(r,t) + E_2^*(r,t) + E_3(r,t) + E_3^*(r,t)\right)^3 \tag{24}$$

This E_4 wave (along $-k_3$) originates in this product from source terms including:

$$E_1(r,t)E_2(r,t)E_3^*(r,t) \tag{25}$$

and thus present an evolution driven by:

$$\frac{\partial A_4}{\partial z} = \frac{3\mathrm{i}\omega}{4n(\omega)c}\chi^{(3)}A_1A_2A_3^* \tag{26}$$

where the phase matching condition is automatically fulfilled. Interesting facts can be described using this expression. First, the radiated amplitude is proportional to the complex conjugate of the E_3 field. This nonlinear process is not only able to reverse the direction of propagation but also the whole phase of an arbitrary incoming beam of light. This 'phase conjugator' can be

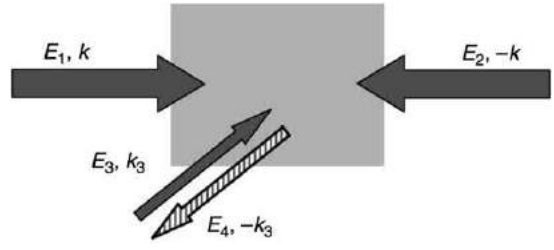


Fig. 3 Two counter-propagating fields (E_1, E_2) and the third field E_3 create a polarization (nonlinear). The sample responds while radiating a fourth field E_4 opposite to the E_3 wave.

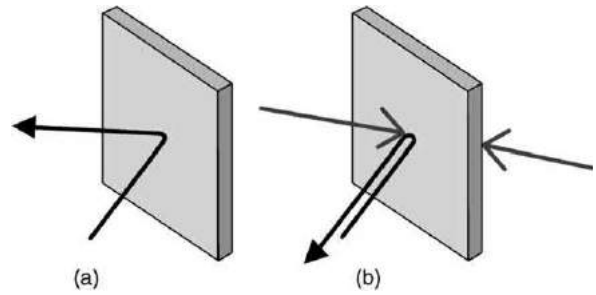


Fig. 4 (a) Snell refraction in a conventional mirror; (b) Phase conjugate mirror.

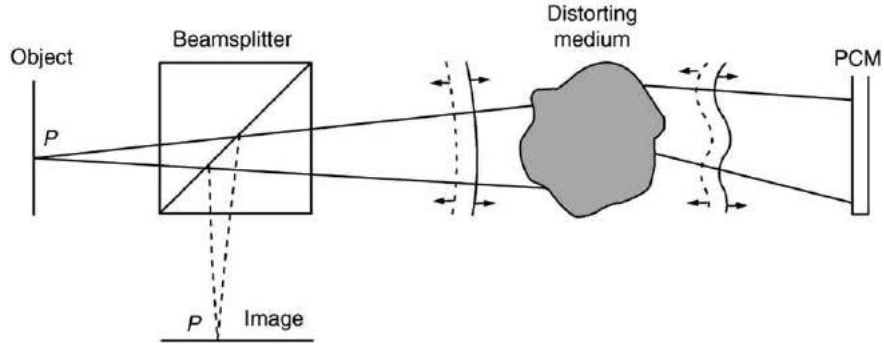


Fig. 5 A basic two-pass system for imaging and distortion compensation using optical phase conjugation.

considered as a kind of mirror with very unusual reflection properties. Unlike a conventional mirror, where a ray is redirected according to the ordinary law of reflection, a phase-conjugate mirror (PCM) retro-reflects all incoming rays back to their origin (Fig. 4).

Second, the reflectivity of this mirror depends on the susceptibility and the amplitude of the first two fields, usually called pump beams. This will allow reflectivity that is much higher than the unity, unlike with conventional mirrors.

The remarkable reflection properties of the PCM have found many important applications. The most useful undoubtedly is related to distortion correction. If the image information has been distorted by the transmitting medium on its way to the PCM, then these aberrations will be corrected when the reflected signal retraces its original path through the medium. Through this amazing 'healing' property of optical phase conjugation, a high-quality optical beam can be double passed through a distorting medium, such as an imperfect imaging device or a turbulent atmosphere, without any loss of beam quality. Fig. 5 shows the basic two-pass geometry for imaging and aberration correction using a PCM.

The spherical wave from the object point P is aberrated by the distorting medium. The wave-front is reversed by the PCM and a second passage through the same distorting path restores the initial field. The resulting spherical wave converges to the image point P', separated from P by the beamsplitter.

Numerous other applications for PCM and the various underlying nonlinear optical interactions have been proposed. These include effects as diverse as parametric oscillation, optical tracking and pointing, spatial and temporal image processing, optical computing, optical filtering, etc. The PCM is also successfully implemented for aberration correction in high-power laser resonators.

Nondegenerate Four-Wave Mixing

If the frequencies of the impinging fields build a nonlinear polarization with an oscillating term close to absorption (or emission) frequencies of the sample, the nonlinear susceptibility increases tremendously. The radiated field amplitude changes by orders of magnitude and allows spectroscopic studies of the sample. The most popular process used to perform gas spectroscopy is CARS, an acronym for coherent anti-Stokes Raman scattering (or spectroscopy). In this case, the three incoming beams have different frequencies and the phase matching condition is eventually fulfilled in the 3D space.

Usually, two beams at the same frequency ω_P (pump beams), are mixed with a third beam at a tunable frequency ω_S, according to the energy scheme in Fig. 6. As the radiated frequency ω_{CARS} is higher, the Raman spectroscopic notation is used and the subscripts 'S' hold for 'Stokes' and 'P' for 'anti-Stokes'.

Whenever $\omega_P - \omega_S = \omega_{molecule}$, then an increase of the CARS intensity by many orders of magnitude is observed. Therefore, a Raman spectrum can be obtained by continuously changing, the Stokes laser and simultaneously recording the intensity of the

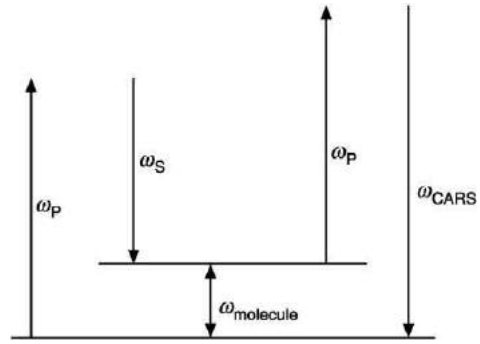


Fig. 6 Energy diagram for a CARS experiment. Four-wave mixing efficiency increases when the frequency difference $\omega_P - \omega_S$ is close to a resonant frequency of the sample, $\omega_{molecule}$.

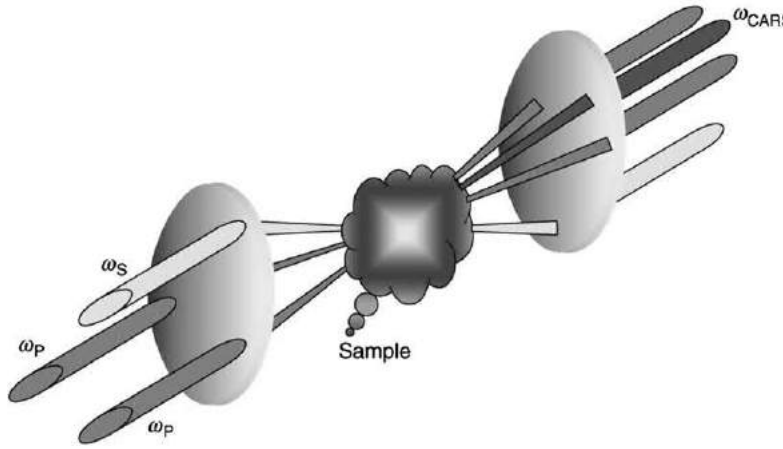


Fig. 7 Layout of the CARS beams near a sample (a flame here). This particular arrangement is called folded (or 3D) Boxcars.

CARS beam. This procedure is called scanning CARS, an interesting technique insofar as it allows taking high-resolution spectra, as well as measuring temperature of the sample while scanning over the Boltzmann distribution of the electronic ground state.

Fig. 7 displays a classical geometrical arrangement where the phase matching condition fully determines the directionality of the radiated field at ω_{CARS}. Although the alignment of the setup can become tedious, this technique has been widely used, even in hostile industrial environments. The coherent aspect of this four-wave mixing process, imbedded in the phase matching condition, has a huge benefit over classical spectroscopic techniques and thus isotropic techniques, such as laser-induced fluorescence or Raman spectroscopy.

The CARS process is an interference process comparable to diffraction of a grating. The two fields at ω_P and ω_S form a laser-induced moving grating and the third field, at ω_P, undergoes a Doppler shift at ω_{CARS}. This scattered field undergoes a coherent amplitude addition in the Bragg direction.

Conclusion

In this article, we have discussed a set of optical nonlinear processes occurring at the third order of the development in series of the material polarization. At this order, wave mixing obviously shows most significant behavior and has been extensively studied and applied in many scientific areas. The general framework presented here can be further investigated using some books listed in Further Reading below.

See also: Nonlinear Optical Phase Conjugation

Further Reading

Bloembergen, N., 1996. Nonlinear Optics. Singapore: World Scientific Pub Co.

Bloom, D.M., Bjorklund, G.C., 1977. Conjugate wave-front generation and image reconstruction by four-wave mixing. Applied Physics Letters 31, 592–594.
Born, M., Wolf, E., 1999. Principles of Optics. Cambridge, UK: Cambridge University Press.
Butcher, P.N., Cotter, D., 1991. The Elements of Nonlinear Optics. Cambridge, UK: Cambridge University Press.
Eaton, D.F., 1991. Nonlinear optical materials. Science 253, 281–287.
Fisher, R.A., 1984. Optical Phase Conjugation. San Diego, CA: Academic.
Flytzanis, C., 1975. Theory of nonlinear optical susceptibilities. In: Rabin, H., Tang, C.L. (Eds.), Quantum Electronics, vol. 1. New York: Academic Press, pp. 9–207.
Hellwarth, R.W., 1977. Generation of time-reversed wave fronts by nonlinear refraction. Journal of the Optical Society of America 67, 1–3.
Jackson, J.D., 1998. Classical Electrodynamics. New York: Wiley.
Marburger, J.H., 1975. Self-focusing: Theory. Progress in Quantum Electronics 4, 35–110.
Mittra, R., Habashy, T.M., 1984. Theory of wave-front-distortion correction by phase conjugation. Journal of the Optical Society of America A 1, 1103–1109.
Rockwell, D.A., 1988. A review of phase-conjugate solid-state lasers. IEEE Journal of Quantum Electronics 24, 1124–1140.
Shen, Y.R., 2002. The Principles of Nonlinear Optics. New York: Wiley.
Yariv, A., 1977. Compensation for atmospheric degradation of optical beam transmission. Optical Communication 21, 49–50.
Yariv, A., 1989. Quantum Electronics, 3rd edn. New York: Wiley.
Yariv, A., Pepper, D.M., 1977. Amplified reflection, phase conjugation, and oscillation in degenerate four-wave mixing. Optical Letters 1, 16–18.
Zel'dovich, B.Ya., Pilipetsky, N.F., Shkunov, V.V., 1985. Principles of Phase Conjugation. Berlin: Springer-Verlag.

Kramers–Krönig Relations in Nonlinear Optics

M Sheik-Bahae, The University of New Mexico, Albuquerque, NM, USA

Nomenclature

α	linear absorption coefficient	n_2	nonlinear refractive index coefficient, coefficient of optical Kerr effect
α_2	nonlinear absorption coefficient	$\Theta(t)$	step function
β	two-photon absorption coefficient	$\wp$	principal value
$\chi^{(n)}$	nth-order nonlinear optical susceptibility	**2PA**	two-photon absorption
		KK relations	Kramers–Krönig relations
n	linear refractive index		

Since their introduction nearly 75 years ago, the Kramers–Krönig (KK) dispersion relations have been widely appreciated and applied in the analysis of linear optical systems. Because they are a consequence of strict causality, the KK relations apply not only to optical systems, but also to any linear, causal system such as electrical networks and particle scattering. In this article, we review the formulation and application of these relations in nonlinear optical systems. Simple logical arguments are used to derive dispersion relations that relate the nonlinear absorption coefficient to the nonlinear refraction coefficient. More general formalisms are then derived that apply to all nonlinear susceptibilities including the harmonic generating cases. Examples of recent successful application of these dispersion relations in analyzing various nonlinear materials will be presented.

The mathematical formalism of the KK dispersion relations in nonlinear optics was studied in the formative days of the field. The great usefulness of these relations was appreciated only recently, however, when they were used to derive the dispersion of the optical Kerr effect in solids from the corresponding nonlinear absorption coefficients, including two-photon absorption.

Before examining the details of KK relations in nonlinear optical systems, it is instructive to revisit the linear dispersion relations and their derivation based on the logic of causality. We will begin this task by introducing the definition of the linear as well as nonlinear susceptibilities $\chi^{(n)}$. In most nonlinear optics texts, the total material polarization (P) that drives the wave equation for the electric field (E) is expressed as

$$\begin{aligned} P_i(t) = \varepsilon_0 \int_{-\infty}^{\infty} R_{ij}^{(1)}(t-t_1)E_j(t_1)\mathrm{d}t_1 + \varepsilon_0 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} R_{ijk}^{(2)}(t-t_1,t-t_2)E_j(t_1)E_k(t_2)\mathrm{d}t_1\mathrm{d}t_2 \\ +\varepsilon_0 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} R_{ijkl}^{(3)}(t-t_1,t-t_2,t-t_3)E_j(t_1)E_k(t_2)E_l(t_3)\mathrm{d}t_1\mathrm{d}t_2\mathrm{d}t_3 + \ldots \end{aligned} \tag{1}$$

where $R^{(n)}$ is defined as the nth-order, time-dependent response function or time-dependent susceptibility. The subscripts are polarization indices indicating, in general, the tensor nature of the interactions. The summation over the various indices j, k, l, … is implied for the various tensor elements of $R^{(n)}$. Upon Fourier transformation, we obtain:

$$\begin{aligned} \vec{P}_i(\omega) = \varepsilon_0 \int_{\infty}^{-\infty} \mathrm{d}\omega_1 \chi_{ij}^{(1)}(\omega_1)\vec{E}_j(\omega_1)\delta(\omega-\omega_1) + \varepsilon_0 \int_{\infty}^{-\infty}\mathrm{d}\omega_1 \int_{\infty}^{-\infty} \mathrm{d}\omega_2 \chi_{ijk}^{(2)}(\omega_1,\omega_2)\vec{E}_j(\omega_1)\vec{E}_k(\omega_2)\delta(\omega-\omega_1-\omega_2) \\ +\varepsilon_0 \int_{\infty}^{-\infty}\mathrm{d}\omega_1 \int_{\infty}^{-\infty}\mathrm{d}\omega_2 \int_{\infty}^{-\infty} \mathrm{d}\omega_3 \chi_{ijkl}^{(3)}(\omega_1,\omega_2,\omega_3)\vec{E}_j(\omega_1)\vec{E}_k(\omega_2)\vec{E}_l(\omega_3)\delta(\omega-\omega_1-\omega_2-\omega_3),\ldots \end{aligned} \tag{2}$$

where δ is the Dirac delta-function. Here the $E(\omega)$ are Fourier transforms of the corresponding electric field. The nth-order susceptibility is defined as the Fourier transform of the nth-order response function:

$$\chi_{ijk\ldots n}^{(n)}(\omega_1,\omega_2,\ldots,\omega_m) = \int_{-\infty}^{+\infty}\mathrm{d}\tau_1 \int_{-\infty}^{+\infty}\mathrm{d}\tau_2 \ldots \int_{-\infty}^{+\infty}\mathrm{d}\tau_n R_{ijk\ldots m}^{(n)}(\tau_1,\tau_2,\ldots,\tau_m)\mathrm{e}^{\mathrm{i}(\omega_1\tau_1+\omega_2\tau_2+\ldots+\omega_m\tau_m)} \tag{3}$$

For simplicity, we drop the polarization indices i, j, … , and thus ignore the tensor properties of $\chi^{(n)}$ as well as the vector nature of the electric fields.

Let us for the moment concentrate on the linear polarization alone and derive the linear KK relations for the first-order susceptibility $\chi^{(1)}(\omega)$. For this, we rewrite **Eq. (3)** for $n=1$:

$$\chi^{(1)}(\omega) = \int_{-\infty}^{\infty} R^{(1)}(\tau)\mathrm{e}^{-\mathrm{i}\omega\tau}\mathrm{d}\tau \tag{4}$$

(As defined above, $\chi^{(1)}(\omega)$ and $R^{(1)}(\tau)$ are not a strict Fourier transform pair because of a missing factor of 2π.) Causality means that the effect cannot precede the cause. This can be restated mathematically as:

$$R^{(1)}(t) = R^{(1)}(t)\Theta(t) \tag{5}$$

i.e., the response to an impulse at $t=0$ must be zero for $t<0$. Here $\Theta(t)$ is the Heaviside step function defined as $\Theta(t)=1$ for $t>0$ and $\Theta(t)=0$ for $t<0$. Upon Fourier transforming this equation, the product in the time domain becomes a convolution integral in

frequency space

$$\begin{aligned}\chi^{(1)}(\omega) &= \chi^{(1)}(\omega)\left[\frac{\delta(\omega)}{2} + \frac{\mathrm{i}}{2\pi\omega}\right] \\ &= \frac{\chi^{(1)}(\omega)}{2} + \frac{\mathrm{i}}{2\pi}\wp\int_{-\infty}^{\infty}\frac{\chi^{(1)}(\omega')}{\omega-\omega'}\mathrm{d}\omega' \\ &= \frac{1}{\mathrm{i}\pi}\wp\int_{-\infty}^{\infty}\frac{\chi^{(1)}(\omega')}{\omega'-\omega}\mathrm{d}\omega'\end{aligned} \tag{6}$$

which is the KK relation for the linear optical susceptibility. The symbol $\wp$ stands for the Cauchy principal value of the integral. The KK relation is thus a restatement of the causality condition **(5)** in the frequency domain. Taking the real part we have,

$$\Re e\left\{\chi^{(1)}(\omega)\right\} = \frac{1}{\pi}\wp\int_{-\infty}^{\infty}\frac{\Im m\{\chi^{(1)}(\omega')\}}{\omega'-\omega}\mathrm{d}\omega' \tag{7}$$

Taking the imaginary part of **Eq. (6)** leads to a similar relation relating the imaginary part to an integral involving the real part. It is conventional to write the optical dispersion relations in terms of the more familiar quantities of refractive index, $n(\omega)$, and absorption coefficient, α (ω). For $|\chi^{(1)}| \ll 1$ then $n-1 = \Re e\{\chi^{(1)}\}/2$ and $\alpha = \omega\Im m\{\chi^{(1)}\}/c$, and **Eq. (7)** is transformed into

$$n(\omega) - 1 = \frac{c}{\pi}\wp\int_{0}^{\infty}\frac{\alpha(\omega')}{\omega' 2 - \omega^2}\mathrm{d}\omega' \tag{8}$$

where we additionally used the reality conditions of $n(\omega) = n(-\omega)$, and $\alpha(\omega) = \alpha(-\omega)$ to change the lower integral limit to 0. More rigorous analysis shows that **Eq. (8)** is general and valid for any value of $|\chi^{(1)}|$. Although the KK dispersion relations and the extent of their applications in linear optics are well understood, some confusion sometimes exists about their applications to nonlinear optics. Causality clearly holds for both linear and nonlinear systems. The question is: what form do the resulting dispersion relations take in a nonlinear system? The linear Kramers–Krönig relations were derived from linear system theory, so it would appear to be impossible to apply the same logic to a nonlinear system. The key insight is that one can linearize the system. This is illustrated in **Fig. 1** where a linear (and of course, causal) optical material is transformed into a 'new' linear system that now contains the material and an external perturbation denoted by ξ. Although we are interested in perturbations of an optical nature, this formalism is general under any type of perturbation. It is important to appreciate the fact that our new system is causal even in the presence of the perturbation. This allows us to write down a modified form of the Kramers–Krönig relation linking the index of refraction to the absorption:

$$[n(\omega) + \Delta n(\omega;\zeta)] - 1 = \frac{c}{\pi}\wp\int_{0}^{\infty}\frac{\alpha(\omega') + \Delta\alpha(\omega';\zeta)}{\omega' 2 - \omega^2}\mathrm{d}\omega' \tag{9}$$

which, after subtracting the linear relation between n and α leaves a relation between the changes in index and absorption:

$$\Delta n(\omega;\zeta) = \frac{c}{\pi}\wp\int_{0}^{\infty}\frac{\Delta\alpha(\omega';\zeta)}{\omega' 2 - \omega^2}\mathrm{d}\omega' \tag{10}$$

where ζ denotes a general perturbation. An equivalent relation also exists whereby the change in absorption coefficient can be calculated from the change in the refractive index. It is essential that the perturbation be independent of frequency of observation, ω', in the integral (i.e., the excitation ζ must be held constant as ω' is varied).

Eq. (10) has been used to determine refractive changes due to 'real' excitations such as thermal and free-carrier nonlinearities in semiconductors. In those cases, ζ denotes either ΔT (change of temperature) or ΔN (change of free-carrier density), respectively. In the former case, one calculates the refractive index change resulting from a thermally excited electron–hole plasma and the temperature shift of the band edge. For cases where an electron–hole plasma is injected (e.g., optically), the change of absorption gives the plasma contribution to the refractive index. In this case, the ζ parameter in **Eq. (10)** is taken as the change in plasma density regardless of the mechanism of generation or the optical frequency.

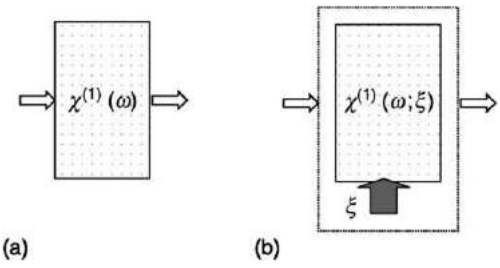


Fig. 1 (a) A causal linear system obeying KK relations. (b) The system in (a) when externally perturbed by ξ. The dotted box now represents our new linear causal system whose altered $\chi^{(1)}$ obeys the KK relations.

Let us now extend this formalism to the case where the perturbation is virtual, occurring at an excitation frequency Ω that is below any material resonance. To the lowest order in the excitation irradiance I_Ω , we write

$$\Delta\alpha(\omega;\zeta) = \Delta\alpha(\omega;\Omega) = 2\alpha_2(\omega;\Omega)I_\Omega \tag{11}$$

and

$$\Delta n(\omega;\zeta) = \Delta n(\omega;\Omega) = 2n_2(\omega;\Omega)I_\Omega \tag{12}$$

where n_2 and α_2 are the nonlinear refractive index and absorption coefficients of the material, respectively. By definition, these coefficients are related to the third-order nonlinear susceptibility $\chi^{(3)}(\omega_1,\omega_2,\omega_3)$ via

$$n_2(\omega;\Omega) = \frac{3}{4\varepsilon_0 n_0(\omega)n_0(\Omega)c}\Re e\left\{\chi^{(3)}(\omega,-\Omega,\Omega)\right\} \tag{13}$$

and

$$\alpha_2(\omega;\Omega) = \frac{3\omega_a}{2\varepsilon_0 n_0(\omega)n_0(\Omega)c^2}\Im m\left\{\chi^{(3)}(\omega,-\Omega,\Omega)\right\} \tag{14}$$

We can therefore write the dispersion relations between α_2 and n_2:

$$n_2(\omega;\Omega) = \frac{c}{\pi}\wp\int_0^\infty \frac{\alpha_2(\omega';\Omega)}{\omega' 2-\Omega^2}\mathrm{d}\omega' \tag{15}$$

Note that even when the degenerate $n_2(\omega) = n_2(\omega;\omega)$ is desired (at a given ω), the dispersion relation requires that we should know the nondegenerate absorption spectrum $\alpha_2(\omega';\omega)$ at all frequencies ω'.

Let us pause here and discuss some physical mechanisms that can be involved for a given system of interest. Consider a material characterized by an optical resonance occurring at, say ω_0 (i.e., a degenerate two-level system). For a solid, this resonance can be regarded as that of the fundamental energy gap; $\omega_0 = \omega_g = E_g/\hbar$ in a two-band system. Now, let us examine how the presence of an optical excitation at $\Omega<\omega_0$ can alter the absorption spectrum (at a variable probe ω'). In the quantum mechanical picture, this gives rise to a 'new' material whose perturbed wave functions are 'dressed' by the intensity and frequency of the applied optical field. The lowest-order correction to the absorption is given by $\alpha_2(\omega';\Omega)$ which involves three major physical processes. Recalling that $\Omega<\omega_0$, these processes include (1) two-photon absorption (2PA) when $\omega' + \Omega \to \omega_0$ and (2) Raman-induced absorption when $\omega' - \Omega \to \omega_0$, both implying an absorption of a photon at the probe frequency ω' (i.e., $\alpha_2 > 0$). The third process can be identified as resulting from the blue-shift (for $\Omega<\omega_0$) of the resonance (known as the quadratic optical Stark effect) caused by the excitation field. For our two-level system, the latter results in a decrease followed by an increase in absorption in the vicinity of ω_0. An example of the overall absorption changes due to such processes is shown in **Fig. 2** where $\alpha_2(\omega';\Omega)$ is qualitatively plotted for a degenerate two-level system. We should note that the relative magnitude of each contribution as well as the width and shape of the resonances are chosen arbitrarily for the purpose of illustration. Using the KK relation in **Eq. (15)**, we can now arrive at the nonlinear index coefficient $n_2(\omega;\Omega)$. The result of this transformation is also given in **Fig. 2**. The above simple example elucidates the key concepts involving the relationship between nonlinear absorption and refraction in materials for third-order processes. These concepts, when applied more rigorously to semiconductors, have been successful in predicting the sign, magnitude, and

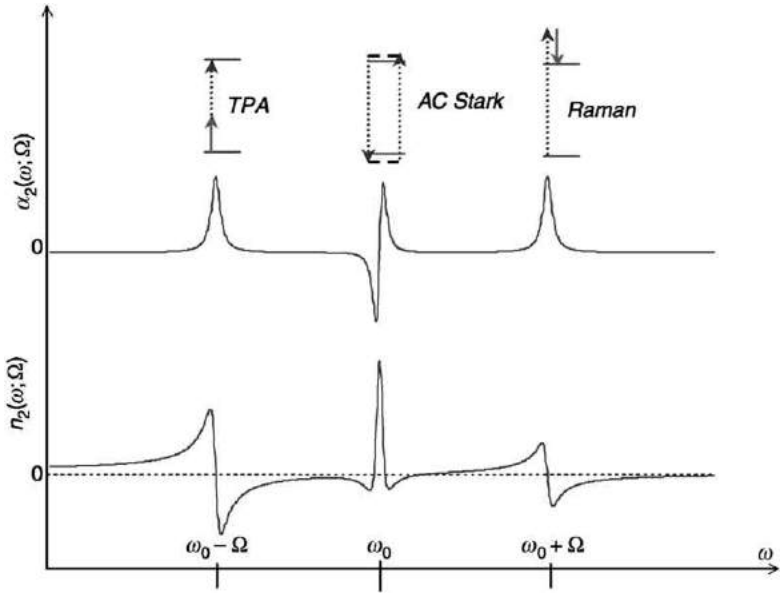


Fig. 2 Upper trace: the nonlinear absorption coefficient in a fictitious 'degenerate' two-level system. Lower trace: the resulting nonlinear refractive index obtained using the KK relations. The insets show the three possible physical mechanisms involved.

dispersion of n_2 due to the anharmonic motion of bound electrons. This will be briefly discussed later. Returning to the mathematical foundation of KK relations, we use **Eqs. (13)** and **(14)** to write **Eq. (15)** in terms of the nonlinear susceptibility $\chi^{(3)}$:

$$\Re e\left\{\chi^{(3)}(\omega_1,\omega_2,-\omega_2)\right\}=\frac{1}{\pi}\wp\int_{-\infty}^{\infty}\frac{\Im m\{\chi^{(3)}(\omega',\omega_2,\omega_2)\}}{\omega'-\omega_1}\mathrm{d}\omega' \tag{16}$$

The above dispersion relation for $\chi^{(3)}$ was obtained using the physical and intuitive arguments that followed the linearization scheme depicted in **Fig. 1**. General dispersion relations can be formulated following a mathematical procedure that is similar to the derivation of the linear KK relations. In this case we apply the causality condition directly to the nth-order nonlinear response $R^{(n)}$. For example, without loss of generality, we can write

$$R^{(n)}(\tau_1,\tau_2,\ldots,\tau_n)=R^{(n)}(\tau_1,\tau_2,\ldots,\tau_n)\Theta(\tau_j) \tag{17}$$

and then calculate the Fourier transform of this equation. Here j can apply to any one of the indices $1,2,\ldots,n$. Following the same procedure as for a linear response, we obtain

$$\chi^{(n)}(\omega_1,\omega_2,\ldots,\omega_j,\ldots,\omega_n)=\frac{-\mathrm{i}}{\pi}\wp\int_{-\infty}^{\infty}\frac{\chi^{(n)}(\omega_1,\omega_2,\ldots,\omega',\ldots,\omega_n)}{\omega_j-\omega'}\mathrm{d}\omega' \tag{18}$$

By separating the real and imaginary parts of this equation, we get the generalized Kramers–Krönig relation pairs for a non-degenerate, nth-order nonlinear susceptibility:

$$\Re e\left\{\chi^{(n)}(\omega_1,\omega_2,\ldots,\omega_j,\ldots,\omega_n)\right\}=\frac{1}{\pi}\wp\int_{-\infty}^{\infty}\frac{\Im m\{\chi^{(n)}(\omega_1,\omega_2,\ldots,\omega',\ldots,\omega_n)\}}{\omega'-\omega_j}\mathrm{d}\omega' \tag{19}$$

and

$$\Im m\left\{\chi^{(n)}(\omega_1,\omega_2,\ldots,\omega_j,\ldots,\omega_n)\right\}=-\frac{1}{\pi}\wp\int_{-\infty}^{\infty}\frac{\Re e\{\chi^{(n)}(\omega_1,\omega_2,\ldots,\omega',\ldots,\omega_n)\}}{\omega'-\omega_j}\mathrm{d}\omega' \tag{20}$$

In particular, for $\chi^{(3)}$ processes having $\omega_1=\omega_a,\omega_2\omega_b$, and ${}_3b$, this becomes identical to **Eq. (16)**.

Note that in describing the nonlinear susceptibilities, no special attention was given to the harmonic generating susceptibility $\chi^{(N)}(N\omega)\equiv\chi^{(N)}(\omega,\omega,\ldots\omega)$, i.e., the susceptibility generating the Nth harmonic at $N\omega$. It turns out that in addition to the KK relations given by **Eqs. (19)** and **(20)**, the real and imaginary parts of $\chi^{(N)}(N\omega)$ can also be related in different sets of dispersion integrals that involve only the degenerate forms of the susceptibilities. A more general yet simple analysis gives the most general form of KK relations for any type of $\chi^{(n)}$:

$$\chi^{(n)}(\omega_1+p_1\omega,\omega_2+p_2\omega,\ldots,\omega_m+p_m\omega)=\frac{1}{\mathrm{i}\pi}\wp\int_{-\infty}^{\infty}\frac{\chi^{(n)}(\omega_1+p_1\Omega,\omega_2+p_2\Omega,\ldots,\omega_m+p_m\Omega)}{\Omega-\omega}\mathrm{d}\Omega \tag{21}$$

for all $p_1,p_2,\ldots,p_m\geq 0$. Setting $\omega_1=\omega_2=\cdots=\omega_m\equiv 0$, and $p_1=p_2=\cdots=p_m=1$ in **Eq. (21)** yields an interesting form of the KK relations for the Nth-harmonic susceptibilities:

$$\Re e\left\{\chi^{(N)}(N\omega)\right\}=\frac{1}{\pi}\wp\int_{-\infty}^{\infty}\frac{\Im m\{\chi^{(N)}(N\omega')\}}{\omega'-\omega}\mathrm{d}\omega' \tag{22}$$

These dispersion relations have allowed calculations of $\chi^{(2)}(2\omega)$ and $\chi^{(3)}(3\omega)$ in semiconductors using full band structures.

At the beginning of this article, it was noted that all the KK relations for nonlinear optics were known in the early days of the field. Their application in unifying nonlinear absorption (in particular two-photon absorption) and the optical Kerr effect (n_2) in solids came only much later. More recent work demonstrated that the KK relations are a powerful analytical tool in nonlinear optics. Following the picture of a degenerate two-level system shown in **Fig. 2**, a simple two-band model has been used to calculate the nonlinear absorption coefficient, $\alpha_2(\omega_1;\omega_2)$, resulting from three mechanisms: 2PA, the Raman absorption process, and the ac Stark effect. The optical Kerr coefficient $n_2(\omega_1;\omega_2)$ was then calculated using **Eq. (15)**. Of particular practical interest is the degenerate case $(\omega_1=\omega_2=\omega)$, from which the 2PA coefficient $\beta(\omega)=\alpha_2(\omega;\omega)$ can be extracted. **Fig. 3** depicts the calculated dispersion of n_2 and β as a function of $\hbar\omega/E_g$ where E_g is the bandgap energy of the solid. The dispersion of n_2 and its sign reversal shown in **Fig. 3** has been observed experimentally in many optical solids.

Finally, let us discuss a related implication of causality in nonlinear optics. The KK dispersion relations are traditionally derived in terms of internal material parameters such as susceptibility, absorption coefficient, and refractive index. Similar to the case of electrical circuits, one can obtain dispersion relations that apply to an external transfer function of the system that relates an input signal to an output signal. In this case, the dispersion of the transfer function includes system structure as well as the intrinsic dispersion of the material. As an optical (and linear) example, consider a Fabry–Perot etalon. The optical transmission of this system has well-known spectral features that are primarily caused by structural dispersion (i.e., interference) in addition to the intrinsic dispersion of the material. Causality still demands that the transmitted signal has a phase variation whose value and dispersion can be determined using a KK relation linking the real and imaginary parts of the transmission coefficient. In other words, the KK relations provide a spectral correlation between the real and imaginary components of the transfer function which in turn may translate to a spectral correlation between the phase and amplitude of the transmitted signal. However, the variations in phase do not necessarily imply the presence of a varying index of refraction, nor does an amplitude variation suggest the existence of material absorption (dissipation). Ultimately, this implies that any mechanism causing a variation in amplitude (including reflection, scattering, or absorption) must be accompanied by a phase variation. (One should note that the reverse of

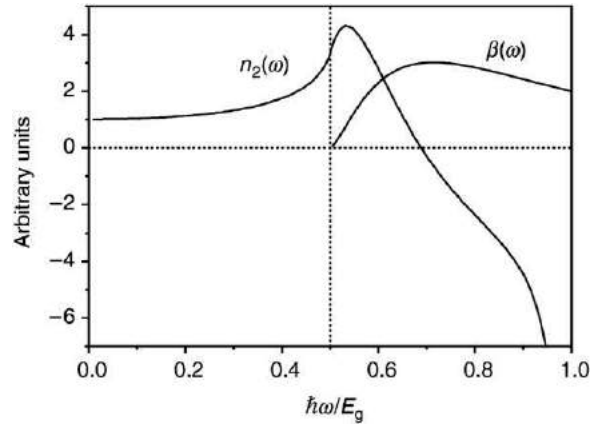


Fig. 3 The two-photon absorption coefficient in semiconductors (β) calculated for a two-band model. The resultant nonlinear refractive index (n_2) obtained using a KK transformation of the calculated nondegenerate nonlinear absorption coefficient includes all major mechanisms.

the previous statement is not necessarily true; i.e., a variation in phase does not have to be accompanied by an amplitude modulation.)

In nonlinear optics with the 'black box' approach of **Fig. 1**, the optical perturbation ξ (with frequency Ω) can render an amplitude variation in the probe (at ω) using various frequency mixing schemes in a noncentrosymmetric material (i.e., with nonzero $\chi^{(2)}$). For instance, the probe at ω can be depleted by nonlinear conversion to $\omega_{\text{sum}} = \omega + \Omega$ via sum-frequency generation involving $\chi^{(2)}(\omega, \Omega)$ and/or to $\omega_{\text{diff}} = \omega - \Omega$ via difference-frequency generation involving $\chi^{(2)}(\omega, -\Omega)$. Such a conversion (or depletion) should be accompanied by a phase variation according to the KK dispersion relations. This type of nonlinear phase modulation is known as a $\chi^{(2)}{:}\chi^{(2)}$ cascaded nonlinearity. Such cascaded processes are routinely (and more simply) analyzed with Maxwell's equations governing the propagation of beams in a second-order nonlinear material. The KK relations, however, provide an interesting physical perspective of the process. We find that cascaded second-order nonlinearities are yet another manifestation of causality in nonlinear optics.

Further Reading

Bassani, F., Scandolo, S., 1991. Dispersion-relations and sum-rules in nonlinear optics. Physical Review B–Condensed Matter 44, 8446–8453.
Caspers, P.J., 1964. Dispersion relations for nonlinear optics. Physical Review A 133, 1249–1251.
Hutchings, D.C., Sheik-Bahae, M., Hagan, D.J., Van Stryland, E.W., 1992. Kramers–Krönig relations in nonlinear optics. Optical and Quantum Electronics 24, 1–30.
Kogan, S.M., 1963. On the electromagnetics of weakly nonlinear media. Soviet Physics JETP 16, 217–219.
Nussenzveig, H.M., 1972. Causality and Dispersion Relations. New York: Academic Press.
Price, P.J., 1964. Theory of quadratic response functions. Physical Review 130, 1792–1797.
Ridener, F.L.J., Good, R.H.J., 1975. Dispersion relations for nonlinear systems of arbitrary degree. Physical Review B 11, 2768–2770.
Sheik-Bahae, M., Van Stryland, E.W., 1999. Optical nonlinearities in the transparency region of bulk semiconductors. In: Garmire, E., Kost, A. (Eds.), Nonlinear Optics in Semiconductors I 58. Academic Press, pp. 257–318.
Sheik-Bahae, M., Hutchings, D.C., Hagan, D.J., Van Stryland, E.W., 1991. Dispersion of bound electronic nonlinear refraction in solids. IEEE Journal of Quantum Electronics 27, 1296–1309.
Smet, F., Vangroenendael, A., 1979. Dispersion-relations for *N*-order non-linear phenomena. Physical Review A 19, 334–337.
Toll, J.S., 1956. Causality and the dispersion relation: logical foundations. Physical Review 104, 1760–1770.

Nonlinear Optical Phase Conjugation

BY Zeldovich, University of Central Florida, Orlando, FL, USA

Phase conjugation (PC) beams and their applications are best illustrated by Fig. 1, which shows the passage of a coherent planewave incident upon a transparent optical element with inhomogeneities of refractive index (Fig. 1(a)). Reversibility of light propagation implies the existence of such an 'anti-distorted' beam, which reverts back to a planewave after reversing through the same inhomogeneities (Fig. 1(b)). A real-valued monochromatic optical wave, $E_{\text{real}}(\mathbf{R}, t)$, may be represented by the complex amplitude $E(\mathbf{R})$ via $E_{\text{real}}(\mathbf{R}, t) = 0.5[E(\mathbf{R})\exp(-i\omega t) + E^*(\mathbf{R})\exp(i\omega t)]$, where $E^*(\mathbf{R})$ represents the complex conjugate of $E(\mathbf{R})$:

$$\begin{aligned} E(\mathbf{R}) &= |E(\mathbf{R})|\exp[i\varphi(\mathbf{R})] \\ E^*(\mathbf{R}) &= |E(\mathbf{R})|\exp[-i\varphi(\mathbf{R})] \end{aligned} \tag{1}$$

Mathematical expression of time-reversal, $E_{\text{real}}(\mathbf{R}, t \to E_{\text{real}}(\mathbf{R}, -t)$, becomes the exchange $E(\mathbf{R}) \leftrightarrow E^*(\mathbf{R})$ for monochromatic waves. The reversibility of propagation means that the complex propagating field $E(\mathbf{R})$ (for example, planewave $\exp(i\mathbf{k}\cdot\mathbf{R})$), is a wave equation solution equivalent to $E^*(\mathbf{R})$ (in this example, $\exp(-i\mathbf{k}\cdot\mathbf{R})$). Conjugation of complex amplitude means reversal of the sign of the phase, and a mixed label PC has been coined and nowadays is firmly established. Terms such as 'wave front reversal' and 'generation of time-reversed replica of the beam' are also used to describe PC.

Fig. 1 also illustrates one of the most important applications of PC. If the element in question is a laser-type amplifier, then double-passage allows the extraction of energy from the optically inhomogeneous laser medium in the form of a perfectly collimated beam of diffraction-limited divergence. A PC device may also serve as a mirror of a laser resonator, resulting in the same beam-correction purpose.

One of the most practical and robust methods of PC is based on stimulated Brillouin back-scattering (SBS). The incident beam illuminates the SBS-active transparent medium, for example, liquids (CS_2, CCl_4, acetone, etc.), compressed gases (CH_4, SF_6, etc.), or solids (fused or crystalline quartz, glass, etc.). This 'pump' beam, $E(\mathbf{R}) \equiv E(\mathbf{r}, z)$, must possess well-developed transverse inhomogeneities of intensity $|E(\mathbf{r}, z)|^2$. Here, $\mathbf{r} = \{x, y\}$ is the part of coordinate vector transverse with respect to the central direction z of the beam in question, and $\mathbf{R} = \{\mathbf{r}, z\}$. These inhomogeneities may constitute a narrow focal waist in the case of weakly distorted focused beams, speckle-inhomogeneities in cases of strong distortions, or a combination thereof; see Fig. 2, where solid lines symbolize the 'rays' of the incident pump.

Spontaneous scattering of the pump results in a multitude of possible transverse profiles $S(\mathbf{r}, z = 0)$ of the signal, whose 'rays' are depicted via the dotted lines on Fig. 2. The signal is amplified exponentially due to SBS processes. This exponent in a simplified

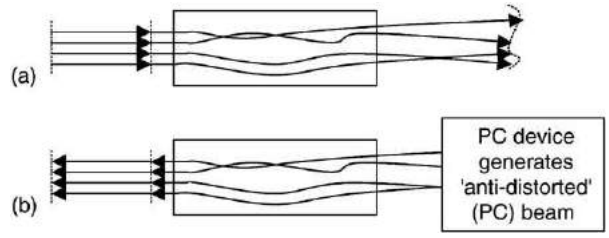


Fig. 1 (a) Collimated beam is distorted by propagation through inhomogeneous medium. (b) Conjugate, a.k.a. antidistorted beam becomes collimated after backward passage through the same medium.

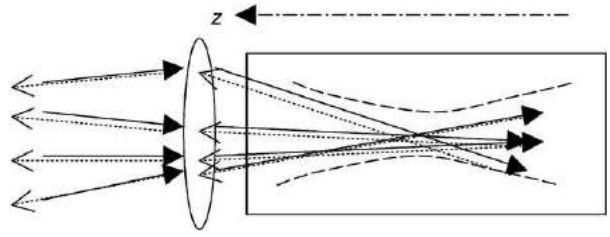


Fig. 2 Stimulated Brillouin Scattering (SBS) method of PC. Solid lines symbolize the rays of incident 'beam', while dotted lines depict the rays of the signal amplified by the SBS process.

 doi:10.1016/B978-0-12-809283-5.00756-4

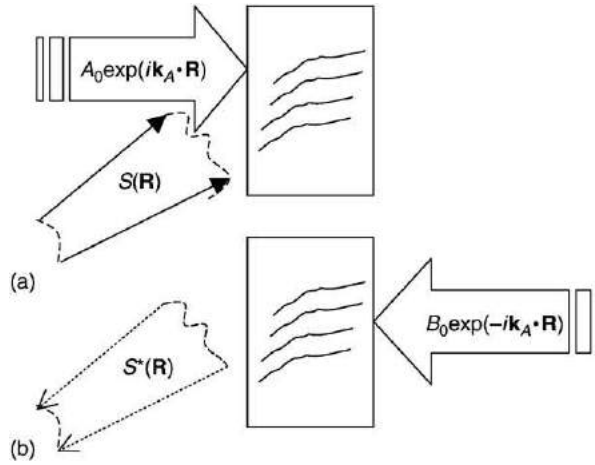


Fig. 3 (a) Recording of a hologram in a photosensitive medium by a reference wave A and signal wave S. (b) Read-out of that hologram by the wave B counter-propagating to A results in generation of the conjugated signal S*.

form may be represented as $|S(\mathbf{r},z)|^2 \propto \exp\left(\int g\mathrm{d}z\right)$, with the gain $g(z)$ being the result of transverse overlapping of the intensity profiles of signal and pump:

$$g(z) \approx \text{const} \cdot \langle |E(\mathbf{r},z)|^2 \cdot |S(\mathbf{r},z)|^2\rangle / \langle |S(\mathbf{r},z)|^2\rangle \tag{2}$$

Thus, the similarity of the intensity profile of the signal to that of the pump is rewarded by the exponential preference in the output signal level. However, this is not enough to guarantee that the output backward-scattered signal $S(\mathbf{r}, z)$ is phase conjugate with respect to the pump. It is here that the mutual reversibility of propagation of strongly inhomogeneous signal and pump becomes crucial. Namely, good transverse overlapping is sustained along the entire interaction length z, only if $S(\mathbf{r},z) \propto E^*(\mathbf{r}, z)$. This constitutes the 'discrimination mechanism' of SBS-PC: the 'conjugate mode' of the signal has a z-sustained advantage in exponential gain and, under appropriate conditions, suppresses all nonconjugate signal configurations in competition for pump power. Nonlinear-optical wave theory of this discrimination process has been developed further.

The holographic mechanism of PC has a different nature. It may take the form of static holography or dynamic holography; the latter essentially may be considered as a part of nonlinear optics. Here is a simplified description of one of the variants. Suppose one wants to obtain a conjugate replica of the incident monochromatic signal, whose complex profile is $S(\mathbf{R})$. A plane reference wave $A(\mathbf{R}) = \exp(i\mathbf{k}_A \cdot \mathbf{R})$ is also directed to the registering medium. If $S(\mathbf{R})$ and $A(\mathbf{R})$ are mutually coherent, then an interference pattern of intensity is produced $\{S^*(\mathbf{R})A(\mathbf{R}) + \text{compl.conj.}\}$. The medium records this pattern in the form of the modulation of refractive index and/or of the absorption coefficient, $\delta\varepsilon(\mathbf{R}) \propto (c_1 + ic_2)\{S^*(\mathbf{R})A(\mathbf{R}) + \text{c.c.}\}$ (**Fig. 3(a)**). At the 'reconstruction' stage, the hologram is illuminated by another plane reference wave, $B(\mathbf{R}) = B_0\exp(i\mathbf{k}_B \cdot \mathbf{R})$. In the specific case, when $\mathbf{k}_B = -\mathbf{k}_A$, the source δD_{conj} of the reconstructed wave becomes, as shown by **Fig. 3(b)**:

$$\delta D_{\text{conj}} = (c_1 + ic_2)[S^*(\mathbf{R})A(\mathbf{R})] \cdot B(\mathbf{R}) \propto S^*(\mathbf{R})(c_1 + ic_2)A_0B_0 \tag{3}$$

Reversibility of propagation laws guarantees that this source will indeed excite a conjugate wave with high efficiency. Both processes, recording of the interference pattern and reconstruction of the conjugate wave, may occur simultaneously, if one deals with the dynamic holography. The same process may also be described as nonlinear-optical four-wave mixing (FWM) via cubic nonlinearity $\chi^{(3)}$:

$$\delta D_{\text{conj}}(\mathbf{R}) = \chi^{(3)}S^*(\mathbf{R})A(\mathbf{R})B(\mathbf{R}) = \chi^{(3)}S^*(\mathbf{R})A_0B_0 \tag{4}$$

with the four waves; A, B, S, S^*.

Important characteristics of a PC device are: fidelity of conjugation, efficiency of returning back-reflected power/energy, and reaction time or build-up time. The fidelity parameter must show how close the output of the device $E_{\text{out}}(\mathbf{r})$ is to the perfectly conjugate profile $S^*(\mathbf{R})$ of the incident signal (up to an arbitrary constant complex factor). Among other definitions, the square modulus of the normalized transverse overlapping integral of the fields is often discussed:

$$f = \frac{|\int\int E_{\text{out}}(x,y)S(x,y)\mathrm{d}x\mathrm{d}y|^2}{\left(\int\int |E_{\text{out}}(x,y)|^2\mathrm{d}x\mathrm{d}y \cdot \int\int |S(x,y)|^2\mathrm{d}x\mathrm{d}y\right)} \tag{5}$$

To achieve good (close to 1) fidelity of PC in the holographic or FWM scheme, one has to guarantee that the reference waves, $A_0\exp(i\mathbf{k}_A\cdot\mathbf{R})$ and $B_0\exp(i\mathbf{k}_B\cdot\mathbf{R})$ are exactly conjugate to each other, $\mathbf{k}_A = -\mathbf{k}_B$. That, in turn, requires the absence of any distortions in the holographic or nonlinear medium. To the contrary, there are very modest requirements on the phase inhomogeneities of the medium in the SBS scheme of PC, and for that reason it is usually labeled as a scheme of self-phase conjugation.

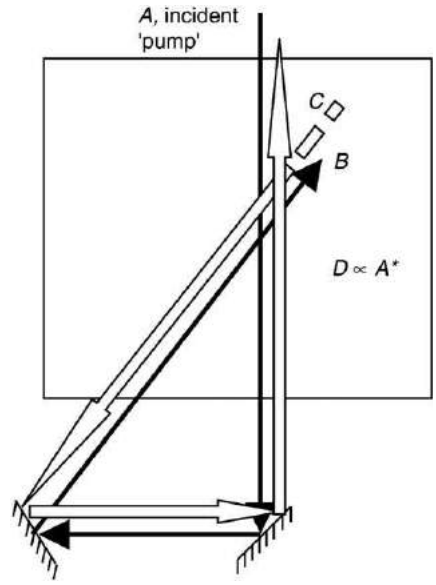


Fig. 4 Tail-biting scheme of phase conjugation.

A hybrid scheme, Brillouin-assisted FWM, has been suggested and implemented, where mutually conjugate reference waves are produced via SBS-PC, while FWM, in a separate medium, exploits SBS nonlinearity. This scheme has produced extremely high, up to about 10^{10}, reflectivity, accompanied by an extremely low-noise detection of incoming signals.

A number of other PC schemes were first realized with the use of an important class of materials for nonlinear optics: photorefractive crystals (PRC). These are crystals where ionization of the dopants by the incident light creates interference pattern of charge separation. The linear electro-optic (Pockels) effect transforms the resulting patterns of the electrostatic field into patterns of refractive index, thus creating dynamic holograms. A remarkable property of these crystals is that most of them act as very good electrical insulators. Therefore, the only source of conductivity, which tends to erase the hologram, is photoconductivity, the value of the latter being proportional to the intensity of incident light itself. For this reason, the steady-state strength of the hologram turns out to be independent of the intensity. It is the build-up time that must be increased, if intensity is low. One can achieve various processes in PRC, including those similar to the $\chi^{(3)}$-type optical Kerr effect, stimulated scattering, etc.

The simplest use of PRC for generation of PC waves is to devise a FWM scheme (see **Fig. 3**). However, a multitude of other, nontrivial schemes have been devised and implemented. The most impressive use of PRC for PC is based on interactions of the beams in the vicinity of a corner of a rectangular crystal, typically $BaTiO_3$, these interactions being of the stimulated scattering type. To honor J. Feinberg's cat, whose image was reconstructed in the first demonstration, this scheme is universally called 'cat conjugator'. It turned out to be a very reliable and robust scheme of self-phase-conjugation (SPC).

Another important group of the SPC scheme is called 'tail-biting'. One of the variants of the latter is shown in **Fig. 4**. Incident 'pump' wave $A(\mathbf{R})$ whose transverse profile one wants to conjugate, enters the medium, and then is redirected by mirrors back into the same medium, to 'bite itself'. For the purpose of discussion, this redirected pump is labeled as wave $B(\mathbf{R})$ on its second arrival into the medium. Spontaneous scattering results in a seed for the amplification of the wave $C(\mathbf{R})$ via the process of stimulated scattering (SS) $A(\mathbf{R}) \rightarrow C(\mathbf{R})$.

Wave $C(\mathbf{R})$ is also redirected into the medium by the same mirrors, and is labeled as wave $D(\mathbf{R})$ on its second arrival into the medium. Both $C(\mathbf{R})$ and $D(\mathbf{R})$ are depicted in **Fig. 4** by white arrows. Wave $D(\mathbf{R})$ is also amplified due to the SS process $B(\mathbf{R}) \rightarrow D(\mathbf{R})$.

As it follows from the macroscopic description of SS process, volume gratings of refractive index are recorded in the medium, $\delta n(\mathbf{R}) \propto -i[A^*(\mathbf{R})C(\mathbf{R}) + B^*(\mathbf{R})D(\mathbf{R})]$. Among all the transverse profiles for the seed $C(\mathbf{R})$ the one that happens to be proportional to $B^*(\mathbf{R})$ will be reflected by the mirrors into the medium in the form $D(\mathbf{R}) \propto A^*(\mathbf{R})$. The latter case is again a consequence of the reversibility of propagation laws, where both gratings have the same profile and add coherently in the form $A^*(\mathbf{R})B^*(\mathbf{R})$. What is even more important, the grating $B^*(\mathbf{R})D(\mathbf{R}) \propto B^*(\mathbf{R})A^*(\mathbf{R})$ serves to close the feedback loop for seeding the appropriate $C(\mathbf{R})$. This, and other variants of tail-biting schemes, have also proved to be reliable and robust. Quite often a laser amplifier is inserted in the path $A \rightarrow B$ and $C \rightarrow D$, thus enhancing the feedback.

Another important scheme is traditionally called 'double PC'; albeit it may be better described as 'mutual PC'. **Fig. 5** should help in understanding that scheme. Consider the beam $A(\mathbf{R})$ incident to a medium active with respect to the stimulated-scattering-type process $A(\mathbf{R}) \rightarrow C(\mathbf{R})$. This process originates from a random seed and results in a fan of different waves $C(\mathbf{R})$. Such a process is characteristic of PRCs and is a consequence of recording dynamic gratings of refractive index $\delta n(\mathbf{R}) \propto -i[A^*(\mathbf{R})C(\mathbf{R})]$. Suppose another beam $B(\mathbf{R})$, typically incoherent with respect to $A(\mathbf{R})$, illuminates the medium from the other side, and also is engaged in the fanning process, this time $B(\mathbf{R}) \rightarrow D(\mathbf{R})$, with the grating of refractive index $\delta n(\mathbf{R}) \propto -i[B^*(\mathbf{R})D(\mathbf{R})]$ involved. Among all the

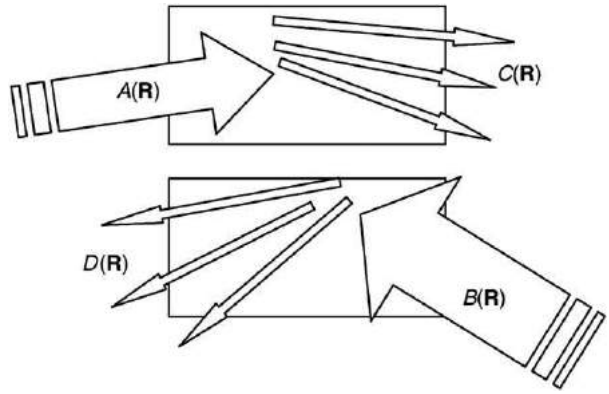


Fig. 5 Towards the mechanism of double PC (mutual PC).

transverse profiles for the seed grating $C(\mathbf{R})A^*(\mathbf{R})$, the one that happens to be proportional to $B^*(\mathbf{R})A^*(\mathbf{R})$ will be supported by similar profiles in the $B(\mathbf{R})\rightarrow D(\mathbf{R})$ scattering. In this way, two fanning processes enhance each other and close the feedback loop. An important consequence is that one obtains the following output waves:

$$\begin{aligned} C(\mathbf{R}) &= \text{const}_1 \times B^*(\mathbf{R}) \\ D(\mathbf{R}) &= \text{const}_2 \times A^*(\mathbf{R}) \end{aligned} \tag{6}$$

One can say that the output wave $C(\mathbf{R})$ presents a conjugate replica of $B(\mathbf{R})$-beam, while the output wave $D(\mathbf{R})$ presents a conjugate replica of $A(\mathbf{R})$-beam; hence the label 'mutual PC'. In some experiments the incident beams, $A(\mathbf{R})$ and $B(\mathbf{R})$, were of somewhat different colors (i.e., from different lasers), or their pulses did not overlap in time in the medium. It is the grating memory that 'informs' one beam about the presence of the other. The mechanism of cat conjugator is considered to be a combination of tail-biting and double PC schemes.

Pumped laser material with saturable gain often serves as an important nonlinear-optical medium for implementing PC. One of the evident advantages is that possible low efficiency of properly nonlinear process may be compensated by the gain in the same piece of material.

Considerable attention is paid nowadays to the chirp reversal of optical pulses that are used in fiber optical communications. Such chirped pulses, $S(t) = \exp\left(-i\omega_0 t + i\beta t^2/2 - t^2/\tau_1^2\right)$, with the value of the 'chirp' $\mathrm{d}\omega/\mathrm{d}t = -\beta$ and an increased pulse duration τ_1, arise as a result of propagation in a fiber which possesses group velocity dispersion (GVD). Pulse stretching and chirp are the consequences of different propagation times for the different frequency constituents of the pulse. If some device changes the sign of that chirp, then subsequent propagation of the pulse, through a piece of fiber with the same GVD, restores the duration of the pulse to original shorter value τ_0. A scheme has been suggested to use nonlinear optical FWM in a $\chi^{(3)}$-medium or three-wave mixing (ThWM) in a $\chi^{(2)}$-medium:

$$\begin{aligned} \delta D_{\text{conj}}(t) &= \chi^{(3)} S^*(t) A(t) B(t) \\ \delta D_{\text{conj}}(t) &= \chi^{(3)} S^*(t) C(t) \end{aligned} \tag{7}$$

Indeed, complex conjugation of the time dependence of the signal's field is equivalent to the change of the chirp sign. The second (ThWM) expression is written assuming that the reference wave $C(t) \propto \exp(-2i\omega_0 t)$, so that is has the frequency double that of the signal. The same process of ThWM has been shown to yield the conjugation of transverse phase profile of the beams, but for a number of reasons it was not applied.

As for the applications of PC, one of them was discussed above (**Fig. 1**: double-pass scheme of a laser amplifier). Lasers with one or two PC mirrors, or with more complicated PC scheme, promise generation of high-transverse-quality beams with the use of realistic laser media, the latter inevitably possessing thermal and other distorting inhomogeneities. Another important potential application is free-space optical communications: Earth–Earth or Earth–satellite through atmosphere, or satellite–satellite communications. An 'interrogating' beam may be sent in one direction through an imperfect optical system and/or through turbulent atmosphere. Conjugation of transverse profile of the 'interrogating' beam at the other end of the communication link leads to an almost perfect redirection of the reflected beam towards the 'beacon' source, while time modulation may carry the information.

Readout of information data from a holographic storage is almost always performed in the regime of reading the conjugate wave, since it corrects for the most part of aberrations of optical systems in question.

To conclude, one may use a citation from the two consecutive *Scientific American* popular papers on PC: 'at present the number of applications would seem to be limited only by imagination.'

Reviews of various aspects of phase conjugation may be found in monographs and *Scientific American* papers listed below. The author of the present article was introduced to PC by his colleague V. V. Ragulskii, and has greatly benefited from collaboration with V. V. Ragulskii and V. V. Shkunov.

See also: Adaptive Optics. Information Theory in Imaging

Further Reading

Bespalov, V.I., Pasmanik, G.A., 1994. Nonlinear Optics and Adaptive Laser Systems. Commack, NY: Nova Science Publishers.
Feinberg, J., 1986. Self-pumped, continuous wave phase conjugator using internal reflection. Optics Letters 7, 486.
Fisher, R.A. (Ed.), 1983. Optical Phase Conjugation. New York: Academic Press.
Gower, M., Proch, D. (Eds.), 1994. Optical Phase Conjugation. Berlin: Springer Verlag.
Hellwarth, R.W., 1982. Optical beam conjugation by stimulated backscattering. Optical Engineering 21, 257.
Nosatch, O.Y., Popovichev, V.I., Ragulskii, V.V., Faizullov, F.S., 1972. Compensation of phase distortions in an amplifying medium by a 'Brillouin mirror'. JETP Letters 16, 435.
Pepper, D.M., 1986. Applications of optical phase conjugation. Scientific American 254, 74–83.
Shkunov, V.V., Zeldovich, B.Y., 1985. Optical phase conjugation. Scientific American 253, 54–59.
Yariv, A., 1978. Phase conjugate optics and real-time holography. IEEE Journal of Quantum Electronics 14, 650.
Zeldovich, B.Y., Mamaev, A.V., Shkunov, V.V., 1995. Speckle-Wave Interactions in Application to Holography and Nonlinear Optics. Boca Raton, FL: CRC Press.
Zeldovich, B.Y., Popovichev, V.I., Ragulskii, V.V., Faizullov, F.S., 1972. On relation between wavefronts of reflected and exciting radiation in stimulated Brillouin scattering. JETP Letters 16, 109.
Zeldovich, B.Y., Shkunov, V.V., Pilipetsky, N.F., 1985. Principles of Phase Conjugation. Berlin: Springer-Verlag.

Photorefraction

M Cronin-Golomb, Tufts University, Medford, MA, USA
B Kippelen, University of Arizona, Tucson, AZ, USA

Nomenclature

γ	Amplitude gain [cm^{-1}]
E_0	Applied dc field [Vm^{-1}]
$\mathbf{j}$	Current [Am^{-2}]
ε	Dielectric permittivity [Farads m^{-1}]
E_D	Diffusion field [Vm^{-1}]
N_A	Electron acceptor density [m^{-3}]
e	Electron charge [C]
N_D	Electron donor density [m^{-3}]
k_g	Grating wavenumber [m^{-1}]
Γ	Intensity gain [cm^{-1}]
ℓ	Interaction length [cm]
N_D^i	Ionized donor density [m^{-3}]
E_Q	Limiting space charge field [Vm^{-1}]
μ	Mobility [$m^2\ V^{-1}\ sec^{-1}$]
E_μ	Mobility field [Vm^{-1}]
I	Normalized intensity [$V^2\ m^{-2}$]
s	Photo-ionization coefficient [$kg\ sec^{-2}$]
τ	Photorefractive time constant [sec]
r	Pump ratio [1]
ω	Radian frequency [sec^{-1}]
γ_R	Recombination coefficient [$m^3\ sec^{-1}$]
A	Slowly varying optical electric field [Vm^{-1}]

Introduction and Basics of the Photorefractive Effect

In 1966 Ashkin and coworkers were pursuing research on optical devices using the electro-optic material lithium niobate. They noticed that the refractive index of lithium niobate would change when it was exposed to laser light, and upset the expected operation of their devices. They called the effect optical damage. Shortly thereafter, Chen, LaMacchia, and Fraser reported on the use of the optical damage effect for holographic data storage. Thus began the field of photorefractive nonlinear optics, which has been used in various applications such as real-time holography, optical data storage, optical image amplification, nondestructive testing, distortion compensation by phase conjugation, pattern recognition, and radar signal processing. Many inorganic and organic materials have been investigated for their photorefractive effects, including ferroelectrics, semiconductors, and sensitized polymers. The most well-known inorganic materials are lithium niobate, bismuth silicon oxide, barium titanate, and strontium barium niobate. Most organic photorefractive materials are based on polymeric photoconductors such as those used in xerography that are doped with electro-active molecules, some plasticizers, and sensitizers.

While in its broadest interpretation, the photorefractive effect occurs whenever light incident on a material causes a refractive index change, one usually applies the term to electro-optic index changes resulting from optically generated space charge fields. Materials that are photorefractive in this sense share the following properties

- high transmission at the operating wavelengths;
- linear electro-optic coefficients or orientational Kerr effects;
- charge carriers that become mobile when optically excited;
- trapping centers for these charge carriers to enable spatially non-uniform redistribution of charge.

Consider two beams from the same laser crossing inside a photorefractive material such as barium titanate. The interference pattern will have bright and dark fringes. Charge carriers are excited where the light is bright, then drift and diffuse to regions of relative darkness where they preferentially recombine into trapping centers. In this way, a net excess charge develops in the dark regions, and a net deficit of charge develops in the bright regions. The spatially varying charge distribution has an electric field associated with it and this electric field causes a spatially varying refractive index profile. Because the space charge, its field, and resulting refractive index have the same spatial periodicity as the original interference pattern we have a holographic phase grating. The diffraction efficiency of the hologram can easily approach 100% for materials such as barium titanate and strontium barium niobate which have high electro-optic coefficients. Likewise, such high diffraction efficiencies are easily obtained in 100 micrometer-thick photorefractive polymer films.

The Standard Rate Equation Model

The development of photorefractive gratings can be modeled using standard semiconductor rate equations. Fig. 1 shows two beams incident symmetrically on a photorefractive crystal or polymer. They form an interference pattern whose intensity may be written:

$$I(x) = I_0\left(1 + m\cos\left(k_g x\right)\right) \tag{1}$$

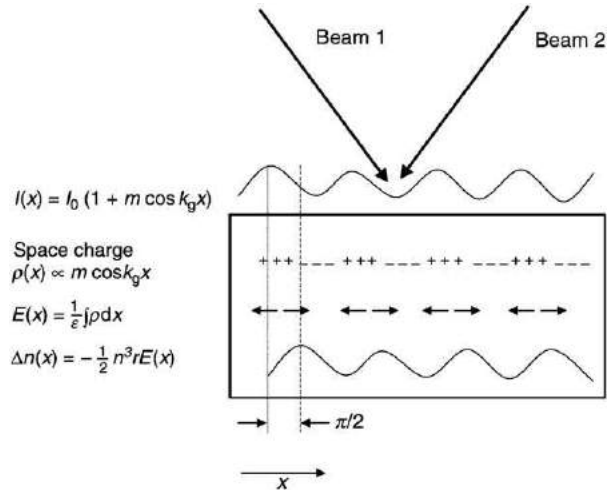


Fig. 1 Diagram of two-beam coupling process showing optical interference pattern, space charge, resultant electric field, and $\pi/2$ phase shifted electro-optically induced refractive index grating.

where x is the direction perpendicular to the interference fringes, I_0 is the average intensity, m is the modulation index, and k_g is the wavenumber of the interference pattern.

The crystal may be considered a wide bandgap semiconductor containing electron donors in the bandgap with density N_D and electron acceptors with density N_A. It is assumed that some electrons ionized from the donors permanently occupy these acceptors so that when charge in the crystal is distributed uniformly in the dark, the number density of ionized donors N_D^i is equal to N_A. Likewise, polymers contain donor and acceptor-like molecules that can be neutral or ionized. The spatially varying light distribution ionizes the donors at the following rate, assuming that $N_D^i \ll N_D$:

$$\frac{\partial N_D^i}{\partial t} = sIN_D - \gamma_R n_e N_D^i \tag{2}$$

where s is a photoionization coefficient, γ_R is a recombination parameter, and n_e is the density of excited charge carriers, which we assume here to be electrons. The model can easily be generalized to include holes. We also use the equation of charge conservation:

$$\frac{\partial N_D^i}{\partial t} = \frac{\partial n_e}{\partial t} - \frac{1}{e}\nabla\cdot\mathbf{j} \tag{3}$$

where e is the charge of an electron, and $\mathbf{j}$ is the current in the conduction band.

$$\mathbf{j} = \mu e n_e \mathbf{E} + k_B T \mu \nabla n_e \tag{4}$$

where μ is the electron mobility, k_B is Boltzmann's constant, and T is the temperature. The electric field obeys Gauss's law:

$$\nabla\cdot\mathbf{E} = -e\left(n_e + N_A - N_D^i\right)/\varepsilon \tag{5}$$

where N_A is the density of acceptors. These equations may be linearized by approximating the electron density, ionized donor density, and electric field with their first Fourier components:

$$\begin{aligned} E &= E_0 + \frac{1}{2}\left(E_1\exp(ik_g x) + E_1^*\exp(-ik_g x)\right) \\ N_D^i &= N_{D0}^i + \frac{1}{2}\left(N_{D1}^i\exp(ik_g x) + N_{D1}^{i*}\exp(-ik_g x)\right) \\ n_e &= n_{e0} + \frac{1}{2}\left(n_{e1}\exp(ik_g x) + n_{e1}^*\exp(-ik_g x)\right) \end{aligned} \tag{6}$$

This assumption is strictly valid only when the modulation index m is much less than unity. Otherwise, a generalization to higher orders in the Fourier series is required. However, the linearized theory is sufficient to illustrate the most important features of the photorefractive effect. The solution for the space charge field E_1 for the case when the interference pattern is applied at time $t=0$ is

$$E_1 = m\frac{-iE_Q(E_0 + iE_D)}{E_0 + i(E_D + E_Q)}\left(1 - \exp(-t/\tau)\right) \tag{7}$$

where E_0 is an externally applied or photovoltaic field (if any), E_D is the diffusion field

$$E_D = \frac{k_B T k_g}{e} \tag{8}$$

and E_Q is the limiting space charge field

$$E_Q = \frac{eN_A}{\varepsilon k_g} \tag{9}$$

Some photorefractive crystals, most notably $LiNbO_3$, exhibit the photovoltaic effect in which optical illumination induces a dc field across the crystal.

The sinusoidally varying space charge field E_1 operates through the linear electro-optic effect with effective coefficient r to produce a sinusoidal variation in the refractive index n of the crystal:

$$n(x) = n_0 + \frac{1}{2}\left(n_1 \exp(ik_g x) + n_1^* \exp(-ik_g x)\right) \tag{10}$$

where

$$n_1 = -\frac{1}{2} r n_0^3 E_1 \tag{11}$$

The effective electro-optic coefficient r may be found from tensor analysis of the electro-optic tensor and the vector space charge and optical fields. Notice that there is a 90-degree phase shift between the interference pattern and the refractive index grating when E_0 is zero. The time constant τ is given by

$$\tau = \frac{N_A}{sN_D I_0} \frac{E_0 + i(E_D + E_\mu)}{E_0 + i(E_D + E_Q)} \tag{12}$$

where E_μ is a mobility field

$$E_\mu = \frac{\gamma_R N_A}{\mu k_g} \tag{13}$$

When E_0 is nonzero, there is an oscillatory component to the time constant.

In contrast to the case of ordinary optical nonlinearities such as the optical Kerr effect, in which the nonlinear coupling strength is proportional to the optical intensity, the steady state strength of the photorefractive effect is independent of optical intensity while the time constant is inversely proportional to intensity in the basic model described above. The time constant varies with the photoconducting performance of a given material. In the fastest polymeric and inorganic materials it is, at the time of writing, of the order of a few milliseconds at 1 W cm^{-2} of incident optical intensity.

At low intensities (below the equivalent dark intensity, 1 W cm^{-2} in barium titanate), the above equations need to be modified to include the effect of dark conductivity. Even in the dark, there will be a few mobile charge carriers in the conduction band that tend to erase the grating. This will result in the effect appearing more Kerr-like, except still with the 90-degree phase shift between the index grating and the interference pattern.

Coupled Wave Equations

The change in refractive index n_1 can be large enough to produce substantial interactions between the writing beams. The writing beams generate a phase grating that diffracts the beams into each other. The grating influences the writing beams, which in turn influence the grating. In the cases where the diffusion field dominates, for example when the externally applied or photovoltaic field E_0 is absent, one beam is amplified by in-phase diffraction of the other beam from the grating. As shown in **Fig. 2**, this amplification results from a 90-degree phase shift due to diffraction from a phase grating coupled with a -90-degree phase shift from the spatial phase shift between the interference pattern and the index grating. The second beam is attenuated by destructive

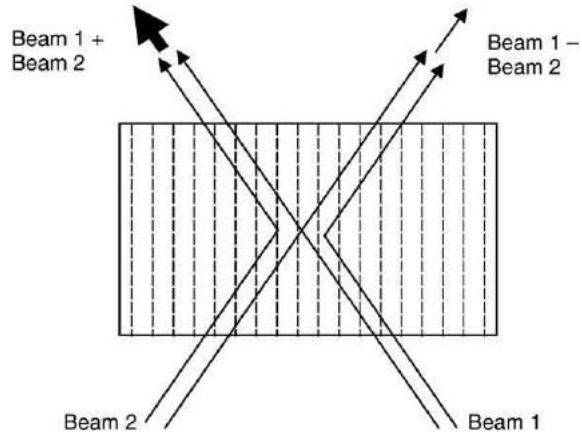


Fig. 2 Two-beam coupling amplification. Beam 1 is amplified by constructive interference. Beam 2 is de-amplified by destructive interference.

interference with the first beam diffracted by the grating. These interactions can be modeled by standard coupled wave theory. Let the electric field amplitude associated with the jth beam be

$$\mathbf{E}_j(\mathbf{r},t) = \mathbf{e}[A_j(\mathbf{r})\exp(\mathrm{i}(\mathbf{k}_j\cdot\mathbf{r}-\omega t) + \mathrm{A}_j^*\exp(-\mathrm{i}(\mathbf{k}_j\cdot\mathbf{r}-\omega t))] \tag{14}$$

where $\mathbf{e}$ is the polarization unit vector. Using the scalar wave equation

$$\nabla^2\mathbf{E} + k^2\mathbf{E} = 0 \tag{15}$$

and the slowly varying envelope approximation

$$\frac{\mathrm{d}^2A_j}{\mathrm{d}z^2} \ll k\frac{\mathrm{d}A_j}{\mathrm{d}z} \tag{16}$$

we find

$$\begin{aligned}\frac{\mathrm{d}A_1}{\mathrm{d}z} &= +\gamma\frac{A_1|A_2|^2}{I_0}\\ \frac{\mathrm{d}A_2^*}{\mathrm{d}z} &= -\gamma\frac{A_2^*|A_1|^2}{I_0}\end{aligned} \tag{17}$$

with the coupling constant γ given by

$$\gamma = -\frac{\mathrm{i}\omega\Delta n_m}{c\ \cos\vartheta} \tag{18}$$

where ϑ is the half-angle between the writing beams and Δn_m is given by

$$\Delta n_m = -\frac{1}{2}n_0^3 r\frac{E_1}{m} \tag{19}$$

Eq. (17) shows that beam 1 is amplified and beam 2 is attenuated. That beam 1 is amplified instead of beam 2 is a result of the choice of crystal orientation and hence the sign of γ. These nonlinear equations can be solved for normalized intensity $I_j = |A_j|^2$:

$$\begin{aligned}I_1(z) &= \frac{I_0}{1+(I_2(0)/I_1(0))\exp(+\Gamma z)}\\ I_2(z) &= \frac{I_0}{1+(I_1(0)/I_2(0))\exp(-\Gamma z)}\end{aligned} \tag{20}$$

where $\Gamma = \gamma + \gamma^*$ is the intensity coupling constant. In the limit where $I_1 \ll I_2$, the weak beam, beam 1, experiences exponential amplification. This amplification can be used to build optical oscillators with many of the same properties as conventional laser oscillators. The solution can be generalized to the case where γ has an imaginary component due to an applied or photovoltaic field E_0. Linear absorption can also be included.

Materials

Photorefractive materials may be separated into two broad classes: inorganic and organic. The first materials investigated were inorganic oxides and semiconductors. Their success led to efforts to endow more easily produced organic materials with the required photoconductivity, charge trapping centers, and electro-optic coefficients.

Inorganic Materials

The first requirement in a classic photorefractive material is a linear electro-optic coefficient, such as appears in sillenites such as bismuth silicon oxide $Bi_{12}SiO_{20}$ and bismuth germanium oxide $Bi_{12}GeO_{20}$. These were some of the first photorefractive materials used for image processing and phase conjugation applications. However, their electro-optic coefficients (a few pm/V) are not large enough to easily give rise to large diffraction efficiencies, or to photorefractive oscillators. Materials at a temperature near a phase transition generally have higher electro-optic coefficients because their crystal structures are on the verge of changing. They are extremely susceptible to the effect of any external influence such as the application of electric fields including photorefractive space charge fields. That is why ferroelectric materials are good candidates for photorefractive materials. These include barium titanate $BaTiO_3$, potassium niobate $KNbO_3$, and strontium barium niobate $Sr_xBa_{1-x}Nb_2O_6$. The mean free path of charge transport is less than that in sillenites, so they require more photons to reach a steady state. Therefore, ferroelectric materials are typically less sensitive than sillenites.

The second requirement is for photoconductivity. This requires the existence of photoexcitable charge carriers, either from valence band to conduction band or from intraband trapping centers. The latter source of photocarriers is used most often because the absorption length of light whose energy is greater than the bandgap is usually only a few micrometers. This places a severe restriction on the beams' interaction length ℓ. Thus there has been considerable research on suitable dopants, most commonly Fe^{2+} and Fe^{3+}. These dopants also act as trapping centers.

Lithium niobate is an example of a material with a large photovoltaic effect. When illuminated, this crystal develops a large dc field, which acts to enhance the grating strength. In those cases where the photorefractive effect is not wanted, such as in the design of lithium niobate waveguide devices, the photovoltaic effect can be greatly diminished by the addition of MgO to the crystal melt during growth.

For photorefractivity in the near infrared, semiconductors such as gallium arsenide and indium phosphide have been used with success. These materials also have the advantage that they can be grown in layered structures to produce, for example, multiple quantum wells that can be used to tailor the characteristics of optical excitation and charge transport.

Organic Materials

First-generation photorefractive polymers were designed to mimic the properties of their inorganic counterparts. Owing to their rich structural flexibility, organic synthetic materials with suitable charge transport, trapping, linear electro-optic effects and low optical absorption, were developed. This approach did build on the know-how developed previously in making photoconducting polymers for xerography and electro-optic polymers for optical modulation. Current polymers are based on an orientational photorefractive effect that leads to higher refractive index changes compared with traditional photorefractive materials. In materials with orientational photorefractivity, the refractive index change is produced by the field induced reorientation of anisotropic conjugated molecules that possess a permanent dipole moment and that have a high polarizability anisotropy. The photorefractive space charge field together with the applied field will periodically reorient these molecules and produce a periodic refractive index modulation through an orientational Kerr effect. The build-up and dynamics of this space-charge field are similar to those of traditional photorefractive crystals and can be described by **Eqs. (2)–(7)**. The time constant of the hologram is a convolution of the build-up time of the space-charge and the orientational diffusion time of the dipolar molecules in the total electric field. In contrast to crystals, polymers are nearly amorphous and the transport properties are described by charge hopping processes instead of band-type transport in crystals. Consequently, the photoconducting properties of polymers are strongly field dependent and photorefractivity is mainly observed under strong applied voltages. Numerous polymer composites have been developed using the hole transport polymer poly(N-vinylcarbazole). Several materials with a refractive index modulation amplitude of the order of a percent and two-beam coupling constants $\gamma > 100\ \mathrm{cm}^{-1}$ have been reported. New polymer composites are continuously being tested. Photorefractivity is also studied in other organic materials including organic crystals, liquid crystals, nanocomposites such as polymer-dispersed liquid crystals, or hybrid materials such as sol-gels.

Applications

Holographic Data Storage

Holographic data storage takes advantage of the Bragg selectivity of thick gratings. This allows many holograms to be superimposed in the same small volume, typically of the order of one cubic centimeter. A page of data is displayed on a spatial light modulator and a laser beam passing through the modulator is holographically recorded in the crystal with a reference beam at a specific angle. Many pages of data can be superimposed by recording many holograms with angularly multiplexed reference beams. Other multiplexing schemes are implemented by changing the shape of the wavefront of the reference beam. In principle, the upper limit of recording density is determined by the wavelength of light: one bit per cubic wavelength. If the recording wavelength is 0.5 μm, then one cubic centimeter can contain 1000 gigabytes of data. In practical circumstances, when noise is taken into consideration, the capacity is more realistically of the order of one gigabyte if a 1000×1000 spatial light modulator is used.

Distortion Compensation by Phase Conjugation

Photorefractive materials can be used to make high-reflectivity phase conjugate mirrors. The phase conjugate of a laser beam is produced when a hologram of the beam is read by another beam traveling in the opposite direction to the original reference beam. The phase conjugate reconstruction is a time-reversed copy of the original beam. If the original beam has passed through distorting optics, then the phase conjugate beam will retrace the path of the original beam through the distortion and emerge in its undistorted original state. In the standard nomenclature of phase conjugation, the input beam is called the signal, or probe, and the two reference beams are called the pumps. The output beam at $z = 0$ is called the phase conjugate beam and has zero amplitude at its input at $z = \ell$, where ℓ is the interaction length. Applications for phase conjugation exist, for example, in signal transmission through distortions and laser cavity design. If a phase conjugate mirror is used as a cavity mirror, then the effects of intracavity distortions are removed.

Since the photorefractive gratings can be very strong, the diffraction efficiency of the counterpropagating reference beam can be so high that the phase conjugate reflectivity can exceed unity. The simplest generalization of **Eq. (17)** to the four-wave mixing phase conjugation case is when only transmission gratings are important, as occurs in many circumstances depending on the mutual

coherence properties of the beams, and the material's properties. The coupled wave equations are

$$\begin{aligned}\frac{dA_1}{dz} &= \gamma\frac{(A_1A_4^* + A_2^*A_3)A_4}{I_0}\\ \frac{dA_2^*}{dz} &= \gamma\frac{(A_1A_4^* + A_2^*A_3)A_3^*}{I_0}\\ \frac{dA_3}{dz} &= -\gamma\frac{(A_1A_4^* + A_2^*A_3)A_2}{I_0}\\ \frac{dA_4^*}{dz} &= -\gamma\frac{(A_1A_4^* + A_2^*A_3)A_1^*}{I_0}\end{aligned} \tag{21}$$

In the undepleted pumps approximation, $(I_1, I_2 \gg I_3, I_4)$, the equations become linear and the solution for phase conjugate reflectivity $R = I_3(0)/I_4(0)$ is

$$R = \frac{\sinh[\gamma\ell/2]}{\cosh\left[(\gamma\ell + \ln r)/2\right]^2} \tag{22}$$

where $r = I_2/I_1$ is the ratio between the intensities of the pumping beams.

The fact that the reflectivity of the phase conjugate mirror can be greater than unity means that we can build an optical oscillator bounded by a regular mirror and a phase conjugate mirror only. Not only does it not require any additional optical gain, but it also compensates for intracavity distortions. The regular mirror can have any shape, provided that it is sufficiently reflective.

Self-pumped Phase Conjugate Mirrors

If a laser beam passes through a crystal placed inside an optical cavity formed by two facing mirrors, light scattered by imperfections in the crystal can be amplified through the photorefractive effect. The cavity provides feedback and optical oscillation can build up. The oscillation beams pump the crystal as a phase conjugate mirror for the incident beam, thus forming a self-pumped phase conjugate mirror. The feedback can even be provided by total internal reflection in the crystal in which case the crystal by itself can become a phase conjugate mirror.

Pattern Recognition and Image Filtering

Photorefractive wave mixing can be used to perform pattern recognition by matched filtering. One example would be to identify a tank in a battlefield scene, another would be to identify all of the occurrences of a particular word on a page of text. Suppose the input beams contain the corresponding pictorial information, such as might be obtained by passing the beams through an image-bearing transparency or spatial light modulator. **Eq. (21)** shows that the source for beam 3 contains a term proportional to the product of the amplitudes of the three input beams. If lenses are placed in the input beams so that the crystal receives the Fourier transforms of those beams, then the output beam at the crystal, beam 3, will be proportional to the product of the Fourier transforms of the input beams 1, 2, and 4. A lens may then be used to perform the inverse Fourier transform of the product of the Fourier transforms of the input beams, producing an output proportional to beam 1 convolved with beam 2 correlated with beam 4. If beam 1 is a point source before its Fourier transforming lens, then it will be a plane wave at the crystal. Beam 3 after inverse Fourier transformation by its lens will be the correlation of beams 4 and 2. For example, suppose we want to find all the occurrences of a particular word, say 'optics' in a given page of text. Then we would make a transparency of the word 'optics' and place that in input beam 4. We would then place an image of the page of text in beam 2. Beam 3 would then contain a field with bright spots at the places containing the word 'optics' in the original text.

The real-time holographic recording properties of photorefractive materials can also be exploited in medical imaging applications by performing time-gated holography. In this method, a hologram is formed by the temporal overlap in the photorefractive sample of a reference pulse and the first-arriving (ballistic photons) light from a stretched image-bearing pulse that has propagated through a scattering medium. The filtering of the useful photons from the scattered ones is achieved by reconstructing the hologram formed with the ballistic photons in a four-wave mixing geometry.

Optical Limiting, the Novelty Filter, and Laser Ultrasonic Inspection

The attenuation of beam 2 in **Eq. (17)** can be used in several applications including optical limiting and novelty filtering. If one wants to protect a sensor from high-intensity laser radiation, then one could split a small portion of the input beam directed at the sensor and use it as beam 1 in a photorefractive recording setup with the input beam acting as beam 2. If the laser intensity is above the equivalent dark intensity such that the optically excited charge density is greater than the thermally excited charge density, the photorefractive effect will be activated and the input beam will be de-amplified by destructive interference with beam 1, thus protecting the sensor. In materials with high gain–length products ($\gamma\ell > 1$), separate provision of beam 1 is unnecessary

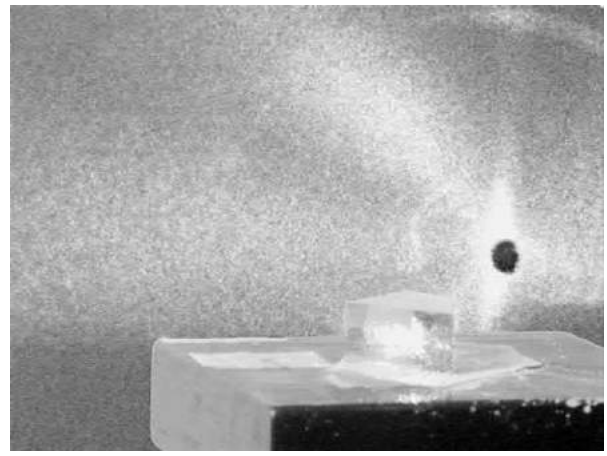


Fig. 3 Photorefractive barium titanate exhibiting amplified scattering. A helium neon laser beam is incident from the lower left, passes through the crystal to a screen where it is blacked out to prevent saturation of the camera. The screen shows brushes of amplified scattering, or fanning.

because light scattered from the input beam by imperfections in the material and other scattering centers will be greatly amplified, often to such an extent that the input beam is almost completely attenuated. This photorefractive amplification of scattered radiation is known as the fanning effect, because the amplified scattered light often appears as a fan, or brush of light, as can be seen in Fig. 3. The effect can also be used to make a novelty filter, which transmits only the moving portion of a scene. The crystal is only fast enough to respond to the slowly changing or static components of an image-bearing beam. Since the grating formed is Bragg-matched only to the slow component, the grating will only attenuate that slow component. Any rapidly changing parts of a scene pass through the crystal unattenuated. Such a filter is useful for picking out moving objects in a static cluttered background, for example a tank on a battlefield, or a micro-organism swimming against a stationary background.

A related application is in laser ultrasonic inspection. Defects in industrial material processing such as welding have characteristic ultrasonic signatures. The part under test is pinged by a pulsed laser and a probe laser is reflected from the part. As the part is shaken by the ultrasound, speckle in the reflected beam vibrates. A photorefractive recording is made of the speckle beam. Electrodes are placed on the photorefractive material, so that any optically induced currents can be detected. If the speckle pattern is not moving, or is moving more slowly than the speed of response of the photorefractive material, the photorefractive grating will be essentially at steady state: the drift currents balance the diffusion currents so there is no net current. There is no signal as the part moves through the process line. However, if the speckle beam is moving faster than the response time of the material, it will move photoexcited charge back and forth past the quasistatic grating and generate a net current for detection via the electrodes.

Adaptive Signal Processing

The relatively slow speed of photorefractive devices can be used to advantage in radio-frequency (RF) signal processing, such as signal extraction and coherent combination of signals from antenna arrays. The signal extraction application depends on grating competition in two-beam coupling. Suppose we wish to separate signals on two different RF carrier frequencies ω_1 and ω_2, respectively. The combined signal is applied to an optical carrier beam using a high-speed modulator. The resulting optical field may be represented as

$$S_1(t)\exp(\mathrm{i}\omega_1 t) + S_2(t)\exp(\mathrm{i}\omega_2 t) \tag{23}$$

It is used to pump a unidirectional ring resonator so that grating competition allows oscillation only on the strongest component of the combined signal, say S_1. The output of the oscillator is proportional to the extracted component S_1. The effectiveness of the intersignal competition is enhanced by placing another photorefractive material in the cavity. A portion of the intracavity beam is picked off by a beamsplitter and used as a two-beam coupling pump in the second material. The crystal is oriented so that the photorefractive grating diffracts the picked-off beam back into the cavity. The return of the picked-off component is most effective for the stronger component S_1 thus decreasing its loss compared to that of the weaker component S_2. This beamsplitter/crystal combination is known as a reflexive coupler.

Photorefractive Solitons

Under favorable conditions, a single beam incident on a photorefractive crystal will induce a positive refractive index change at the center of the beam. This provides a self-focussing tendency that counteracts the beam's divergence due to diffraction. When these two effects balance each other, the beam can propagate with a constant diameter. Such a beam is known as a spatial soliton in

analogy with temporal solitons in optical fibers and can occur when there is a component of the photorefractive response due to drift. The drift component of the photorefractive effect appears when a dc field E_0 is applied to the material. Potential applications are optically written waveguides and couplers.

See also: Nonlinear Optical Phase Conjugation

Further Reading

Coufal, H.J., Psaltis, D., Sincerbox, G.T. (Eds.), 2000. Holographic Data Storage. Berlin: Springer.
Gunter, P., Huignard, J.-P. (Eds.), 1988. Photorefractive Materials and their Applications I. Berlin: Springer-Verlag.
Gunter, P., Huignard, J.-P. (Eds.), 1989. Photorefractive Materials and their Applications II. Berlin: Springer-Verlag.
Nalwa, H.S., Miyata, S. (Eds.), 1997. Nonlinear Optics of Organic Molecules and Polymers. Boca Raton: CRC Press.
Nolte, D.D. (Ed.), 1995. Photorefractive Effects and Materials. Kluwer.
Pepper, D.M., Feinberg, J., Kukhtarev, N.V., 1990. The photorefractive effect. Scientific American 263, 62.
Solymar, L., Webb, D.J., Grunnet-Jepsen, A., 1996. The Physics and Application of Photorefractive Materials. Oxford: Clarendon Press.
Yeh, P., 1993. Introduction To Photorefractive Nonlinear Optics. New York: Wiley.
Yu, F., Yin, S. (Eds.), 2000. Photorefractive Optics: Materials, Properties, and Applications. San Diego: Academic Press.

Ultrafast and Intense-Field Nonlinear Optics

AL Gaeta, Cornell University, Ithaca, NY, USA
RW Boyd, University of Rochester, Rochester, NY, USA

Introduction

The tremendous development of high-powered femtosecond laser systems since the 1980s, has opened up new areas of research in nonlinear optics, plasma physics, atomic and molecular dynamics, and intense-field physics. For many of these applications, it is important to understand how ultrashort light pulses propagate through a medium under conditions in one or more of the processes in which nonlinear optics play an important role.

The starting point for the modeling of light propagation under these conditions is Maxwell's wave equation for the electric field $E(\mathbf{r}, t)$ which is given in Gaussian units as

$$\nabla^2 E - \frac{\partial^2 E}{\partial t^2} = 4\pi \frac{\partial^2 P}{\partial t^2} \tag{1}$$

where the $P(\mathbf{r}, t)$ is the polarization inside the medium. Typically, the polarization is separated into a term P_l that depends linearly on the field E and into a term P_{nl} that depends nonlinearly on the applied field. The Fourier transform of the linear polarization can be expressed as $\tilde{P}_l(\mathbf{r}, \omega) = \chi^{(1)}(\omega)\tilde{E}(\mathbf{r}, \omega)$, where $\tilde{E}(\mathbf{r}, \omega)$ is the Fourier transform of electric field $E(\mathbf{r}, t)$. For the laser–matter interactions that we consider here, we assume that the linear susceptibility $\chi^{(1)}(\omega)$ is real, in which case the wave equation can be expressed as

$$\nabla^2 \tilde{E} + \left[\frac{n_l(\omega)\omega}{c}\right]^2 \tilde{E} = -\frac{4\pi\omega^2}{c^2}\tilde{P}_{\mathrm{nl}} \tag{2}$$

where $n_l(\omega) = \sqrt{1 + 4\pi\chi^{(1)}(\omega)}$ is the frequency-dependent linear refractive index of the medium.

For light pulses that are longer than an optical period, the envelope description is valid and the electric field can be described by an amplitude envelope $A(\mathbf{r}, t)$ and a carrier frequency ω_0 such that

$$E(\mathrm{r}, t) = A(\mathrm{r}, t)\mathrm{e}^{\mathrm{i}(k_0 z - \omega_0 t)} + c.c \tag{3}$$

where $k_0 = k(\omega_0) = n_0\omega_0/c$ is the wavevector amplitude and $n_0 = n_l(\omega_0)$. This envelope description is advantageous for performing analytical and numerical studies of pulse propagation. However, for sufficiently short laser pulses, where the shape of the envelope function does not depend on the carrier phase of the carrier wave, such a description is no longer applicable. Nevertheless the envelope description can be made valid for pulses that are nearly as short as a single cycle or, alternatively, that have spectral bandwidths that are comparable to the central frequency ω_0. To derive an equation for the spatio-temporal evolution of the pulse envelope, the relation for the electric field is substituted into the wave equation (**Eq. (2)**) and the following two approximations are made: i) $k_0\partial A/\partial z \ll \partial^2 A/\partial z^2$, which signifies that the envelope varies slowly in space over a distance compared to the central wavelength; ii) $v_{\mathrm{ph}} \sim v_{\mathrm{gr}}$ where $v_{\mathrm{ph}} = c/n_0$ is the phase velocity and $v_{\mathrm{gr}} = (\mathrm{d}k/\mathrm{d}\omega)^{-1}$ is the group velocity. This latter approximation is invariably well satisfied when the frequencies contained in the laser field are highly nonresonant with any transition frequencies of the medium.

For an input pulse at $z=0$ with a peak amplitude A_0, and characteristic widths in space and time given by w_0 and τ_p, respectively, the equation for the normalized amplitude $u(r_\perp, z, t) = A(r_\perp, z, t)/A_0$ can be expressed as

$$\frac{\partial u}{\partial \xi} = \frac{i}{4}\left(1 + \frac{i}{\omega_0\tau_p}\frac{\partial}{\partial\tau}\right)^{-1}\nabla_\perp^2 u - i\sum_{n=2}\frac{L_{\mathrm{ds}}}{n!L_{\mathrm{ds}}^{(n)}}\frac{\partial^n u}{\partial\tau^n} + i\left(1 + \frac{i}{\omega_0\tau_p}\frac{\partial}{\partial\tau}\right)p^{\mathrm{nl}} \tag{4}$$

where $\tau = (t - z/v_g)/\tau_p$ is the normalized retarded time for the pulse traveling at the group velocity v_g, $L_{\mathrm{ds}}^{(n)} = \tau_p^n/\beta_n$ is the nth-order dispersion length, $\beta_n (n \geq 2)$ is the nth-order dispersion constant, $L_{\mathrm{ds}} = |L_{\mathrm{ds}}^{(2)}|$ is the dispersion length, $\zeta = z/L_{\mathrm{df}}$ is the normalized distance, $L_{\mathrm{df}} = kw_0^2/2$ is the diffraction length, p_{nl} is the normalized nonlinear polarization, and $\nabla_\perp^2$ is the transverse Laplacian. The presence of the operator $(1 + i\partial/\omega_0\partial\tau)$ in the diffraction term $(\nabla_\perp^2 u)$ of **Eq. (4)** leads to space–time focusing, while its presence in the nonlinear term results in self-steepening and both these terms arise from not making the slowly varying envelope approximation in time (i.e., $k_0\partial A/\partial t \ll \partial^2 A/\partial t^2$) in deriving **Eq. (4)**. For a self-consistent analysis of pulse propagation in the nonlinear regime, both the effects of space–time focusing and self-steepening must be included.

Nonlinear Refractive Index

For an isotropic medium in the highly nonresonant limit, the third-order term is the lowest-order contribution to the nonlinear susceptibility. This term gives rise to the nonlinear refractive index, that is, the index of refraction depends on the local intensity of

the laser field. For many materials there are two contributions to the nonlinear refractive index: i) an instantaneous part that arises from the electronic response of the medium; and ii) a noninstantaneous contribution due to the nuclear motion of the molecules (i.e, the Raman contribution). For such a medium, the nonlinear polarization may be expressed as

$$p_{\mathrm{nl}}(\zeta,\tau)=\frac{L_{\mathrm{ds}}}{L_{\mathrm{nl}}}\left[(1-f)|u(\zeta,\tau)|^2+f\gamma_R\int_{-\infty}^{\tau}\mathrm{d}\tau' R(\tau-\tau')|u(\zeta,\tau')|^2\right]u(\zeta,\tau) \tag{5}$$

where $L_{\mathrm{nl}}=(c/\omega_0 n_2 I_0)$ is the nonlinear length, $I_0=n_0 c|A_0|^2/2\pi$ is the peak input intensity, f is the fraction of the Raman contribution to the nonlinear refractive index, and $R(\tau)=\left\{\left[1+(\Omega_R\tau_R)^2\right]/\Omega_R\tau_R\right\}\exp(-\tau/\tau_R)\sin(\Omega_R\tau)$ is the Raman response function, τ_R is the characteristic Raman response time, Ω_R is the characterstic Raman frequency, and $\gamma_R=\tau_p/\tau_R$. For example, for fused silica $f=0.15$, $\tau_R=50$ fs, and $\Omega_R\tau_R=4.2$. For a noble gas such as Xe, there is no Raman contribution and $f=0$.

Self Focusing, Supercontinuum Generation, and Filamentation

The presence of the nonlinear refractive index with $n_2>0$ can lead to self-focusing of a laser beam. For sufficiently long pulses such that dispersion, self-steepening, and space–time focusing effects can be neglected, it is found that laser beams with input powers greater than the critical power $P_{\mathrm{cr}}=\alpha\lambda^2/4\pi n_0 n_2$ will undergo catastrophic self-focusing collapse. The dimensionless parameter $\alpha\geq 1.86$ depends on the spatial profile of the input beam and for a Gaussian input beam is given by $\alpha=1.9$, in which case the ratio of the input power P to the critical power satisfies the relation $P/P_{\mathrm{cr}}=1.055L_{\mathrm{df}}/L_{\mathrm{nl}}$. Extensive studies have been made on the dynamics of laser beams undergoing self-focusing. For example, it has been shown that the shape of the collapsing beam evolves to a radially symmetric profile as it approaches the collapse point and that the total power contained in the collapsing portion always corresponds to the minimum value (i.e., $\alpha\sim 1.86$) regardless of the initial power in the beam.

For light pulses shorter than a picosecond, the role of material dispersion plays an important role and completely alters the dynamics of the self-focusing process. These dispersive effects lead to a temporal splitting of the pulse into two pulses and the arrest of its collapse. At even higher powers, other phenomena can occur, such as 'optical shock' formation at the rear edge of the pulse, due to self-steepening and space–time focusing. Shock formation leads to the emission of a broad spectrum of radiation extending from the ultraviolet to the mid-infrared, known as supercontinuum generation (SCG). This phenomenon was first observed in 1970, and since then it has been observed in many different solids, liquids, and gases, under a wide variety of experimental conditions.

If the peak intensity of the pulse approaches intensities of $<10^{13}$ W/cm^2, either through self-focusing or external focusing, multiphoton ionization occurs and a plasma is formed in the medium. (See section below on Plasma Nonlinearities and Relativistic Effects.) The generation of the plasma lowers the refractive index and effectively counteracts the self-focusing process, resulting in the spatial confinement of the pulse for distances far beyond what would be allowed by ordinary linear diffraction and has been observed in gases, liquids, and solids. A striking example of this apparent self-waveguiding is the observation of 'light strings' in air which can extend more than 10 km into the atmosphere. This phenomenon was first observed with 100 fs laser pulses in the near infrared ($\lambda=800$ nm). Researchers found that pulses with energies greater than 10 mJ undergo self-focusing collapse in air and produce a highly intense ($>10^{13}$ W/cm^2) 100-micron-diameter light filament tens of meters long.

Multiphoton Absorption

Multiphoton absorption is a process in which an atom or molecule makes a transition from a ground state to an excited state by means of the simultaneous absorption of N photons. In the lowest order of perturbation theory, such a process can be described by means of a susceptibility of order $(2N-1)$, that is, by $\chi^{(2N-1)}$. Alternatively, this process can be described in terms of an N-photon cross-section $\sigma^{(N)}$ defined such that the transition rate per atom is given by

$$R^{(N)}=\sigma^{(N)}I^N \tag{6}$$

where I is the intensity of the laser field. Quantum mechanical expressions for the N-photon cross-sections are readily obtained. One finds, for instance, that

$$R_{\mathrm{ng}}^{(2)}=\left|\sum_m\frac{\mu_{\mathrm{nm}}\mu_{\mathrm{mg}}E^2}{\hbar^2(\omega_{\mathrm{mg}}-\omega)}\right|^2 2\pi\rho_{\mathrm{f}}(\omega_{\mathrm{ng}}-2\omega)$$

$$R_{\mathrm{og}}^{(3)}=\left|\sum_{mn}\frac{\mu_{\mathrm{on}}\mu_{\mathrm{nm}}\mu_{\mathrm{mg}}E^3}{\hbar^3(\omega_{\mathrm{ng}}-2\omega)(\omega_{\mathrm{mg}}-\omega)}\right|^2 2\pi\rho_{\mathrm{f}}(\omega_{\mathrm{og}}-3\omega) \tag{7}$$

In each of these expressions, the quantity ρ_{f} represents the density of final states, or equivalently the atomic lineshape function, evaluated at the N-photon transition frequency.

Optical Damage

High-intensity laser fields can produce unwanted damage to optical materials. As a point of reference, the threshold for laser damage to fused silica at a wavelength of 1.05 micrometers for a pulse of 30 ps duration corresponds to an intensity of 230 GW/cm^2 or a fluence of 7 J/cm^2. Over a wide range of pulselengths (approximately 1 ps to 1 μs), the threshold intensity for laser damage decreases with pulse length T as $T^{-1/2}$ and correspondingly the threshold fluence for laser damage increases with pulse length $T^{1/2}$. In this range of pulse durations, the dominant mechanism of laser damage is avalanche breakdown. In this process, free electrons are accelerated by the laser field until they acquire sufficient energy to impact-ionize other atoms in the sample. These additional electrons are similarly accelerated and create still more free electrons. The combined action of the breaking of chemical bonds and the deposition of heat energy leads to the fracturing of the optical material. For pulses shorter than 1 ps, processes such as multiphoton absorption and multiphoton dissociation contribute to the mechanism of optical damage. For laser pulses longer than approximately 1 μs (including continuous wave laser beams), the dominant damage mechanism is direct heating of the optical material by linear absorption.

High-Harmonic Generation

Let us consider how nonlinear optical effects are modified when excited by a super-intense pulse. Nonlinear optical effects are traditionally modeled using a power-series expansion, such as

$$P = \chi^{(1)}E + \chi^{(2)}E^2 + \chi^{(3)}E^3 + \ldots \tag{8}$$

but this series is not expected to converge if the laser field strength E exceeds the atomic unit of field strength $E_{\text{at}} = e/a_0^2 = 2 \times 10^7$ statvolt/cm $= 6 \times 10^9$ V/cm. This field strength corresponds to a laser intensity of $I_{\text{at}} = 4 \times 10^{16}$ W/cm^2, which constitutes the threshold intensity for exciting nonperturbative nonlinear optical response.

One of the consequences of excitation with intensities comparable to the atomic unit of intensity I_{at}, is the occurrence of high-harmonic generation. In a typical experimental arrangement, a gas jet is irradiated by high-intensity laser radiation, and all odd harmonics of the fundamental laser frequency, up to some maximum value N_{max}, are observed. The various harmonics below N_{max} are typically emitted with approximately equal intensity; such an observation is incompatible with a perturbative explanation of this phenomenon. Recent work has demonstrated harmonic generation with N_{max} as large as 341.

The phenomenon of high-harmonic generation can be understood in terms of a simple physical model. One imagines an atomic electron that has received kinetic energy from the laser field and is excited to a highly elliptical orbit. The positively charged atomic nucleus is at one focus of this ellipse, and each time the electron passes near the nucleus it undergoes strong acceleration and emits a short pulse of radiation. This radiation will occur in the form of a train of short pulses; the spectrum of the radiation will be the square of the Fourier transform of this pulse train, which will contain the odd harmonics of the oscillation frequency up to some maximum frequency, that is approximately the inverse of the time the electron spends near the atomic core. This argument can be made quantitative to show that the maximum harmonic number is given by

$$N_{\text{max}}\hbar\omega = 3.17K + U_p \tag{9}$$

where $K = e^2E^2/m\omega^2$ is the 'ponderomotive energy' (the kinetic energy of an electron in a laser field) and U_p is the ionization energy of the atom.

Plasma Nonlinearities and Relativistic Effects

The process of multiphoton ionization can liberate a sufficient number of free electrons to transform the optical medium into a plasma, that is, a fully or partially ionized gas. The process of plasma formation is described by the equation

$$\frac{dN_e}{dt} = \frac{dN_i}{dt} = (N_T - N_i)\sigma^{(N)}I^N - rN_eN_i \tag{10}$$

where N_e is the number density of electons, N_i is the number density of ions, N_T is the total number of atoms (both ionized and un-ionized) in the material, and r is the electron–ion recombination coefficient. The optical properties of plasmas are very different from those of typical dielectric materials; the plasma contribution to the dielectric constant is given by

$$\varepsilon(\omega) = 1 - \frac{\omega_p^2}{\omega^2} \tag{11}$$

where $\omega_p = \sqrt{4\pi Ne^2}/m$ is known as the plasma frequency.

Nonlinear effects can occur in the propagation of light through a plasma. One example is the nonlinear response resulting from the relativistic change in mass of the electron due to the large velocity that it can attain in the field of an intense

laser beam. Detailed consideration of this effect shows that the nonlinear change in refractive index can be described as $\Delta n = n_2 I$ where

$$n_2 = \frac{2\pi\omega_p^2 e^2}{n_0^2 m^2 c^3 \omega^4} = \frac{1}{2\pi n_0^2}\left(\frac{\omega_p}{\omega}\right)^2 1.1 \times 10^{-26}\,\frac{\text{cm}^2}{\text{W}} \tag{12}$$

Nonlinear Quantum Electrodynamics

One can imagine an electric field so intense that it could lead to the spontaneous creation of electron–positron pairs. Such a field would have a strength of the order of $E_{\text{QED}} = mc^2/e\lambda\!\!\!^{-}_c$ where $\lambda\!\!\!^{-}_c = \hbar/mc$ is the reduced Compton wavelength of the electron. The intensity of a light beam with a peak field amplitude of E_{QED} is $I_{\text{QED}} = 4 \times 10^{29}$ W/cm^2. Detailed consideration shows that even for fields weaker than I_{QED} there will be a field-induced change in the dielectric tensor given by

$$\varepsilon_{ik} = \delta_{ik} + \frac{e^4\hbar}{45\pi m^4 c^7}\left[2(E^2 - B^2)\delta_{ik} + 7B_i B_k\right] \tag{13}$$

Because of the unusual tensor properties of this relation, it displays a different polarization dependence than typical optical nonlinearities. Nonetheless, to an order of magnitude one can describe the strength of this response as

$$n_2 = \frac{7}{15c}\frac{e^2}{\hbar c}\frac{1}{E_{\text{QED}}^2} = 5.6 \times 10^{-34}\text{cm}^2/\text{W} \tag{14}$$

See also: Attosecond Spectroscopy

Further Reading

Agostini, P., Fabre, F., Mainfray, G., Petite, G., Rahman, N.K., 1979. Free-free transitions following six photon ionization of xenon. Physics Review Letters 42, 1127.
Alfano, R.R. (Ed.), 1989. The Supercontinuum Laser Source. New York: Springer-Verlag.
Bloembergen, N., 1997. A brief history of light breakdown. Journal of Nonlinear Optical Physics and Materials 6, 377.
Boyd, R.W., 2003. Nonlinear Optics, Second Edition. Section 12.5. Amsterdam: Academic Press.
Braun, A., Korn, G., Liu, X., Du, D., Squier, J., Mourou, G., 1995. Self-channeling of high-peak-power femtosecond laser pulses in air. Optical Letters 20, 73.
Brabec, T., Krausz, F., 2000. Intense few-cycle laser fields: Frontiers of nonlinear optics. Reviews of Modern Physics 72, 545.
Chang, Z., Rundquist, A., Wang, H., Murnane, M.M., Kapteyn, H.C., 1997. Generation of coherent soft X-rays at 2.7nm using high harmonics. Physics Review Letters 79, 2967.
Corkum, P.B., 1993. Plasma perspective on strong field multiphoton ionization. Physics Review Letters 71, 1994.
Corkum, P.B., Rolland, C., Srinivasanrao, T., 1986. Supercontinum generation in gases. Physics Review Letters 57, 2268.
Gaeta, A.L., 2003. Collapsing light really shines. Science 301, 54.
Kasparian, J., Rodriguez, M., Mejean, G., Yu, J., Salmon, E., Wille, H., Bourayou, R., Frey, S., Andŕe, Y.-B., Mysyrowicz, A., Sauerbrey, R., Wolf, J.-P., Woste, L., 2003. White light filaments for atmospheric analysis. Science 301, 61.
Lewenstein, M., Balcou, P., Ivanov, M.Y., L'Huillier, A., Corkum, P.B., 1994. Theory of high-harmonic generation by low-frequency laser fields. Physics Review A 49, 2117.
Max, C.E., Arons, J., Langdon, A.B., 1974. Self-modulation and self-focusing of electromagnetic waves in plasmas. Physics Review Letters 33, 209.
Moll, K.D., Fibich, G., Gaeta, A.L., 2003. Self-similar wave collapse: observation of the Townes profile. Physics Review Letters 90, 203902.
Monot, P., Auguste, T., Gibbon, P., Jakober, F., Mainfray, G., Dulieu, A., Louis-Jacquet, M., Malka, G., Miquel, J.L., 1995. Experimental demonstration of relativistic self-channeling of a multiterawatt laser pulse in an underdense plasma. Physics Review Letters 74, 2953.
Rairoux, P., Schillinger, H., Niedermeier, S., *et al.*, 2000. Applied Physics B—Lasers Optics 71, 573.
Ranka, J.K., Schirmer, R.W., Gaeta, A.L., 1996. Observation of pulse splitting in nonlinear dispersive media. Physics Review Letters 77, 3783.
Rothenberg, J.E., 1992. Pulse splitting during self-focusing in normally dispersive media. Optical Letters 17, 583.
Sprangle, P., Tang, C.-M., Esarey, E., 1987. Relativistic self-focusing of short-pulse radiation beams in plasmas. IEEE Transactions on Plasma Science 15, 145.
Stuart, B.C., Feit, D., Rubenchik, A.M., Shore, B.W., Perry, M.D., 1995. Laser-induced damage in dielectrics with nanosecond to subpicosecond pulses. Physics Review Letters 74, 2248.
Wagner, R., Chen, S.-Y., Maksemchak, A., Umstadter, D., 1997. Electron acceleration by a laser wakefield in a relativistically self guided channel. Physics Review Letters 78, 3125.
Wood, R.M., 1986. Laser Damage in Optical Materials. Bristol, UK: Adam Hilger.

Electromagnetically Induced Transparency

JP Marangos, Imperial College of Science, Technology and Medicine, London, UK

Introduction

In this article we examine how the process of electromagnetically induced transparency (EIT) can be used to increase the efficiency of non-linear optical frequency mixing schemes. We illustrate the basic ideas by concentrating upon the enhancement of resonant four-wave mixing schemes in atomic gases. To start we introduce the quantum interference phenomenon that gives rise to EIT. The calculation of EIT effects in a three-level atomic medium using a density matrix approach is presented. Next we examine the modified susceptibilities that result and that can be used to describe nonlinear mixing, and show how large enhancements in nonlinear conversion efficiency can result. Specific examples are briefly discussed along with a further discussion of the novel aspects of refractive index and pulse propagation modifications that result from EIT. The potential benefits to nonlinear optics of electromagnetically induced transparency are discussed. In an article of this nature it is not possible to credit all of the important contributions to this field, but we include a bibliography for further reading indicating some of the seminal contributions.

Electromagnetically induced transparency is the term applied to the process by which an otherwise opaque medium can be rendered transparent through laser-induced quantum mechanical interference processes. The linear susceptibility is substantially modified by EIT. Absorption is cancelled due to destructive interference between excitation pathways. The dispersion of the medium is likewise modified such that where there is no absorption the refractive index is unity and there can be a large value of normal dispersion leading to low values of group velocity. In contrast to the linear susceptibility the nonlinear susceptibility involving these same resonant laser fields will undergo constructive rather than destructive interference. This leads to a strong enhancement in nonlinear frequency mixing. For a normally transparent medium the nonlinear optical couplings will be small unless large-amplitude fields are applied. The most important consequence of EIT, in contrast to the usual case when a medium is transparent, is that the dispersion is not vanishing and the nonlinear couplings can be very large.

To understand how this comes about we must consider the system illustrated in **Fig. 1(a)** where because of the resonant condition satisfied for the applied fields the atom can be simplified to just three-levels. The important parameters in this model system are that state $|2>$ is metastable and has a dipole-allowed coupling to state $|3>$. The coupling between states $|1>$ and $|3>$ is also dipole-allowed and as a consequence state $|3>$ can decay radiatively to either $|1>$ or $|2>$. Let a coupling field be applied resonantly to the $|2> - |3>$ transition, which may be cw or pulsed but should in the latter case be transform-limited. We define the Rabi coupling as $\Omega_{23} = \frac{\mu_{23}E}{\hbar}$ where E is the laser electric field amplitude and μ_{23} is the dipole moment of the transition. The Rabi frequency needs to be larger than the inhomogeneous broadening of the sample. A second field, the probe, typically of much lower intensity, is then applied in resonance with the $|1> - |3>$ transition. If the condition $\Omega_{23} \gg \Omega_{13}$ is satisfied then it is convenient to replace the bare atomic states $|2>$ and $|3>$ with the dressed states (see **Fig. 1(b)**):

$$|a> = \frac{1}{\sqrt{2}}[|2> + |3>] \tag{1a}$$

$$|b> = \frac{1}{\sqrt{2}}[|2> - |3>] \tag{1b}$$

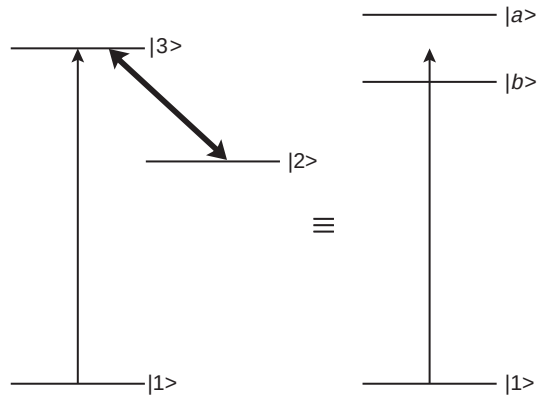


Fig. 1 A three-level atomic system coupled to laser fields in the lambda configuration is shown on the left-hand side. In the limit of a strong coupling field this is equivalent to the dressed states $|a>$ and $|b>$ coupled to the ground state by the probe field alone as shown on the right-hand side.

Transitions from state $|1>$ induced by the probe field to this pair of near-resonant dressed states are subject to exact cancellation at resonance if $|2>$ is perfectly metastable. This is because the only contribution to the excitation amplitude comes from the matrix elements $<1|\mathbf{r}\cdot\mathbf{E}|3>$ as the $|1>-|2>$ transition is dipole forbidden. The contributions from the equally detuned states $|a>$ and $|b>$ thus enter into the overall amplitude with opposite signs and equal magnitude as can be seen by inspection of Eqs. (1a) and (1b). This leads to cancellation of the absorption amplitude. This type of absorption cancellation is well known and closely related to the so-called *dark states*.

Theoretical Treatment of EIT in a Three-Level Medium

It was realized by several workers in the 1970s that laser-induced interference effects could lead to a cancellation of absorption at certain frequencies. To gain a more quantitative understanding of the effects of the coupling field upon the optical properties of a dense ensemble of three-level atoms we require a treatment that computes the optical susceptibility of the medium. A treatment originally carried out by Harris *et al.* for a Λ scheme similar to that illustrated in Fig. 1 was the first to derive the modified susceptibilities that will be discussed below. In that treatment the state amplitudes in the three-level atom were solved in the steady-state limit and from these the linear and nonlinear susceptibilities (see below) are then derived. In what follows we adopt a density matrix treatment as employed by various workers. This readily allows the inclusion of dephasing as well as population decay terms. The critical parameters in this system are the strengths of the fields (described in terms of the Rabi couplings), the detuning of the fields from resonance $\Delta\omega_{13}=\omega_{13}-\omega_P$ and $\Delta\omega_{23}=\omega_{23}-\omega_C$ (see Fig. 2), the radiative decays from $|3>$ to $|1>$ and $|2>$, γ_c and γ_d and from $|2>-|1>\gamma_a$ (although the latter is anticipated to be smaller). Extension to other configurations of the three states, such as a V or ladder scheme is implicit within this general treatment.

Fig. 2 illustrates the system considered. This treatment will also address the nonlinear response in a four-wave mixing scheme completed by the two-photon resonant coupling applied between state $|1>$ and $|2>$.

We write the interaction Hamiltonian as:

$$H = H_0 + V \tag{2}$$

where H_0 is the unperturbed Hamiltonian of the system and is written as

$$H_0 = \hbar\omega_1|1\rangle\langle 1| + \hbar\omega_2|2\rangle\langle 2| + \hbar\omega_3|3\rangle\langle 3| \tag{3}$$

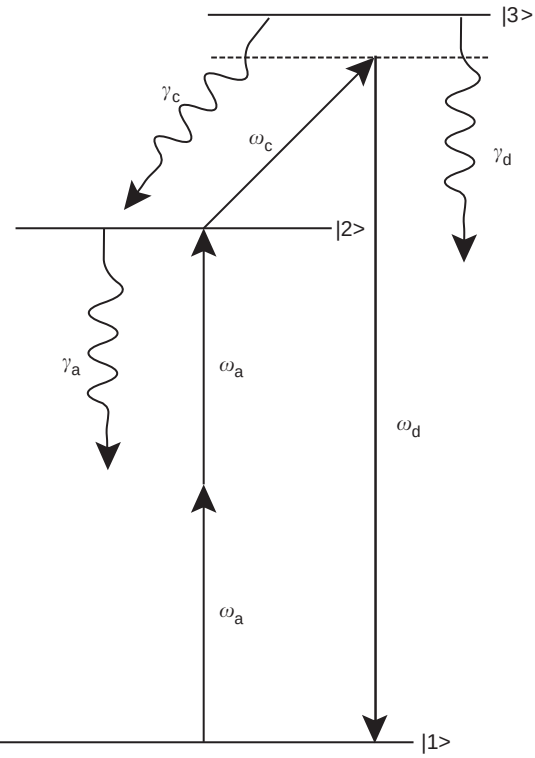


Fig. 2 A four-wave mixing scheme incorporating the three-level lambda system in which electromagnetically induced transparency has been created for the generated field ω_d by the coupling field ω_c. The decay rates γ_i from the three atomic levels are also shown. For a full explanation of this system see text.

and V is the interaction Hamiltonian and can be expressed

$$V = \hbar\Omega_a e^{-i\omega_a t}|2\rangle\langle 1| + \hbar\Omega_c e^{-i\omega_c t}|2\rangle\langle 3| + \hbar\Omega_d e^{-i\omega_d t}|3\rangle\langle 1| + c.c \tag{4}$$

Note that the Rabi frequencies in the equation can be described as $\hbar\Omega_{ij} = \mu_{ij}|E(\omega_{ij})|$, where μ_{ij} is the dipole moment of the transition between states |i> and |j>, the Rabi frequency Ω_a is a two-photon Rabi frequency that characterizes the coupling between the laser field a and the atom for this two-photon transition. We have assumed for simplicity that $\omega_a = \omega_b = \omega_{12}/2$.

Assuming all the fields lie close to the resonance, the rotating wave approximation can be applied to the interaction picture and the interaction Hamiltonian V^I is given as

$$V^I = \hbar\Omega_a e^{i\Delta_a t}|2\rangle\langle 1| + \hbar\Omega_c e^{i\Delta_c t}|2\rangle\langle 3| + \hbar\Omega_d e^{i\Delta_d t}|3\rangle\langle 1| + c.c \tag{5}$$

where Δ_a, Δ_c and Δ_d refer the detunings of the fields and can be written as:

$$\begin{aligned}\Delta_a &= \omega_{12} = 2\omega_a \\ \Delta_c &= \omega_{32} - \omega_c \\ \Delta_d &= \omega_{13} - \omega_d\end{aligned} \tag{6}$$

For the evaluation of the density matrix with this interaction V^I, the Schrödinger equation can be restated in terms of the density matrix components. This form is called the Liouville equation and can be written as:

$$\hbar\frac{\partial}{\partial t}\varrho_{ij}(t) = -i\sum_k H_{ik}(t)\varrho_{ij}(t) + i\sum_k \varrho_{ij}(t)H_{ik}(t) + \Gamma_{ij} \tag{7}$$

where Γ_{ij} represents phenomenologically added decay terms (i.e. spontaneous decays, collisional broadening, etc.). This formalism leads to a set of nine differential equations for nine different density matrix elements that describe the three-level system.

To remove the optical frequency oscillations, a coordinate transform is needed and to incorporate the relevant oscillatory detuning terms into the off-diagonal elements we make the substitution:

$$\begin{aligned}\tilde{\varrho}_{12} &= \varrho_{12}e^{-i\Delta_a t} \\ \tilde{\varrho}_{23} &= \varrho_{23}e^{-i\Delta_c t} \\ \tilde{\varrho}_{31} &= \varrho_{31}e^{-i\Delta_d t}\end{aligned} \tag{8}$$

This operation conveniently eliminates the time dependencies in the rate equations and the equations of motion for the density matrix are given by:

$$\begin{aligned}\frac{\partial}{\partial t}\varrho_{11} &= i\frac{1}{2}\Omega_a\tilde{\varrho}_{12} + i\frac{1}{2}\Omega_d\tilde{\varrho}_{13} - i\frac{1}{2}\Omega_a^*\tilde{\varrho}_{21} - i\frac{1}{2}\Omega_d^*\tilde{\varrho}_{31} + \Gamma_{11} \\ \frac{\partial}{\partial t}\varrho_{22} &= i\frac{1}{2}\Omega_c^*\tilde{\varrho}_{23} + i\frac{1}{2}\Omega_a^*\tilde{\varrho}_{21} - i\frac{1}{2}\Omega_c\tilde{\varrho}_{32} - i\frac{1}{2}\Omega_a\tilde{\varrho}_{12} + \Gamma_{22} \\ \frac{\partial}{\partial t}\varrho_{33} &= i\frac{1}{2}\Omega_c^*\tilde{\varrho}_{23} - i\frac{1}{2}\Omega_d^*\tilde{\varrho}_{31} + i\frac{1}{2}\Omega_d\tilde{\varrho}_{13} - i\frac{1}{2}\Omega_c\tilde{\varrho}_{32} + \Gamma_{33} \\ \frac{\partial}{\partial t}\varrho_{23} &= i\frac{1}{2}\Omega_c\tilde{\varrho}_{22} - i\frac{1}{2}\Omega_c\tilde{\varrho}_{33} - i\Delta_a\tilde{\varrho}_{23} + i\frac{1}{2}\Omega_d\tilde{\varrho}_{21} - i\frac{1}{2}\Omega_c\tilde{\varrho}_{13} + \Gamma_{23} \\ \frac{\partial}{\partial t}\varrho_{21} &= i\Omega_a\tilde{\varrho}_{22} + \Omega_a\tilde{\varrho}_{33} + i\Omega_d\tilde{\varrho}_{23} - i\Delta_c\tilde{\varrho}_{21} - i\frac{1}{2}\Omega_c\tilde{\varrho}_{31} - i\frac{1}{2}\Omega_a\tilde{\varrho}_{31} + \Gamma_{21} \\ \frac{\partial}{\partial t}\varrho_{31} &= i\frac{1}{2}\Omega_d\tilde{\varrho}_{22} - i\Omega_d\tilde{\varrho}_{33} + i\Omega_d\tilde{\varrho}_{23} - i\Delta_c\tilde{\varrho}_{21} - i\frac{1}{2}\Omega_c\tilde{\varrho}_{31} - i\frac{1}{2}\Omega_a\tilde{\varrho}_{31} + \Gamma_{31}\end{aligned} \tag{9}$$

Using the set of equations above and the relation $\tilde{\varrho}_{ij} = \tilde{\varrho}_{ij}^*$we obtain equations for $\tilde{\varrho}_{12}$, $\tilde{\varrho}_{32}$, and $\tilde{\varrho}_{13}$. For the incoherent population relaxation the decays can be written:

$$\begin{aligned}\Gamma_{11} &= \gamma_a\varrho_{22} + \gamma_d\varrho_{33} \\ \Gamma_{22} &= -\gamma_a\varrho_{22} + \gamma_c\varrho_{33} \\ \Gamma_{33} &= -(\gamma_c + \gamma_d)\varrho_{33}\end{aligned} \tag{10}$$

and for the coherence damping:

$$\begin{aligned}\Gamma_{21} &= -\left\{\frac{1}{2}[\gamma_a + \gamma_c] + \gamma_{21}^{col}\right\}\varrho_{21} \\ \Gamma_{23} &= -\left\{\frac{1}{2}[\gamma_a + \gamma_c + \gamma_d] + \gamma_{23}^{col}\right\}\varrho_{23} \\ \Gamma_{31} &= -\left\{\frac{1}{2}[\gamma_d + \gamma_c] + \gamma_{31}^{col}\right\}\varrho_{31}\end{aligned} \tag{11}$$

where γ_{ij}^{col} represents collisional dephasing terms which may be present.

This system of equations can be solved by various analytical or numerical methods to give the individual density matrix elements. Analytical solutions are possible if we can assume for instance that $\Omega_c \gg \Omega_a, \Omega_d$ and that there are continuous-wave fields so a steady-state treatment is valid. For pulsed fields or in the case where the generated field may be significant numerical solutions are in general required.

We are specifically interested in the optical response to a probe field at ω_d (close to resonance with the $|1> - |3>$ transition) that is governed by the magnitude of the coherence ϱ_{13}. We find ϱ_{13} making the assumption that only the coupling field is strong, i.e., that $\Omega_c \gg \Omega_a, \Omega_d$ holds. From this quantity the macroscopic polarization is obtained from which the susceptibility can be computed. The macroscopic polarization P at the transition frequency ω_{13} can be related to the microscopic coherence ϱ_{13} via the expression:

$$P_{13} = N\mu_{13}\varrho_{13} \tag{12}$$

where N is the number of equivalent atoms in the ground state within the medium, and μ_{13} is the dipole matrix element associated with the transition. In this way real and imaginary parts of the linear susceptibility χ at frequency ω can be directly related to ϱ_{13} via the macroscopic polarization since the latter can be defined as:

$$P_{13}(\omega) = \varepsilon_0 \chi(\omega) E \tag{13}$$

where E is the electric field amplitude at frequency ω_d. The linear susceptibility (real and imaginary parts) is given by the following expressions:

$$\mathrm{Re}\chi_D^{(1)}(-\omega_d, \omega_d) = \frac{|\mu_{13}|^2 N}{\varepsilon_0 \hbar}\left[\frac{-4\Delta_{21}(|\Omega_2|^2 - 4\Delta_{21}\Delta_{32}) + 4\Delta_{31}\gamma_a^2}{(4\Delta_{31}\Delta_{21} - \gamma_d\gamma_a - |\Omega_2|^2)^2 + 4(\gamma_d\Delta_{21} + \gamma_a\Delta_{31})^2}\right] \tag{14}$$

$$\mathrm{Im}\chi_D^{(1)}(-\omega_d, \omega_d) = \frac{|\mu_{13}|^2 N}{\varepsilon_0 \hbar}\left[\frac{8\Delta_{21}^2\gamma_d + 2\gamma_a(|\Omega_2|^2 + \gamma_d\gamma_a)}{(4\Delta_{31}\Delta_{21} - \gamma_d\gamma_a - |\Omega_2|^2)^2 + 4(\gamma_d\Delta_{21} + \gamma_a\Delta_{31})^2}\right] \tag{15}$$

We now consider the additional two-photon resonant field ω_a. This leads to a four-wave mixing process that generates a field at ω_d (i.e., the probe frequency). The nonlinear susceptibility that describes the coupling of the fields in a four-wave mixing process is given by the expression:

$$\chi_D^{(3)}(-\omega_d, \omega_a, \omega_a, \omega_c) = \frac{\mu_{23}\mu_{13}N}{6\varepsilon_0\hbar^3((\Delta_{21} - j\gamma_a/2)(\Delta_{31} - j\gamma_d/2) - (|\Omega_c|^2/4))^*}\sum_i \mu_{i1}\mu_{i3}\left(\frac{1}{\omega_i - \omega_a} + \frac{1}{\omega_i - \omega_b}\right)$$

Nonlinear Optical Processes

The dependence of the susceptibilities upon the detuning is plotted in **Fig. 3**. In this plot the effects of inhomogeneous (Doppler) broadening are also incorporated by computing the susceptibilities over the inhomogeneous profile (see below for more details).

By inspection of **Eq. (14)** we see that the absorptive loss at the minimum of $\mathrm{Im}[\chi^{(1)}]$ varies as γ_a/Ω_c^2. This dependence is a consequence of the interference discussed above. In the absence of interference (i.e., just simply Autler–Townes splitting) the minimum loss would vary as γ_d/Ω_c^2. Since $|3> - |1>$ is an allowed decay channel (in contrast to $|2> - |1>$) it follows that $\gamma_b \gg \gamma_a$ and so the absorption is much less as a consequence of EIT. $\mathrm{Re}[\chi^{(1)}]$ in **Eq. (15)** and **Fig. 3** is also modified significantly. The resonant value of refractive index is equal to the vacuum value where the absorption is a minimum, the dispersion is normal in this region with a gradient determined by the strength of the coupling laser, a point we will return to shortly. For an unmodified system the refractive index is also unity at resonance but in that case there is high absorption and steep anomalous absorption. Reduced group velocities result from the steep normal dispersion that accompanies EIT. Inspection of the expression **(16)** and **Fig. 3** shows that $\chi^{(3)}$ is also modified by the coupling field. The nonlinear susceptibility depends upon $1/\Omega_c^2$ as is expected for a laser dressed system, however in this case there is not destructive but rather constructive interference, between the field-split components. This result is of great importance: it ensures that the absorption can be minimized at frequencies where the nonlinear absorption remains large.

As a consequence of constructive interference the nonlinear susceptibility remains resonantly enhanced whilst the linear susceptibility vanishes or becomes very small at resonance due to destructive interference. Large nonlinearity accompanying vanishing absorption (transparency) of course match conditions for efficient frequency mixing as a large atomic density can then be used. Moreover the dispersion (controlling the refractive index) also vanishes at resonance; this leads to perfect phase matching (i.e., zero wavevector mismatch between the fields) in the limit of a simple three-level system. As a result of these features a large enhancement of the conversion efficiency in this type of scheme can be achieved.

To compute the generated field strength Maxwell's equations must be solved using these expression for the susceptibility to describe the absorption, refraction, and nonlinear coupling in the medium. We will treat this within the slowly varying envelope approximation since the fields are cw or nanosecond pulses. To be specific we assume that the susceptibilities are time independent, i.e., that we are in the steady-state (cw) limit. We make the assumptions also that there is no pump field absorption and

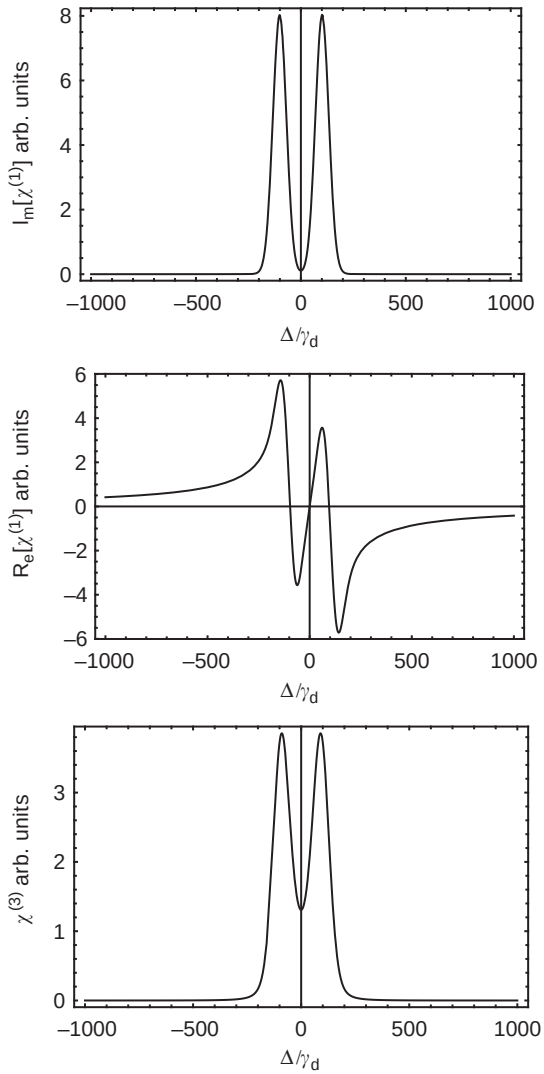


Fig. 3 Electromagnetically induced transparency is shown in the case of significant Doppler broadening. The Doppler averaged values of the imaginary Im[$\chi^{(1)}$] and real Re[$\chi^{(1)}$] parts of the linear susceptibility and the nonlinear susceptibility $\chi^{(3)}$ are plotted in the three frames. The Doppler width is taken to be $20\gamma_d$, $\Omega_c = 100\gamma_d$ and $\gamma_d = 50\gamma_a$, $\gamma_c = 0$.

that we have plane waves. Under these assumptions the generated field amplitude A_d is given in terms of the other field amplitudes A_i by:

$$\frac{\partial}{\partial z}A_{\mathrm{d}} = \mathrm{i}\frac{\omega_{\mathrm{d}}}{4c}\chi^{(3)}A_{\mathrm{a}}^2 A_{\mathrm{c}}\mathrm{e}^{-\mathrm{i}\Delta k_{\mathrm{d}}z} - \frac{\omega_{\mathrm{d}}}{2c}\mathrm{Im}[\chi^{(1)}]A_{\mathrm{d}} + \mathrm{i}\frac{\omega_{\mathrm{d}}}{2c}\mathrm{Re}[\chi^{(1)}]A_{\mathrm{d}} \tag{16}$$

where the wavevector mismatch is given by:

$$\Delta k_{\mathrm{d}} = k_{\mathrm{d}} + k_{\mathrm{c}} - 2k_{\mathrm{a}} \tag{17}$$

The wavevector mismatch will be zero on resonance for the three-level atom considered in this treatment. In fact the contribution to the refraction from all the other atomic levels must be taken into account whilst computing Δk_{d} and it is implicit that these make a finite contribution to the wavevector mismatch.

We can solve this first-order differential equation with the boundary condition $A_{\mathrm{d}}(z=0)=0$ to find the generated intensity $I(\omega_{\mathrm{d}})$ after a length z:

$$I(\omega_{\mathrm{d}}) = \frac{3n\omega_{\mathrm{d}}^2}{8Z_0c^2}|\chi^{(3)}|^2|A_{\mathrm{a}}|^4|A_{\mathrm{c}}|^2\frac{\left\{1+\mathrm{e}^{-\frac{\omega_{\mathrm{d}}}{c}\mathrm{Im}[\chi^{(1)}]z} - 2\mathrm{e}^{-\frac{\omega_{\mathrm{d}}}{c}\mathrm{Im}[\chi^{(1)}]z}\cos\left(\Delta k + \frac{\omega_{\mathrm{d}}}{2c}\mathrm{Re}[\chi^{(1)}]\right)z\right\}}{\frac{\omega_{\mathrm{d}}^2}{4c^2}\mathrm{Im}[\chi^{(1)}]^2 + \left\{\Delta k + \frac{\omega_{\mathrm{d}}}{2c}\mathrm{Re}[\chi^{(1)}]\right\}^2}$$

where Z_0 is the impedance of free space. This expression is quantitatively correct for the case of the assumptions made. More generally the qualitative predictions and general scaling remain valid in the limit where the pulse duration is significantly longer than the time required to traverse the medium. Note that both real and imaginary parts of $\chi^{(1)}$ and $\chi^{(3)}$ play an important role, as

we would expect for resonant frequency mixing. The influence of that part of the medium refraction which is due to other levels is contained in the terms with Δk. In the case of a completely transparent medium this becomes a major limit to the conversion efficiency.

Propagation and Wave-Mixing in a Doppler Broadened Medium

Doppler shifts arising from the Maxwellian velocity distribution of the atoms in the medium lead to a corresponding distribution in the detunings for the various members of the atomic ensemble. The response of the medium, as characterized by the susceptibilities, must therefore include the Doppler effect by performing a weighted sum over possible detunings. The weighting is determined by the Gaussian form of the Maxwellian velocity distribution. From this operation the effective values of the susceptibilities at a given frequency are obtained, and these quantities can be used to calculate the generated field. This step is of considerable practical importance as in most up-conversion schemes it is not possible to achieve Doppler-free geometries and the use of laser cooled atoms, although in principle possible, limits the atomic density that can be employed. The interference effects persist in the dressed profiles providing the coupling laser Rabi frequency is comparable to or larger than the inhomogeneous width. This is because the Doppler profile follows a Gaussian distribution which falls off much faster in the wings of the profile than the Lorentzian profile due to the natural broadening.

In the case considered with weak probe field, excited state populations and coherences remain small. The two-photon transition need not be strongly driven (i.e., a small two-photon Rabi frequency can be used) but a strong coupling laser is required. The coupling laser must be intense enough that its Rabi frequency is comparable to or exceeds the inhomogenous widths in the system (i.e., Doppler width), and a laser intensity of above 1 MW cm^{-2} is required for a typical transition. This is trivially achieved even for unfocused pulsed lasers, but does present a serious limit to cw lasers unless a specific Doppler-free configuration is employed. The latter is not normally suitable for a frequency up-conversion scheme if a significant up-conversion factor is required, e.g., to the vacuum ultraviolet (VUV); however recent experiments report significant progress in cw frequency up-conversion using EIT and likewise a number of other possibilities, e.g., laser-cooled atoms and standing-wave fields, have been proposed.

A transform-limited single-mode laser pulse is essential for the coupling laser field since a multimode field will cause an additional dephasing effect on the coherence, resulting in a deterioration of the quality of the interference. In contrast, whilst it is advantageous for the field driving the two-photon transition to be single mode (in order to achieve optimal temporal and spectral overlap with the EIT hole induced by the dressing laser), this is not essential since this field does not need to drive the coherence responsible for interference.

When a pulsed laser field is used additional issues must be considered. The group velocity is modified for pulses propagating in the EIT large reductions, e.g., by factors down to $<c/100$, in the group velocity have been observed. Another consideration beyond that found in the simple steady-state case is that the medium can only become transparent if the pulse contains enough energy to dress all the atoms in the interaction volume. The minimum pulse energy to prepare a transparency is:

$$E_{\text{preparation}} = \frac{f_{13}}{f_{23}} NL\hbar\omega \tag{18}$$

where f_{ij} are the oscillator strengths of the transitions and NL the product of the density and the length. Essentially the number of photons in the pulse must exceed the number of atoms in the interaction volume to ensure all atoms are in the appropriate dressed state. This puts additional constraints on the laser pulse parameters.

Up-conversion to the UV and vacuum UV has been enhanced by EIT in a number of experiments. Only pulsed fields have so far been up-converted to the VUV with EIT enhancement. The requirements on a minimum value of $\Omega_c > \Delta_{\text{Doppler}}$ constrains the conversion efficiency that can be achieved because the $1/\Omega_c^2$ factor in Eq. (17) ultimately leads to diminished values of $\chi^{(3)}$. The use of gases of higher atomic weight at low temperatures is therefore highly desirable in any experiment utilizing EIT for enhancement of four-wave mixing to the VUV. Conversion efficiencies, defined in terms of the pulse energies by E_d/E_a or E_d/E_c of a few percent have been achieved using the EIT enhancement technique. It is typically most beneficial to maximize the conversion efficiency defined by the first ratio since ω_a is normally in the UV and is the lower energy of the two applied pulses.

Nonlinear Optics with a Pair of Strong Coupling Fields in Raman Resonance

An important extension of the EIT concept occurs when two strong fields are applied in Raman resonance between a pair of states in a three-level system. Considering the system illustrated in Fig. 1 we can imagine that both applied fields are now strong. Under appropriate adiabatic conditions the system evolves to produce the maximum possible value for the coherence $\varrho_{12} = 0.5$. Adiabatic evolution into the maximally coherent state is achieved by adjusting either the Raman detuning or the pulse sequence (counter to-intuitive order). The pair of fields may also be in single-photon resonance with a third level, in which case the EIT-like elimination of absorption will be important. This situation is equivalent to the formation of a darkstate, since neither of the two strong fields is absorbed by the medium. For sufficiently strong fields the single-photon condition need not be satisfied and a maximum coherence will still be achieved.

An additional field applied to the medium can participate in sum- or difference-frequency mixing with the two Raman resonant fields. The importance of the large value of coherence is that it is the source polarization that drives the new fields generated in the frequency mixing process. Complete conversion can occur over a short distance that greatly alleviates the constraints usually set by phase-matching in nonlinear optics. Recently near unity conversion efficiencies to the far-UV were reported in an atomic lead system where maximum coherence had been created. In a molecular medium large coherence between vibrational or rotational levels has also been achieved using adiabatic pulse pairs. Efficient multi-order Raman sideband generation has been observed to occur. This latter observation may lead the way to synthesizing very short duration light pulses since the broadband Raman sideband spectrum has been proved to be phase-locked.

Pulse Propagation and Nonlinear Optics for Weak CW Fields

In a Doppler-free medium a new regime can be accessed. This is shown in Fig. 4 where the possibility now arises of extremely narrow transparency dips since very small values of Ω_c are now sufficient to induce EIT. The widths of these features are typically subnatural and are therefore accompanied by very steep normal dispersion, which corresponds to a much reduced group velocity. The ultraslow propagation of pulses is one of the most dramatic manifestations of EIT in this regime. Nonlinear susceptibilities are now very large as there is constructive interference controlling the value and the splitting of the two states is negligible compared to their natural width. Nonlinear optics at very low light levels, i.e., at the few-photon limit, is possible in this regime.

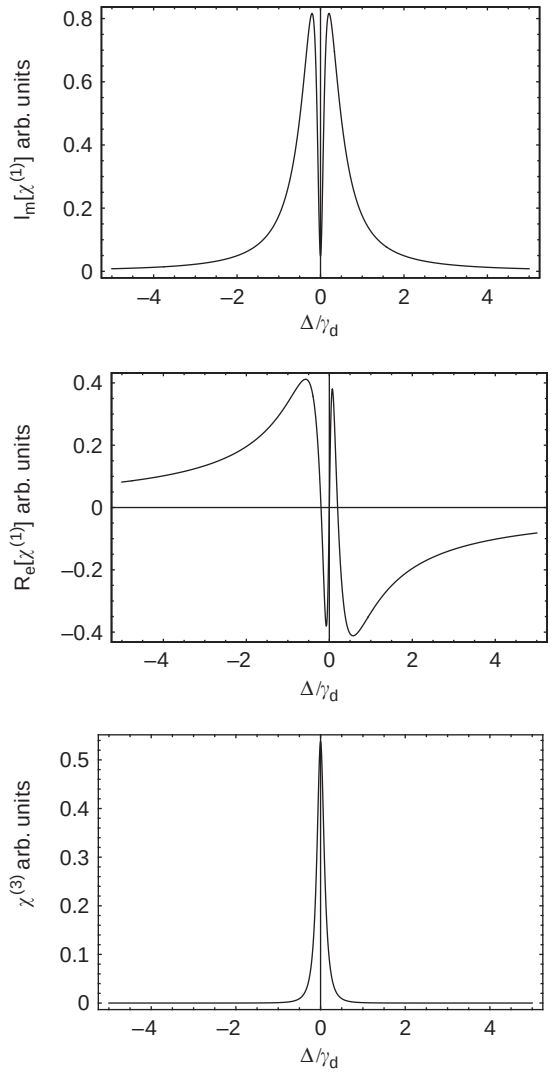


Fig. 4 Electromagnetically induced transparency is shown in the case where there is no significant Doppler broadening. The values of the imaginary $\mathrm{Im}[\chi^{(1)}]$ and real $\mathrm{Re}[\chi^{(1)}]$ parts of the linear susceptibility and the nonlinear susceptibility $\chi^{(3)}$ are plotted in the three frames. We take $\Omega_c = \gamma_d/5$ and $\gamma_d = 50\gamma_a$, $\gamma_c = 0$.

Propagation of pulses is significantly modified in the presence of EIT. Fig. 4 shows the changes to $\mathrm{Re}[\chi^{(1)}]$ that arise. Within the transparency dip there exists a region of steep normal dispersion. In the vicinity of resonance this is almost linear and it becomes reasonable to consider the leading term only that describes the group velocity. An analysis of the refractive changes has been provided by Harris who expanded the susceptibilities (both real and imaginary parts) of the dressed atom around the resonance frequency to determine various terms in $\mathrm{Re}[\chi^{(1)}]$. The first term of the series (zero order) $\mathrm{Re}[\chi^{(1)}](\omega_{13})=0$ corresponds to the vanishing dispersion at resonance. The next term $\partial[\mathrm{Re}\chi^{(1)}](\omega)/\partial\omega$ gives the slope of the dispersion curve; at ω_{13} this takes the value:

$$\frac{\partial \mathrm{Re}\chi^{(1)}(\omega_{13})}{\partial\omega}=\frac{|\mu_{13}|^2N4(\Omega_c^2-\gamma_a^2)}{2\pi\varepsilon_0\left(\Omega_c^2+\gamma_a\gamma_d\right)} \tag{19}$$

The slope of the dispersion profile leads to a reduced group velocity v_g:

$$\frac{1}{v_g}=\frac{1}{c}+\frac{\pi}{\lambda}\frac{\partial\chi^{(1)}}{\partial\omega} \tag{20}$$

From the expression for $\partial\chi/\partial\omega$ we see that this slope is steepest (and so v_g mimimum) in the case where $\Omega_c\gg\Gamma_2$ and $\Omega_c^2\gg\Gamma_2\,\Gamma_3$ but is still small compared to Γ_3 (i.e., $\Omega_C<\Gamma_3$) and hence $\partial\chi/\partial\omega\propto 1/\Omega_c^2$. In the limit of small Ω_c the following expression for v_g therefore holds:

$$v_g=\frac{\hbar c\varepsilon_0[\Omega_c]^2}{2\omega_d|\mu_{13}|^2N} \tag{21}$$

Extremely low group velocities, down to a few meters per second, are achieved in this limit using excitation of the hyperfine split ground states in either laser cooled atomic samples or Doppler-free configurations in finite temperature samples. Recently similar light slowing has been observed in solids. Storage of the optical pulse within the medium has also been achieved by adiabatically switching off the coupling field and thus trapping the optical excitation as an excitation within the hyperfine ground states for which the storage time can be very long (>1 ms) due to very low dephasing rates. Since the storage scenario should be valid even for single photons this process has attracted considerable attention recently as a means to enable quantum information storage.

Extremely low values of Ω_c are sufficient to induce complete transparency (albeit in a very narrow dip) and at this location the nonlinear response is resonantly enhanced. Very high efficiency nonlinear frequency mixing and phase conjugation at low light levels have been reported under these conditions. It is expected that high-efficiency nonlinear optical processes will persist to ultralow intensities (the few photon level) in an EIT medium of this type.

One example of highly enhanced nonlinear interactions is the predicted large values of the Kerr type nonlinearity (nonlinear refractive index). The origin of this can be seen by considering the steep dispersion profile in the region of the transparency dip in Fig. 4. Imagine that we apply an additional field ω_f, perhaps a very weak one, at a frequency close to resonance between state $|2>$ and a fourth level $|4>$. The ac Stark shift caused by this new field to the other three level, although small, will have a dramatic effect upon the value of the refractive index because of the extreme steepness of the dispersion profile. Essentially even a very small shift of the resonant wavelength causes a large change in the refractive index for a field applied close to this frequency. It is predicted that strong cross-phase modulations will be created in this process between the fields ω_f and ω_d, even in the quantum limit for these fields. This is predicted to lead to improved schemes for quantum nondemolition measurements of photons through the measurement of the phaseshifts they induce (through cross-phase modulation) on another quantum field. This type of measurement has direct application in quantum information processing.

Further Reading

Arimondo, E., 1996. Coherent population trapping in laser spectroscopy. Progress in Optics 35, 257–354.
Harris, S.E., 1997. Electromagnetically induced transparency. Physics Today 50, 36–42.
Harris, S.E., Field, J.E., Imamoglu, A., 1990. Non-linear optical processes using electromagnetically induced transparency. Physical Review Letters 64, 1107–1110.
Harris, S.E., Hau, L.V., 1999. Non-linear optics at low light levels. Physical Review Letters 82, 4611–4614.
Hau, L.V., Harris, S.E., Dutton, Z., Behroozi, C.H., 1999. Light speed reduction to 17 metres per second in an ultra-cold atomic gas. Nature 397, 594–598.
Marangos, J.P., 2001. Electromagnetically induced transparency. In: Bass, M. (Ed.), Handbook of Optics, vol. IV, ch. 23. New York: McGraw-Hill.
Merriam, A.J., Sharpe, S.J., Xia, H., *et al.*, 1999. Efficient gas-phase generation of vacuum ultra-violet radiation. Optics Letters 24, 625–627.
Schmidt, H., Imamoglu, A., 1996. Giant Kerr nonlinearities obtained by electromagnetically induced transparency. Optics Letters 21, 1936–1938.
Scully, M.O., 1991. Enhancement of the index of refraction via quantum coherence. Physical Review Letters 67, 1855–1858.
Scully, M.O., Suhail Zubairy, M., 1997. Quantum Optics. Cambridge, UK: Cambridge University Press.
Sokolov, A.V., Walker, D.R., Yavuz, D.D., *et al.*, 2001. Femtosecond light source for phase-controlled multi-photon ionisation. Physical Review Letters 87, 033402-1.
Zhang, G.Z., Hakuta, K., Stoicheff, B.P., 1993. Nonlinear optical generation using electromagnetically induced transparency in hydrogen. Physical Review Letters 71, 3099–3102.

Nonlinear Optics, Basics: Nomenclature and Units

MP Hasselbeck, The University of New Mexico, Albuquerque, NM, USA

Nomenclature

CARS Coherent anti-Stokes Raman scattering
c.c. Complex conjugate
CSRS Coherent Stokes Raman scattering
DFG Difference frequency generation
DFM Difference frequency mixing
DFWM, FWM (Degenerate) Four wave mixing
DRO Doubly resonant OPO
EIT Electromagnetically induced transparency
ESA Excited state absorption
GVD Group velocity dispersion
GVM Group velocity mismatch
NLA Nonlinear absorption
NLR Nonlinear refraction
OFID Optical free-induction decay
OPA Optical parametric amplification/amplifier
OPG Optical parametric generation/generator
OPL Optical power limiter
OPO Optical parametric oscillation/oscillator
2PA, TPA Two-photon absorption
PCM Phase conjugate mirror
QPM Quasi phase matching
RIKES Raman-induced Kerr effect spectroscopy
RSA Reverse saturable absorption
SBS Stimulated Brillouin scattering
SFG Sum frequency generation
SHG Second harmonic generation
SIT Self-induced transparency
SRO Singly resonant OPO
SRS Stimulated Raman scattering
SRWS Stimulated Rayleigh wing scattering
SVEA, SVAA Slowly varying envelope (amplitude) approximation
THG Third harmonic generation
TPF Two-photon fluorescence

Nomenclature Associated with the Excitation Light

Particular care must be used when characterizing the excitation beam in nonlinear optical experiments compared to linear measurements. By definition, nonlinear optical phenomena depend on the electric field to high order. The higher the order, the more sensitive the observed behavior depends on the input. Extracting a representative nonlinear coefficient from an experiment, for example, becomes progressively more difficult as the order of the optical nonlinearity increases. In other words: the errors associated with optical beam characterization get magnified by the order of the nonlinearity under study.

There is an extensive nomenclature for characterizing the light (almost always laser light) interacting with the nonlinear optical medium. Different descriptions may be used depending on whether the excitation light is pulsed, continuous, or a continuous train of pulses. When a laser beam is constant or continuous, it is often described by its 'cw power'. 'Cw' stands for 'continuous wave', an acronym taken from the nomenclature of electronics. The cw power of a laser is determined by placing a power meter in the path of a beam. Repetitively pulsed lasers can be characterized in the same way, provided the response time of the power meter is slower than the pulse separation period. This will almost certainly be the case with a continuously pumped mode-locked laser such as a dye laser, fiber laser, or Ti:sapphire laser, which produce pulses at repetition frequencies of tens of megahertz. Cw power may also be suitable for describing pulsed flashlamp lasers, gas discharge lasers, or any laser that is excited in a periodic fashion.

When the optical output can be distinguished as individual pulses, additional metrics are used. A pulse, by definition, exists during a window of time, i.e., there is a time when light is present and a time when light is absent. How one characterizes the time light is present – the pulse duration – is critically important and not always obvious. There are a myriad of ways the pulse temporal envelope can manifest itself; common examples include square pulses, triangular pulses, Gaussian pulses, and hyperbolic secant pulses. These names refer to the mathematical waveforms that map the pulse envelope as a function of time.

A pulse waveform that exhibits symmetry about the peak in its temporal envelope is commonly characterized by its 'full-width, half-maximum', abbreviated FWHM. One obtains the full-width, half-maximum by locating the two points on the pulse profile that are at half the peak value. The temporal separation of these two points is the FWHM. One uses this measure because in a strict mathematical sense, waveforms such as the Gaussian or hyperbolic-secant exist even at times $t = \pm \infty$. A less common term is the 'half-width, half-maximum' or HWHM. As the name implies, the HWHM is exactly half the FWHM value. Not all light-pulses are temporally symmetric, however, so greater detail may be needed when giving a mathematical description of asymmetric pulses.

Light pulses are also characterized by their energy and peak power. The temporal envelope recorded by a so-called square-law detector (all laboratory detectors are square-law detectors) shows the power of the pulse as a function of time, provided the detector response time is sufficiently fast. Integrating this waveform gives the pulse energy. It makes no sense to talk about the energy of a cw beam of light, unless that beam is composed of repetitive, distinguishable pulses. Each pulse in the train carries a distinct amount of optical energy.

'Ultrafast' or 'ultrashort' laser pulses generally refer to light pulses that are of such small duration they cannot be measured directly with detectors and oscilloscopes because of bandwidth limitations. To infer the pulse duration, so-called intensity autocorrelation measurements are often made. The temporal power profile is then deduced from the autocorrelation signal with the appropriate mathematical conversion factor. Equally important is the spectrum of a short pulse. The optical bandwidth determines the lower limit of temporal compression; such ideally compressed pulses are said to be 'transform limited'. If different components of the spectrum can be distinguished at specific positions on the pulse temporal profile, then the pulse is 'chirped'.

Nonlinear optical effects take place when light is concentrated on a target – there is a cross-section or 'footprint' of the beam on the medium. The interaction of the light and material takes place in a region called the beam area, beam cross-section, focal area, or focal volume. One must then define the appropriate interaction area or volume.

Specification of the beam area allows one to define two important quantities used frequently in nonlinear optics. The power/area is known as irradiance or power density. This quantity is commonly called the intensity, although the strict radiometric definition of intensity is power/(solid angle). The second important quantity is energy/area, which is called the energy density or fluence.

The convention in nonlinear optics is to define the optical electric field as:

$$\mathbf{E}(\mathbf{z},t)=\frac{1}{2}E_0\exp[\mathrm{i}(\mathbf{k}\cdot\mathbf{z}-\omega t)]+\text{c.c.} \tag{1}$$

where E_0 is the peak field, $\mathbf{k}$ is the propagation vector, ω is the angular frequency of the light, and c.c. stands for complex conjugate. This represents a forward- and backward-propagating (in the z-direction) infinite, transverse plane wave. Other definitions are also used. In SI/mks units and using the convention of **Eq. (1)**, the irradiance (I in units of W/m^2) inside a material of refractive index n is related to the electric field vector of the light ($\mathbf{E}$ in units of V/m) as follows:

$$I=\frac{1}{2}cn\varepsilon_0|\mathbf{E}|^2 \tag{2}$$

Here c is the speed of light (2.998×10^8 m/s) and ε_0 is the permittivity of free space (8.854×10^{-12}F/m). Nonlinear optics has an unfortunate tradition of mixing mks and cgs units, resulting in a lot of confusion. Irradiance is usually expressed in units of W/cm^2. One can calculate the peak electric field (in units of V/cm) of a laser beam given its irradiance (in units of W/cm^2) by using this simple formula:

$$E_0=38.82\sqrt{\frac{I}{n}}\left(\frac{\mathrm{V}}{\mathrm{cm}}\right) \tag{3}$$

The root-mean-square value of the electric field is obtained by replacing the factor 38.82 by 27.45. In Gaussian/cgs units, irradiance is expressed in units of ergs/(cm^2 sec) and is related to the field as:

$$I=\frac{cn}{8\pi}|\mathbf{E}|^2 \tag{4}$$

where the field is in units of statvolt/cm and $c=2.998\times10^{10}$ cm/sec. Conversion between mks and cgs for the electric field is 3×10^4 V/m$=1$ statvolt/cm; irradiance is converted using 1 erg /(cm^2 sec)$=1\times10^{-3}$ W/m^2. The following point must be emphasized: depending on how the electric field is defined, there can be factors of 2 or even 4 discrepancies in the irradiance values quoted by different authors. The common definitions are used here, although they are certainly not universal.

A very common realization of the spatial irradiance profile of a laser beam is the radially symmetric Gaussian function:

$$I(r)=I_0\exp\left(-\frac{2r^2}{w^2}\right) \tag{5}$$

where I_0 is the peak irradiance at the center of the beam ($r=0$) and w is the 'spot size'. The spot size changes continuously as the beam propagates and the minimum spot size is known as the waist. One often hears the spot size referred to as the radius, but the radius of a Gaussian beam traditionally denotes the curvature of the phase front. By definition, the phase front radius is infinite at the waist. The factor of 2 appearing in the argument of the exponential in **Eq. (5)** stems from the fact that the spot size w is conventionally defined for the electric field of the Gaussian beam. When the field is squared to obtain the irradiance, the factor of 2 appears. Using this function for the irradiance spatial profile leads to the common parameter '1/e^2 diameter'. The irradiance falls to 1/e^2 (0.1353) of its peak value when $r=w$.

The Gaussian spatial profile or 'Gaussian beam' results when the laser has been designed to operate in the lowest-order transverse mode, usually denoted TEM$_{00}$. The TEM$_{00}$ Gaussian beam is particularly convenient because the same relative power profile is maintained in the near- and far-field, whether or not the beam is collimated or focused.

The cross-sectional area of a TEM$_{00}$ Gaussian beam normally incident on a surface is found as follows. The power in the beam is obtained by integrating the irradiance profile (I) over the surface:

$$P=\int_{\pm\infty}\int I\mathrm{d}A \tag{6}$$

Referring to **Eq. (5)**, the integral becomes:

$$P=\int_0^{2\pi}\mathrm{d}\theta\int_0^{\infty}r\mathrm{d}rI_0\exp\left(-\frac{2r^2}{w^2}\right)=I_0\frac{\pi w^2}{2} \tag{7}$$

Hence the effective area of a TEM_{00} Gaussian beam normally incident on a surface is:

$$A = \frac{\pi w^2}{2} \tag{8}$$

There are many situations where the Gaussian formulation of Eq. (5) is not suitable. Beams produced by unstable laser resonators, for example, are not Gaussian. A multitransverse mode beam profile obtained from a stable resonator does not usually lend itself to a simple mathematical characterization.

Nomenclature Associated with the Nonlinear Optical Medium

In nonlinear optics texts, the subject is often introduced by writing the macroscopic polarization in Maxwell's equations as a power series expansion in the electric field. This approach, proposed by Bloembergen and coworkers in the early 1960s, has been spectacularly successful for interpreting experiments, though it has also led to confusion in the definition of nonlinear optical coefficients. The confusion, which stems from arbitrary definitions and nomenclature used by different authors, cannot be resolved here. The reader should, however, be alert for these discrepancies. Comparison of results published by different laboratories requires that the fundamental equations for extracting the nonlinear coefficients from their data are known. As the discipline has matured, the nonlinear optics community largely recognized the ambiguities and confusion; precise definitions of terms and coefficients are now commonly provided in the research literature.

It should also be noted that the power-series framework may not always be the best formulation for modeling and characterizing nonlinear optical effects. A carefully designed optical limiter, for example, may exhibit an abrupt decrease of transmission at a specific 'threshold irradiance' or 'threshold fluence', perhaps at a given laser pulse duration. The physical phenomenon or phenomena driving this behavior may not readily succumb to characterization by an nth-order coefficient in a power series expansion. In this situation it may be appropriate to specify a threshold optical input parameter. Another example is a homogeneously broadened saturable absorber, in which the absorption of an optical medium decreases with increasing input irradiance. The irradiance-dependent absorption coefficient α (refer to Eq. (26)) is characterized by a 'saturation irradiance' I_{sat}:

$$\alpha(I) = \frac{\alpha_0}{1 + \frac{I}{I_{sat}}} \tag{9}$$

where α_0 is the linear absorption coefficient.

Nonlinear Susceptibility

The power-series expansion of the macroscopic polarization is the standard approach for modeling nonlinear optical behavior and categorizing the various phenomena. A common way to write the polarization, nonlinear in the electric field, is (mks units):

$$\mathbf{P} = \varepsilon_0(\chi^{(1)} \cdot \mathbf{E} + \chi^{(2)} : \mathbf{EE} + \chi^{(3)} \vdots \mathbf{EEE} + \ldots) \tag{10}$$

where the polarization $\mathbf{P}$ is a time- and space-dependent vector and the terms $\chi^{(n)}$ are the various orders of the nonlinear susceptibility (the cgs equation is obtained by dropping the free-space permitivitty coefficient ε_0). A second-order effect is associated with $\chi^{(2)}$, a third-order with $\chi^{(3)}$, and so on. In general, there are distinguishable electric field vectors (in Eq. (10), the fields have not been individually designated for clarity). The susceptibility terms are generally tensors, which means the medium is sensitive to the orientation of the input fields. The dots in Eq. (10) indicate tensor products. The input field vectors can have different frequencies and the presence of complex conjugates in Eq. (1) indicates that the frequency components will have both + and − signs, i.e. the jth electric field term will have associated frequency factors $\exp(\pm i\omega_j t)$.

The nomenclature used here can be illustrated by example. In the case of the second-order nonlinear polarization (mks units):

$$\mathbf{P}_j = \varepsilon_0 \chi^{(2)}_{jkl} : \mathbf{E}_k \mathbf{E}_l \tag{11}$$

Note that subscripts have been introduced. The indices correspond to Cartesian space vectors. This equation says that the polarization in the j-direction results from the tensor product of the appropriate $\chi^{(2)}$ with the input fields at $\mathbf{E}_k$ and $\mathbf{E}_l$. The jth component of polarization has an oscillation frequency that is determined by mixing of the input fields. In this example, the indices j, k and l can each take the value of x, y, and z. Let each input field be at one of two possible frequencies, call them ω_1 and ω_2. There are an enormous number of permutations (81 to be exact) of Eq. (11). One term is:

$$P_x(\omega_1 + \omega_2) = \varepsilon_0 \chi^{(2)}_{xzy}(\omega_1 + \omega_2) E_z(\omega_1) E_y(\omega_2) \exp[-i(\omega_1 + \omega_2)t] \tag{12}$$

Another possibility is:

$$P_z(\omega_2 - \omega_1) = \varepsilon_0 \chi^{(2)}_{zyx}(\omega_2 - \omega_1) E_y(\omega_2) E_x(-\omega_1) \exp[-i(\omega_2 - \omega_1)t] \tag{13}$$

Eq. (12) corresponds to sum frequency generation ($\omega_1 + \omega_2$) and Eq. (13) shows difference frequency generation ($\omega_2 - \omega_1$).

Not explicitly written on the right-hand side of the above equations is the exponential containing the mixing of the propagation vectors. To maximize the nonlinear polarization, which is the source of nonlinear behavior in Maxwell's equations, the

vector sum $\sum \mathbf{k}_j$ for *all* the interacting fields should be close to zero. Arranging the propagation vectors to accomplish this is known as 'phase matching'.

It is important to point out that there are eight more terms describing the nonlinear polarization at $P_x(\omega_1 + \omega_2)$ in addition to Eq. (12); likewise for $P_z(\omega_2 - \omega_1)$. Also note that the indices k and l may be identical, hence a nonlinear polarization driven by $\chi^{(2)}_{xyy}$ for example, is allowed. Higher-order nonlinear polarizations get progressively more complicated. The general third-order nonlinear polarization is

$$\mathbf{P}_j = \varepsilon_0 \chi^{(3)}_{jklm} : \mathbf{E}_k \mathbf{E}_l \mathbf{E}_m \tag{14}$$

Sometimes the second-order nonlinear susceptibility is written with the coefficient d instead of $\chi^{(2)}$. The relation between the d-coefficient and $\chi^{(2)}$ depends arbitrarily on how it is defined; a common definition is

$$d_{jkl} = \frac{1}{2}\chi^{(2)}_{jkl} \tag{15}$$

but the reader is cautioned that the factor $\frac{1}{2}$ is sometimes missing. In the preceding discussion it was pointed out how there can, in principle, be a large number of tensor components involved in the expansion of the nonlinear polarization. Simplification occurs when it is realized that there is no discernible physical difference between the frequency terms ω_j and $-\omega_j$ and that ordering of the fields in Eq. (14) – which suggests a time order in the arrival of the fields – is irrelevant. These symmetry arguments show that $d_{jkl} = d_{jlk}$, which reduces the number of independent d-coefficients from 81 to 18. A simpler subscript notation is then introduced that uses the integers 1–6 to represent pairs of Cartesian components k and l. This reduces the number of subscripts from three to two. An example of this notation is $d_{xzz} = d_{14}$.

In certain situations, one can take advantage of crystal symmetry to further reduce the complicated summations to a single scalar coefficient for the nonlinear susceptibility, which is referred to as 'd-effective'. For these specialized cases, the nonlinear polarization is

$$P = \varepsilon_0 d_{\text{eff}} E_1 E_2 \tag{16}$$

In SI/mks units, the polarization is in units of coul/m^2. The second-order susceptibility (i.e. $\chi^{(2)}$, d_{jkl}) must therefore have units of m/volt. The units of $\chi^{(3)}$ are m^2/volt^2, $\chi^{(4)}$ is m^3/volt^3, and so on. In Gaussian/cgs units, $\chi^{(2)}$ has dimensions cm/statvolt, $\chi^{(3)}$ will be cm^2/statvolt^2, etc. although sometimes in the Gaussian system *all* the coefficients $\chi^{(n)}$ are discussed with shorthand 'electrostatic units' or 'esu'.

Complex Quantities

In linear optics, the susceptiblity has real and imaginary parts

$$\chi^{(1)} = \chi^{(1)}_{\text{real}} + \mathrm{i}\chi^{(1)}_{\text{imaginary}} \tag{17}$$

The complex linear index is, in turn, written as the sum of real and imaginary components, derived from the linear susceptibility as follows (mks units):

$$\sqrt{1 + \chi^{(1)}} = n_0 + \mathrm{i}\kappa \tag{18}$$

where n_0 is the linear refractive index and κ is the imaginary term leading to absorption of light. In general, the nonlinear susceptibilty $\chi^{(n)}$ is also a complex number:

$$\chi^{(n)} = \chi^{(n)}_{\text{real}} + \mathrm{i}\chi^{(n)}_{\text{imaginary}} \tag{19}$$

As in linear optics, the complex notation is a bookkeeping method that conveniently accounts for what is known as 'resonant enhancement'. Nonlinear optics research has revealed that the nonlinear susceptibility $\chi^{(n)}$ can be a strong function of frequency. Specifically, this quantity will be strongly enhanced when sums and/or differences of the photon energies in the interacting light beams coincide with quantum mechanical energy resonances in the material. When a resonance condition is achieved, $\chi^{(n)}$ will be dominated by its imaginary component. Far from the resonances, $\chi^{(n)}$ behaves more like a real quantity. These complex terms directly enter the wave equations describing how light propagates in a nonlinear medium and are particularly important for $\chi^{(3)}$; the complex quantity $\chi^{(3)}$ determines whether light will be refracted or absorbed and by how much. The real part of $\chi^{(3)}$ drives nonlinear refraction while the imaginary portion characterizes nonlinear absorption (e.g., two-photon absorption) and the inverse effect – gain.

Nonlinear Refraction

The most common manifestation of nonlinear refraction arises from the third-order nonlinear susceptibility, the so-called optical Kerr effect. In this case, the refractive index is linearly proportional to the irradiance (I) of a monochromatic light beam. The irradiance-dependent refractive index is

$$n(I) = n_0 + n_2 I \tag{20}$$

where n_0 is the linear, irradiance-independent refractive index and n_2 is the nonlinear refractive index coefficient. Because n is a dimensionless quantity, n_2 must be in units of area/power.

The optical Kerr effect is usually written in mks units with the expression shown in **Eq. (20)**. In cgs units, the common (but by no means universal) convention is to write:

$$n(E) = n_0 + \frac{1}{2}\tilde{n}_2|\mathbf{E}|^2 \tag{21}$$

where the field is in units of statvolts/cm. The nonlinear coefficient $\tilde{n}_2$ (cgs) must therefore be in units of cm^2/statvolt^2, which is sometimes abbreviated to 'esu'. The conversion for n_2 between these two equations is:

$$\tilde{n}_2(\text{cgs}) = \frac{n_0 c}{40\pi} n_2(\text{mks}) \tag{22}$$

where n_2(mks) is in units of m^2/W and c is in m/sec. Also useful is the relation between n_2 and $\chi^{(3)}$, written in mks units as:

$$\Delta n = n_2 I = \frac{\text{Re}\left(\chi^{(3)}\right)}{4\varepsilon_0 n^2 c} I \tag{23}$$

where 'Re' denotes the real part of $\chi^{(3)}$. In cgs units we have:

$$\Delta n = \frac{1}{2}\tilde{n}_2|\mathbf{E}|^2 = \frac{4\pi^2 \text{Re}(\chi^{(3)})}{n^2 c} I \tag{24}$$

where I is in units of ergs/(cm^2 sec) defined by **Eq. (4)**. The reader is again advised that different authors will show discrepancies of 2, 4, or even 8 when writing these equations. The susceptibilities $\chi^{(3)}$ are related as follows:

$$\tilde{\chi}^{(3)}(\text{cgs}) = \frac{910^8}{4\pi}\chi^{(3)}(\text{mks}) \tag{25}$$

If the coefficients n_0 and n_2 exist, does the nomenclature imply there are terms n_1, n_3, and others? These coefficients are certainly allowed, but not often seen in discussions of nonlinear refraction. When one refers to nonlinear refraction, the common understanding is that the index depends linearly on irradiance, which is conveniently modeled by **Eq. (20)**. But there are other physically relevant situations to consider. The coefficient n_1, for example, describes the linear electro-optic effect in which the change in index is linearly proportional to the electric field (although it is called the linear electro-optic effect, it is actually derived from the second-order nonlinear susceptibility). The electric field can be the oscillating field of a laser beam, for example, or a dc field applied to an electro-optic crystal (e.g., Pockel's cell). When other terms are added to **Eqs. (20)** or **(21)**, the units of the coefficients must be chosen to keep the equation dimensionless.

In the preceding discussion, there is an implication that the nonlinear index is an instantaneous function of the irradiance or field. Although the material system can never respond instantaneously, this is a good approximation in many situations. Sometimes it is not. Consider a pulsed laser beam that heats the material. When the local temperature increases, the linear refractive index can change. A short, Q-switched laser can have a pulse duration far less than a microsecond, while the temperature change it induces can last many orders of magnitude longer. This means optical modification of the refractive index may persist long after the exciting pulse has vanished, i.e., when the irradiance is at zero.

Another example is when a laser beam promotes electrons from their ground state to higher energy states of the material system. These excited electrons may modify the refractive index of the material. In general, the excited electrons remain in high-energy states for some period of time before relaxing to the ground level – if this time is longer than the excitation, the change of the refractive index persists beyond the time the laser beam is present. In these situations, the model of nonlinear refraction suggested by the above equations is inaccurate. One cannot obtain the nonlinear refraction from a simple algebraic analysis. A system of time-dependent equations – describing the dynamics of excitation and relaxation – must be solved using numerical procedures. Such phenomena are sometimes loosely categorized as 'dynamic linear optical effects' or 'effective third-order nonlinearities'.

Nonlinear Absorption

Consider a single-frequency light beam passing through an optically absorbing region of length L. For simplicity, neglect reflections caused by surfaces that may define the region of interest, i.e., ignore reflections at surfaces that may be located on the optical axis z at points $z=0$, $z=L$, or any other point in the path. Linear absorption means that the optical power extracted from the light beam as it traverses the absorbing medium is a direct function of the power at a given point. This is described mathematically by an elementary, linear differential equation known as Beer's law:

$$\frac{\text{d}}{\text{d}z} I(z) = -\alpha I(z) \tag{26}$$

The constant of proportionality is α, which is the linear absorption coefficient. **Eq. (23)** is solved by direct integration:

$$\int_{I(0)}^{I(L)} \frac{\text{d}I}{I} = -\alpha \int_0^L \text{d}z \tag{27}$$

Table 1 Dimensions of optical absorption coefficients

Process	*Coefficient*	*Units*
Linear absorption	α, K_1	$(\text{length})^{-1}$
Two-photon absorption	β, K_2	length/power
Three-photon absorption	γ, K_3	$\text{length}^3/\text{power}^2$
Four-photon absorption	K_4	$\text{length}^5/\text{power}^3$

This has the solution:

$$\frac{I(L)}{I(0)} = \exp(-\alpha L) \tag{28}$$

which means the irradiance decreases exponentially as a function of propagation distance in a linearly absorbing medium.

In the nonlinear regime, we don't expect a classical Beer's law model to hold. By definition, the absorption will be a nonlinear function of irradiance at a given point. One makes the following power-series expansion to describe nonlinear absorption of monochromatic light:

$$\frac{\mathrm{d}}{\mathrm{d}z}I = -\alpha I - \beta I^2 - \gamma I^3 - \cdots \tag{29}$$

The coefficient β corresponds to a two-photon absorption process and γ is the coefficient of three-photon absorption. What about four-photon and even higher-order processes? These are rarely encountered, but certainly possible. To make these distinctions, **Eq. (29)** is sometimes written with coefficients K_i or α_i instead of the sequential Greek alphabet:

$$\frac{\mathrm{d}}{\mathrm{d}z}I = -K_1 I - K_2 I^2 - K_3 I^3 - K_4 I^4 - \cdots \tag{30}$$

The units of the various absorption coefficients must maintain the dimensional consistency of **Eqs. (29)** and **(30)**, which is irradiance/length or power/$(\text{length})^3$. These are shown in **Table 1**.

If there is linear absorption, the nonlinear processes at that wavelength are often (but not always) negligibly weak. This means we have the following inequalities:

$$K_1 I \gg K_i I^i, \quad \text{where } i \geq 2 \tag{31}$$

Nonlinear absorption is generally observed in the wavelength region where the medium is transparent to low-irradiance light, i.e., where linear absorption is negligible. It should be emphasized that there are situations where linear absorption is large and in fact a crucial component of the aggregate nonlinear effect. We will return to this point later in the discussion. For the moment, we neglect linear absorption. If the energy of the incident photons and energy levels of the system permit it, the lowest-order nonlinear process is two-photon absorption, described by the equation

$$\frac{\mathrm{d}}{\mathrm{d}z}I = -K_2 I^2 \tag{32}$$

which can be solved by elementary integration:

$$\frac{I(L)}{I(0)} = \frac{1}{1 + I(0)K_2 L} \tag{33}$$

Unfortunately, the situation is rarely this convenient. The above integration has ignored the fact that nonlinear absorption may enable significant linear absorption. This is possible because nonlinear absorption promotes electrons from low- to high-energy states in the medium. The change in excited electron density associated with absorption of light (both linearly and nonlinearly) is called 'photocarrier generation'. These photocarriers (e.g., photo-electrons) may modify the linear absorption as well as the refractive properties of the material. If two-photon absorption causes the linear absorption to increase, for example, the assumption that allowed us to ignore K_1 when writing **Eq. (32)** ceases to be valid. While **Eq. (32)** was appropriate at the very start of the light–matter interaction, the generation of photocarriers after the passage of time may change that condition. The behavior of the system is therefore time-dependent, i.e., it is dynamic. Simple analytic solutions are almost always not possible when dealing with nonlinear absorption.

The time rate of change of the photocarrier population (N) is modeled by the following equation:

$$\frac{\partial}{\partial t}N = \frac{K_1 I}{\hbar\omega} + \frac{K_2 I^2}{2\hbar\omega} + \frac{K_3 I^3}{3\hbar\omega} + \frac{K_4 I^4}{4\hbar\omega} + \cdots - \text{recombination} - \text{diffusion} \tag{34}$$

where ω is the angular frequency ($\omega = 2\pi f$) of the incident light. Also included are place-holders for loss processes such as recombination and diffusion, which may themselves have complicated mathematical descriptions. The coefficients K_i have exactly the same units, dimensions, and interpretation as in **Eq. (30)** and **Table 1**; in the spirit of the preceding discussion, these coefficients may be time dependent. Note that the dimensions of **Eq. (30)** describe absorption of light (irradiance/length), while **Eq. (34)** models the photocarrier population density (length^{-3} time^{-1}). The units of these two equations are very different.

Further Reading

Bloembergen, N., 1992. Nonlinear Optics. Redwood City, CA: Addison-Wesley.
Boyd, R.W., 1991. Nonlinear Optics. Boston, MA: Academic Press.
Levenson, M.D., Kano, S., 1988. Introduction to Nonlinear Laser Spectroscopy. Boston, MA: Academic Press.
Miller, A., Miller, D.A.B., Smith, S.D., 1981. Dynamic nonlinear optical processes in semiconductors. Advances in Physics 30, 697–800.
Mills, D.L., 1998. Nonlinear Optics: Basic Concepts. New York: Springer.
Newell, C., Moloney, J.V., 1992. Nonlinear Optics. Redwood City, CA: Addison-Wesley.
Sauter, E.G., 1996. Nonlinear Optics. New York: Wiley.
Shen, Y.R., 1984. The Principles of Nonlinear Optics. New York: Wiley.
Sutherland, R.L., 1996. Handbook of Nonlinear Optics. New York: Marcel Dekker.
Yariv, A., 1991. Optical Electronics, 3rd edn. New York: Holt, Rinehart, and Winston.
Yariv, A., 1989. Quantum Electronics. New York: Wiley.
Zernike, F., Midwinter, J.E., 1973. Applied Nonlinear Optics. New York: Wiley.

Raman Spectroscopy

R Withnall, University of Greenwich, Chatham, UK

Introduction

Inelastic light scattering, the optical analogue of Compton scattering, had been predicted to occur by Smekal in 1923, but it was Chandresekar Venkataraman Raman who provided the first experimental demonstration of the phenomenon in February 1928. Only a few months later in 1928 the Russian scientists, Landsberg and Mandelstam observed inelastic light scattering from a quartz crystal. In spite of initial claims to the contrary, this was the same effect that had been observed by Raman, although the Russian scientists did not acknowledge this, preferring to call the effect 'combinatorial scattering'. It was not until the 1970s that the effect became known as Raman scattering everywhere, including Russia.

As typically practised, Raman spectroscopy involves laser excitation of a sample and measurement of the wavelength and intensity distribution of the scattered Raman light. When a microscope is used for delivering laser excitation and/or collecting the inelastically scattered light, the technique is referred to as Raman microscopy. However, it is important to recognize that, apart from the differences in sampling configuration, there are no fundamental differences between Raman microscopy and Raman spectroscopy; the terms merely identify the different sampling techniques.

Since its invention in 1960, major advances in the technology of the laser have occurred, resulting in the wide variety of laser light sources that are currently available for Raman spectroscopy. Characteristic properties of laser systems for Raman microscopy will be described in this chapter, following an introduction to the technique and its applications in the next section below. Then, the commonly used laser sources for Raman microscopy are categorized according to the wavelength regions of their emissions, along with their merits and drawbacks for each application. The chapter will consider mainly continuous wave laser sources, because pulsed laser beams that are tightly focused with microscope optics give very high irradiance that would destroy most samples.

Raman Microscopy

The first experimental Raman microscopes were developed in the mid-1970s by two independent groups, and it was not long before the first generation of Raman microscopes were subsequently commercialized. Many of these early instruments simply consisted of commercial optical microscopes coupled to Raman spectrometers, as shown in Fig. 1. It was recognized that the use of a microscope for sampling in this way can facilitate the Raman examination of tiny particles of micrometer dimensions.

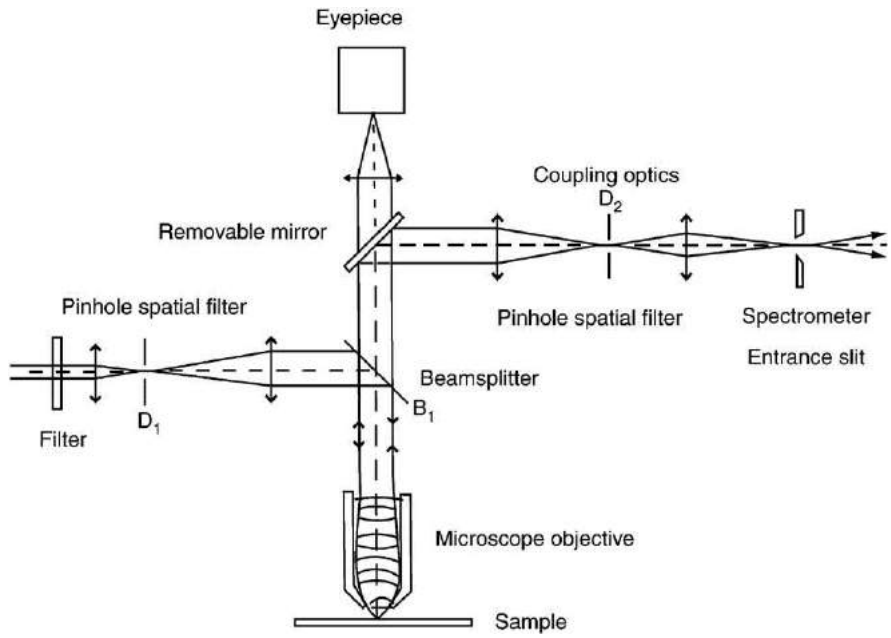


Fig. 1 Diagram of the optical layout of a visible Raman microscope. Reproduced from Turrell G, Delhaye M and Dhamelincourt P (1996) Characteristics of Raman microscopy. In: Turrell G and Corset J (eds) *Raman Microscopy. Development and Applications*, 2. London: Academic Press.

 doi:10.1016/B978-0-12-809283-5.00960-5

The optical lay-out in Fig. 1 shows an infinity corrected microscope. The advantage of using this, rather than a standard tube length microscope, is that its length can be extended in order to incorporate additional optical elements that are required for coupling to the Raman excitation and collection optics. In the optical layout, the incident laser beam is spatially filtered by a pinhole, D_1, in order to remove the diffraction rings and give a point source. The laser light is then partially reflected by beamsplitter, B_1, and focused on to the sample by the microscope objective. The back-scattered Raman radiation is collected by the same microscope objective, partially transmitted by the beamsplitter, and directed into the entrance slit of the Raman spectrometer by means of coupling optics. The accurate location of an aperture, D_2, at the focal point improves the spatial resolution and allows depth profiling of transparent samples. The two pinholes, D_1 and D_2, act as spatial filters and are referred to as confocal diaphragms. The Raman microscope is described as being confocal, because out-of-focus light, collected from outside the focal volume, is rejected.

The use of the beamsplitter, in the optical path of early Raman microscope designs, gives rise to significant losses of the precious Raman scattered light. If a 50:50 beamsplitter is used then half of the incoming laser light is lost by transmission at B_1 and, more significantly, half of the Raman scattered light is lost by reflection at B_1. As long as a higher laser power can be employed, it is better to use a 90:10 beamsplitter that transmits 90% of the light and reflects 10%; then, only 10% of the incoming laser light is directed by B_1 on to the sample, but 90% of the Raman scattered light is transmitted by B_1 towards the detector and only 10% of the Raman light is lost by reflection at B_1. The efficiency of a beamsplitter can be defined as the product of the reflectivity at the laser wavelength and the transmission at the Raman-shifted wavelength. Thus, conventional 50:50 and 90:10 beamsplitters would have efficiencies of 25% and 9%, respectively. Recently, however, holographic notch filters have been developed which can be used as higher efficiency beamsplitters. These typically have a laser rejection contrast ratio of better than 10^4, reflecting ca. 90% of the laser excitation and transmitting ca. 90% of the Raman scattered light, thus giving an efficiency of ca. 81%. Due to these considerations, higher throughput, commercial Raman microscope designs usually incorporate a holographic notch filter.

The importance of Raman microscopy stems from the fact that it is the only microanalytic method available today, by use of which it is possible to identify, or characterize, small particles of micrometer dimensions *in situ*. It is also advantageous that no sample preparation is required when performing Raman microscopy.

The fundamental limit of the lateral spatial resolution (i.e., for a diffraction limited focus) is the separation at which the maximum of one Airy disc function just touches the first minimum of an adjacent Airy disc, given by:

$$\text{Lateral spatial resolution} = \frac{1.22\lambda}{2\text{NA}} \tag{1}$$

where λ is the wavelength of the light and NA is the numerical aperture of the microscope objective.

The axial Airy disc function determines the resolution in the longitudinal direction. A good estimate of the axial resolution limit for low numerical aperture is given by:

$$\text{Axial spatial resolution} = \frac{2\lambda}{(\text{NA})^2} \tag{2}$$

When the sample is heterogeneous and exhibits fluorescence that is not evenly distributed within the sample, a region of the sample can be selected with the microscope that shows the minimum amount of fluorescence. If the fluorescence is intrinsic to the sample itself, then it is possible to use the shift subtract technique and/or temporal discrimination between the fluorescence and the Raman scattering. The latter can be achieved by Kerr gate fluorescence rejection, in which a pulse of light is used to close a Kerr gate shutter. The problem with this approach, however, is that it uses a laser pulse, and even pulses of modest energy will have extremely high irradiance when focused to a tiny spot by a microscope objective. Such laser pulses would inevitably destroy the sample under a microscope.

Another approach for reducing fluorescence is to use near infrared excitation, which is low enough in energy so that absorption cannot occur to promote electronic transitions. The technique of Fourier Transform Raman (FTR) spectroscopy often offers the best chance of obtaining Raman spectra from fluorescent samples for this reason. Low energy excitation of wavelength equal to 1064 nm, provided by a Nd:YAG laser, is normally used for FTR. Disadvantages are that the Raman scattering efficiency is low relative to that obtained with visible excitation, due to the ν^4 dependency (ν^4 refers to Rayleigh scattering of the Raman light). This is compensated to some extent by employing a high throughput (Jacquinot advantage) interferometer. However, this works better for macro rather than micro samples. This is because the coupling of a microscope to an FT Raman spectrometer has a fundamental drawback; the microscope is throughput limited, so the high throughput advantage of the FTR spectrometer cannot be realized in FTR microscopy. The image of the Jacquinot stop (the large circular aperture which is the entrance to the interferometer) at the sample plane typically has a diameter of a few hundred μm, which is much larger than particles with diameters of ca. 1 μm having similar dimensions to the waist diameter of the 1064 nm light excitation at its diffraction-limited focus. Consequently the throughput advantage of the interferometer is not fully exploited and there is a trade-off of spatial resolution with signal-to-noise at the detector. For this reason, the typical spatial resolution that is achieved is in the range of 15–100 μm rather than 4 μm, as claimed in the literature. Bruker has commercialized an FTR microscope by coupling an optical microscope to an FTR spectrometer with near infrared optical fibers.

An approach that offers more promise for reducing fluorescence and achieving spatial resolution close to the diffraction limit is to use near infrared excitation in conjunction with dispersive Raman microscopy. For example, semiconductor lasers operating

from 785 to 852 nm can be used in conjunction with sensitive multichannel silicon-based CCD arrays. If longer excitation wavelengths are necessary, in order to overcome the problems with fluorescence, then diode lasers can be used that emit at longer wavelengths in conjunction with dispersive Raman spectrometers equipped with multichannel near infrared (NIR) detectors, e.g., Ge- or InGaAs-array detectors.

The applications of Raman microscopy cover many areas, including material science, the earth sciences, environmental science, biology and medicine, forensic science, and even the analysis of artworks. These areas are too wide ranging to be described in detail here, but the interested reader is referred to the Further Reading section at the end of this article.

Characteristics of Laser Sources

The laser is an excellent light source for Raman spectroscopy to such an extent that the terms 'Raman spectroscopy' and 'laser Raman spectroscopy' are synonymous for all but the most esoteric experiments, for example, with synchotron sources. Indeed, the advent of the laser was the stimulation for the renaissance of Raman spectroscopy in the 1960s, given its special properties, such as monochromaticity, high intensity, beam collimation, and coherence.

The characteristics of laser light and the advantages it offers to Raman microscopy are now considered.

Beam Quality

The light inside a laser tube is formed from a number of standing waves having distinct vibrational modes. There are a limited number of these modes transverse to the beam and these are characterized by the TEM_{pq} number (where p and q can be 0, 1, 2,…) where TEM is an acronym for 'transverse electromagnetic'. When a laser is operating in its fundamental TEM_{00} mode, light rays are reflected on axes between the end mirrors of the laser cavity. The '00' indicates that there are no nodes in the beam profile (Fig. 2(a)), and the laser beam has a Gaussian intensity profile in the radial direction:

$$I(r) = I_0 \exp\left(-\frac{2r^2}{w^2}\right) \tag{3}$$

where $I(r)$ is the irradiance at a radial distance r from the axis of the beam, I_0 is the axial irradiance, and w is the beam radius.

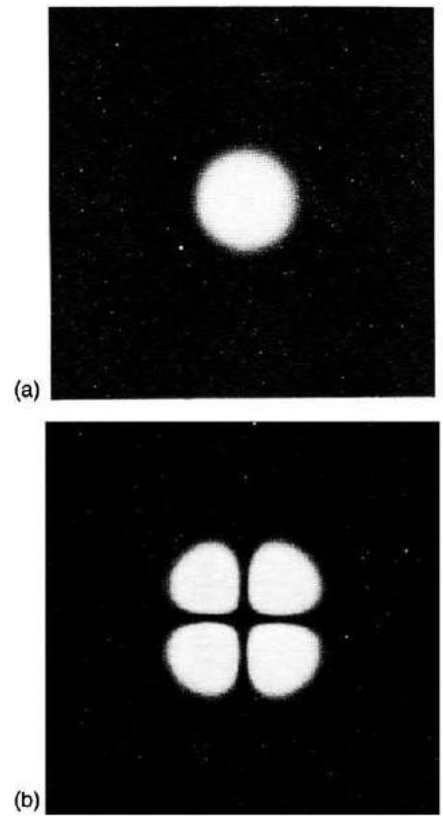


Fig. 2 Transverse electromagnetic modes formed with confocal concave mirrors, (a) TEM_{00}, and (b) TEM_{11}. Reproduced with permission from Young M (1977) *Optics and Lasers: an Engineering Approach.* Berlin, Heidelberg: Springer-Verlag.

For higher-order modes, a number of nodes are observed in the beam profile, which arise from off-axis light rays being reflected between the end mirrors, for example, the TEM_{11} mode has two nodes which are mutually perpendicular (Fig. 2(b)).

The propagation characteristics of a Gaussian beam can be fully defined, either by the diameter of the beam waist or by the far-field divergence. Consequently, it is only necessary to know the diameter of the beam waist ($2w_0$), or the diameter of the beam ($2w_z$) at a longitudinal distance z from the waist, in order to determine the propagation characteristics of the beam:

$$w(z) = w_0\sqrt{1 + \left(\frac{z}{z_R}\right)^2} \tag{4}$$

$$R(z) = z\left[1 + \left(\frac{z_R}{z}\right)^2\right] \tag{5}$$

where the quantity $z_R = \pi w_0^2/\lambda$ is known as the Rayleigh range of the beam, λ is the wavelength of the laser radiation, and $R(z)$ is the radius of curvature of the wavefront at a distance z from the beam waist.

The wavefront is planar at the minimum beam waist and the Rayleigh range is the distance from the beam waist to the location at which the wavefront is most curved (Fig. 3), the region from the waist to the Rayleigh range being the near field. Beyond approximately ten times the Rayleigh range, in the far field, the beam diverges as a cone with approximately straight sides. It can be seen by substituting $z = z_R$ into Eq. (4) that the beam diameter at the Rayleigh range is $\sqrt{2}$ times the waist diameter.

Unfortunately laser beams do not conform to pure Gaussian functions and therefore they do not propagate according to the above equations. As a result, a dimensionless beam propagation parameter was devised in the early 1970s. This parameter, known as the M^2 factor or the 'times diffraction limit', is based on the brightness theorem, which states that for any laser beam the product of the beam diameter and the far-field divergence is a constant. Thus, M^2 is defined as the ratio of the laser beam's multimode diameter-divergence product to the ideal diffraction-limited (TEM_{00}) beam diameter-divergence product:

$$M^2 = \left(\frac{2w_m\Theta_m}{2w_0\theta_0}\right) \tag{6}$$

where Θ_m is the laser beam's multimode divergence, θ_0 is the theoretical diffraction-limited divergence, $2w_m$ is the laser beam's multimode waist diameter, and $2w_0$ is the ideal diffraction-limited beam waist diameter.

Alternatively, M^2 can be defined as the square of the ratio of the multimode beam diameter ($2w_m$) to the diffraction-limited beam diameter ($2w_0$):

$$M^2 = \left(\frac{2w_m}{2w_0}\right)^2 \tag{7}$$

The intensity of the beam has a Gaussian distribution in the radial direction, but the accepted definition of beam diameter is the distance at which the intensity has fallen to $1/e^2$ (i.e., ca. 13.5%) of its axial intensity, as can be seen by setting $r = w$ in Eq. (3) above. This definition only applies properly to Gaussian beams; for other beam profiles the diameter can be calculated using a second moment measurement of the entire beam.

The real beam, then, can be treated as Gaussian by substituting an artificial wavelength $M^2\lambda$ into the equations that apply to an ideal diffraction-limited TEM_{00} beam. The ideal TEM_{00} Gaussian can be superimposed as an 'embedded Gaussian' on the optical

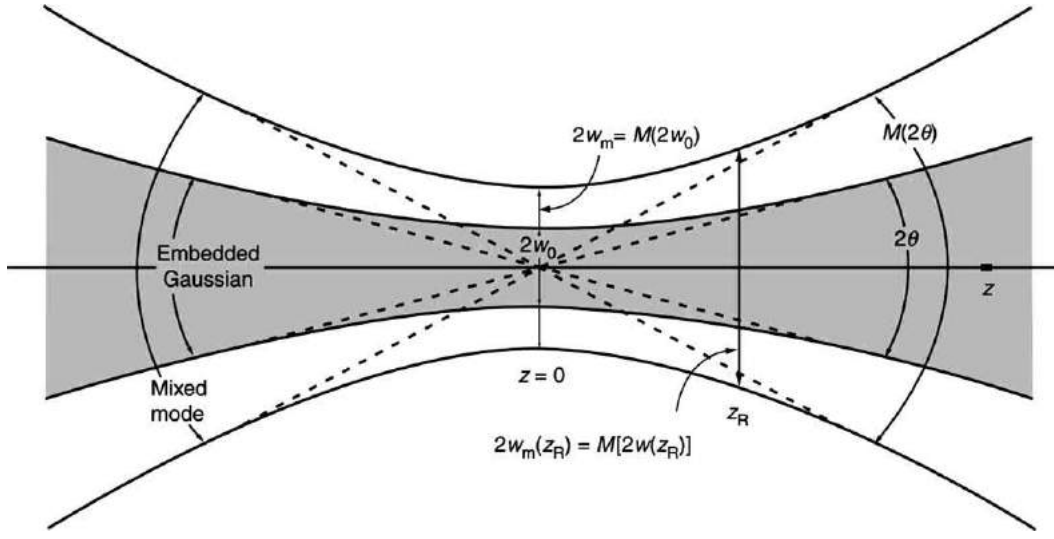


Fig. 3 The Rayleigh range and embedded Gaussian. Adapted from http://beammeasurement.mellesgriot.com/tut_m2.asp.

diagram of the real multimode beam (Fig. 3). It can be seen that the multimode beam has a waist diameter and a divergence that are both M times larger than those of the embedded Gaussian.

The M^2 factor has a value of unity for an ideal diffraction-limited TEM_{00} beam, and values of greater than unity for all other beams; it can be as high as several hundred for a distorted beam of poor quality. This is an important parameter where a tightly focused laser spot (e.g., confocal Raman microscopy) or low divergence (e.g., remote sensing) are required.

Polarization

Although polarization is not an inherent property of laser light, lasers provide plane polarized light when the end windows of the laser tube are mounted at Brewster's angle (θ_B):

$$\theta_B = \tan^{-1} n \tag{8}$$

where n is the refractive index of the window material for the appropriate wavelength of laser light. For this angle, reflected light is completely plane polarized and consequently the laser light that is generated is also plane polarized. Brewster windows are often used inside the laser cavity in order to reduce reflection losses. This is important in low-gain lasers such as HeNe lasers where laser action could be prevented by reflection losses. Air-cooled ionic gas lasers typically deliver highly linearly polarised output with ratios exceeding 1000:1.

The polarization property of the exciting light is used a great deal in Raman spectroscopy. In combination with the polarization properties of the Raman scattered light, it can be used to determine the depolarization ratios, ρ, of Raman bands in the solution state. In solid state studies it can be used to determine the degree of orientation, for example, in polymeric fibers or films, and in single crystal studies it can be used to determine the symmetry classes of the Raman active vibrations. In such studies a linear polarizer is used to improve the degree of plane polarization of the laser light, and a half-wave plate is used to rotate the plane of polarization by 90°.

Plane polarized laser light can be converted to circularly polarized laser light which is required for Raman optical activity (ROA) studies. For example, plane polarized 514.5 nm laser light provided by an argon ion laser can be converted to circularly polarized light by an electro-optic modulator. As CCD detectors are usually employed in present-day ROA experiments and they have relatively long sampling times, slow polarization modulation is required which can be achieved by periodically rotating a quarter-wave plate or by applying a square quarter-wave voltage to a Pockels cell.

Whereas optical isomers or enantiomers give identical Raman spectra, they are distinguished in the ROA experiment by the sign of the ROA signal. Indeed, the sign can be used to determine the absolute configuration of a chiral molecule. Also the enantiomeric excess of mixtures of enantiomers can be determined in such experiments, with obvious applications in the pharmaceutical industry, and the conformations of biological molecules, for example, proteins, nucleic acids, sugars, and viruses, can be determined.

Up until now, the ROA technique has been practised by no more than a handful of research groups world-wide, but it is expected that this situation will change in the near future, since the first commercial ROA spectrometer (the ChiralRAMAN spectrometer) has recently been launched by Biotools Inc. This spectrometer makes use of a solid state laser as the source of the polarized laser light.

Beam Divergence

Referring to Fig. 3, the full divergence angle, Θ, for the fundamental TEM_{00} Gaussian beam is given by:

$$\Theta = \lim_{z \to \infty} \frac{2w(z)}{z} = \frac{2\lambda}{\pi w_0} \tag{9}$$

From Eq. (4) it can be seen that the spot size increases linearly with the longitudinal distance z and, from Eq. (9), that it diverges at a constant cone angle, Θ. It can also be seen from Eq. (9) that the smaller the spot size, w_0, the greater the beam divergence angle, Θ.

Indeed, a highly divergent beam is a problem with edge-emitting laser diodes due to the small dimensions of the light source. Furthermore, it is much smaller in the vertical direction than in the horizontal, giving rise to an elliptical output beam that diverges much more rapidly in the vertical direction than in the horizontal. The highly divergent, elliptical beam can be corrected, to an extent, with a cylindrical lens, but the inherent problem of a small, elliptical source can never be completely corrected.

In contrast, vertical-cavity surface-emitting semiconductor lasers (VCSELs) do not have the same limitations on beam divergence because the cavity is in the vertical direction and the light emission occurs from the surface, which has a much larger area than the light source of an edge emitter. Indeed, frequency-doubled green and blue semiconductor lasers are now available, having beams with M^2 values of less than 1.1 and beam divergences of a few milliradians. In terms of these beam characteristics, it appears that these lasers are beginning to challenge the argon ion laser.

Emission Linewidth

The laser linewidth limit, $\Delta\nu_L$, is given by the Schawlow–Townes expression:

$$\Delta\nu_L = \frac{2\pi h\nu(\Delta\nu_c)^2}{P_{out}} \tag{10}$$

where $\Delta\nu_c$ is the linewidth of the passive resonator, $h\nu$ is the photon energy, and P_{out} is the laser power output.

The power output, P_{out}, is equal to the number of photons, N_p, in the resonator times the energy per photon, hv, divided by the photon lifetime, τ_c:

$$P_{out} = \frac{N_p h\nu}{\tau_c} \tag{11}$$

Furthermore, the width $\Delta\nu$ of the resonance curve, at which the intensity has fallen off to half the maximum value, is given by:

$$\Delta\nu_c = (2\pi\tau_c)^{-1} \tag{12}$$

Substituting for P_{out} and $\Delta\nu_c$ from **Eqs. (8)** and **(9)** into **Eq. (10)** gives the limit for the linewidth, $\Delta\nu_L$:

$$\Delta\nu_L = \frac{\Delta\nu_c}{N_p} \tag{13}$$

The linewidth of a laser is characterized by its Q factor:

$$Q = \frac{\nu}{\Delta\nu} \tag{14}$$

where ν is the frequency of the laser line and $\Delta\nu$ is the linewidth.

Other useful relationships involving linewidth parameters are:

$$Q = \frac{\lambda}{\Delta\lambda} \tag{15}$$

and

$$\Delta\tilde{\nu} = -\frac{\Delta\lambda}{\lambda^2} \tag{16}$$

where λ and $\Delta\lambda$ are the wavelength and linewidth (in units of metres), respectively, and $\Delta\tilde{\nu}$ is the linewidth expressed as a wavenumber (in units of cm^{-1}).

The Q factor can be as high as 10^8 which is of great use in high resolution spectroscopy.

Visible Lasers

Helium-Neon Laser

The helium-neon laser is continuously pumped electrically using a high dc voltage of up to 1 kV. The gain medium consists of a gas mixture of about 10 parts helium to each part of neon at a pressure of about 3 Torr. The gain of this medium is extremely low, hence the requirement for Brewster-angle windows to eliminate reflection loss of the light polarized in the plane that includes the axis of the laser and the normal to the Brewster window. Due to this requirement, the light output is necessarily plane polarized.

The pumping excites helium atoms by electron impact, and resonant energy transfer to neon atoms then occurs via collisions of the gaseous atoms (**Fig. 4**). This creates a population inversion in the neon atoms, enabling the laser transition to occur at 632.8 nm. Following emission, the neon atoms decay nonradiatively down to their metastable $2p^5 3s^1$ level from which they decay back down to the ground state via collisional de-excitation with the walls of the tube. This mechanism restricts the maximum achievable power output to ca. 50 mW, because of the need to depopulate the metastable neon atoms by wall collisions.

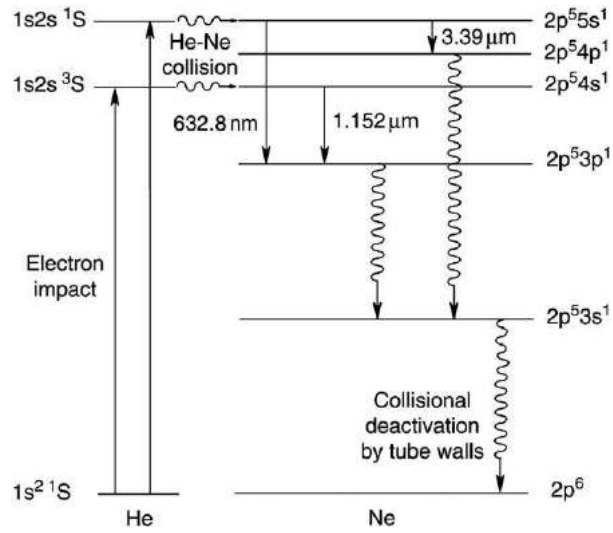


Fig. 4 Energy levels of the He-Ne laser.

An advantage of using helium-neon lasers for Raman spectroscopy is that they generally operate in the fundamental TEM_{00} mode, which is critically dependent on the ratio of the length of the tube to its diameter. For this reason they are designed to have small tube diameters of around a few millimetres and lengths of 0.15–0.50 m, the small diameter also aiding collisional deactivation with the walls due to the relatively large surface to volume ratio of the tube. Another advantage of this laser for Raman spectroscopy is that the linewidth of the 632.8 nm emission line is ca. 1.5 GHz.

A disadvantage of the helium-neon laser is that it emits a large number of spontaneous emission lines originating from the excited neon atoms. These plasma emissions are observed in the Raman spectrum as sharp lines unless they are filtered out, for example, by a pre-monochromator or an interference filter. It should be mentioned, in passing, that these plasma lines are not always unwanted, because they can be very useful for wavenumber calibration of the Raman spectrometer.

The low output power of <50 mW is not generally a disadvantage for Raman microscopy since higher laser powers can often cause photolytic or thermal degradation of the sample. This is due to the high irradiance at the sample when the laser light is focused to a tight spot by the microscope objective. The excitation wavelength of 632.8 nm is suitable for combined use of the laser with a silicon-based CCD detector in Raman spectroscopy, because large Stokes shifts in the 3000–4000 cm^{-1} region lie well within the quantum efficiency curve of this detector. It is also possible, though less common, to operate the helium-neon laser on weaker transitions that include the green 543.5 nm line.

Argon Ion Laser

The argon ion laser was one of the first lasers to be discovered following the invention of the laser and up until now it has been used extensively for Raman spectroscopy, among many other applications. Excitation is provided by a continuous electrical discharge, and because of the high energy required to ionize the argon atoms and then promote the ions to an excited state, the efficiency of the laser is very small (ca. 0.1%). In spite of this, once population inversion has been achieved, the gain of the laser is very high and output powers of up to 25 W can be obtained for the strong lines at 488.0 and 514.5 nm and up to 50 W for multiline operation.

An advantage of the argon ion laser is that it can provide emission at more than 35 discrete wavelengths, the strongest of which are listed in **Table 1**. These lie in the green, blue, and near ultraviolet regions of the spectrum, a number of the ultraviolet lines only being obtained on the larger frame argon ion lasers. Discrete laser emission lines are selected by tuning a prism or grating inside the cavity. The argon ion laser can also be operated in multiline mode, for example, for pumping dye or Ti:sapphire lasers.

The linewidths of the argon ion emission lines at 488.0 and 514.5 nm are around 4.0 GHz. The high temperature laser tube has a diameter of approximately 12 mm and a length in the range of 0.1 to 1.8 m. A 240 V three phase power supply and ca. 35 A are required for pumping a medium power 5 W argon ion laser. The tube also requires water cooling at a flow rate of ca. 10 L/min and a pressure of 25 psi because of the large amount of heat dissipation.

Table 1 Continuous wave argon ion laser wavelengths and output powers for the Coherent Inc. Innova 300 argon ion laser system

Wavelength (nm)	*Power (mW)*
228.9	30
238.3	100
244.0	400
248.3	180
257.3	750
275.4	5
300–305.5	20
333.4	40
333.8	30
335.8	20
351.1	200
351.4	60
363.8	240
454.5	140
457.9	420
465.8	180
472.7	240
476.5	720
488.0	1800
496.5	720
501.7	480
514.5	2400
528.7	420

The five lines below 260 nm are frequency doubled.

For Raman microscopy smaller, air-cooled argon ion lasers, which only require a 240 V single phase power supply, can be used to provide output of a few hundred milliwatts on the 488.0 and 514.5 nm lines. For this application, the excitation is provided by argon ion lasers, which are designed to operate in the TEM_{00} mode.

Until recently, the argon ion laser has been a workhorse as the most common excitation source for Raman spectroscopy, but it is now losing ground to solid state lasers. The principle reasons for this are that the latter are much more efficient and consequently do not in general have three phase power and water-cooling requirements. Nevertheless, argon ion lasers still have a niche when high excitation powers on the order of watts are required in conjunction with a good beam quality ($M^2 < 1.1$), for example, for Raman spectroscopy of gaseous samples. Another advantage of argon ion lasers over solid state lasers is their multiline capability.

Krypton Ion Laser

The krypton ion laser has useful discrete emission lines in the near ultraviolet, blue, yellow, red, and near infrared regions of the electromagnetic spectrum; the strongest lines of a Coherent Inc. Innova 400 krypton ion laser are listed in **Table 2**.

The argon ion and krypton ion lasers are close relatives, thus the large frame krypton ion laser has similar power and water-cooling requirements to those mentioned above for the argon ion laser. Also, like the argon ion laser, air-cooled models with lower power output are available. The characteristics of the two lasers are also similar; for example, the linewidths of the krypton ion transitions at 530.9, 568.2, and 647.1 nm are about 4.0 GHz, the laser tube has similar dimensions (0.1–2.0 m in length) and the laser can be operated in either TEM_{00} or multimode.

The krypton ion laser is even less efficient than the argon ion laser, consequently lower-power outputs of up to about 20 W can be achieved when operated in multiline mode.

Mixed argon and krypton ion laser tubes are also commonly used that provide laser lines originating from both argon ion and krypton ion transitions. As with the individual argon ion and krypton ion lasers, sharp plasma lines, due to spontaneous emission from the rare gas ions, are observed in the Raman spectrum unless they are filtered out. If interference filters are used for this, each laser line requires its own filter that is tailored to the specific emission wavelength.

HeCd Laser

Laser lines at 325 and 442 nm, both capable of providing milliwatts of output power, can be obtained from the helium-cadmium laser. These emission lines result from electronic transitions in free cadmium atoms.

It is obviously advantageous to use blue, rather than longer wavelength, excitation (e.g., provided by the 442 nm line of the HeCd laser) for off-resonance Raman spectroscopy, due to the v^4 dependence of the Raman light scattering efficiency, provided that the efficiencies of the illumination/collection optics and Raman spectrometer throughput are optimized in the blue region of the spectrum. Unfortunately, far more samples fluoresce when excited with blue light than with red or near infrared radiation, which explains why red or near infrared lasers (e.g., HeNe, semiconductor lasers) are more commonly used for general Raman applications.

Table 2 Continuous wave krypton ion laser wavelengths and output powers for the Coherent Inc. Innova 400 krypton ion laser system

Wavelength (nm)	*Power (mW)*
206.5	4
234	8
337.5–356.4	2000
406.7	900
413.1	1800
415.4	280
468.0	500
476.2	400
482.5	400
520.8	700
530.9	1500
568.2	1100
647.1	3500
676.4	900
720.8	45
752.5	1200
793.1–799.3	300

The 206.5 and 234 nm lines are frequency doubled.

Near Infrared Lasers

Although the traditional laser systems for Raman spectroscopy have been the argon ion, krypton ion, and helium-neon lasers for discrete excitation wavelengths, the diode laser has gained popularity over recent years.

The principal advantage of near infrared semiconductor lasers for Raman microscopy is that they generally excite less fluorescence in the Raman spectrum than visible lasers due to their longer wavelength. Commonly available wavelengths are 670, 785, 830, and 852 nm and, of these, the first two can be used in conjunction with silicon-based CCD detectors to give Stokes Raman shifts over the whole range of fundamental vibrations, up to 4000 cm^{-1}. However, for excitation wavelengths of 830 and 852 nm, only Stokes Raman bands in the fingerprint region can be detected with a silicon-based CCD detector because higher wavenumber Raman shifts approach the bandgap of the silicon semiconductor (λ>ca. 1050 nm).

The disadvantages of diode lasers are that they cannot supply high power of narrow linewidth in single-mode and they are susceptible to mode hopping which gives rise to a shift in the excitation wavelength. This latter drawback is particularly disadvantageous for Raman spectroscopy because it results in a shift in the wavenumber positions of the Raman bands. When good beam quality is not required, broad stripe, high power laser diodes can be used; these can have output powers that are greater than 1 W but their emission linewidths are equal to ca. 2 nm, which is far too wide for Raman spectroscopic applications. These laser diodes find applications as pumps of solid state lasers, for example, Nd:YAG, Nd:YVO_4 lasers, however. Additionally, amplified stimulated emission (ASE) gives an unwanted background signal that can be very broad (ca. 20–30 nm) and 0.1–1% of the intensity of the laser line. This necessitates the use of bandpass filters in order to reduce this unwanted background.

Although laser diodes having an output of less than 200 mW can operate in TEM_{00} and in single longitudinal modes, mode hopping can occur due to optical feedback or fluctuations in environmental factors and this causes severe broadening of the linewidth. The sensitivity to optical feedback can be greatly reduced, hence the laser linewidth can be narrowed appreciably, by confining the frequency of the photons either in an internal cavity containing a small diffraction grating or in an external cavity. The former types of laser are sometimes called 'distributed feedback' (DFB) lasers because the feedback is distributed over the length of the grating, rather than occurring all at once at a mirror. The wavelength that is fed back is determined by the period of the grating. Usually, a DFB laser has a grating fabricated into the entire length of the laser. A variation referred to as a distributed Bragg reflector (DBR) laser has a distinct grating fabricated into the substrate on each side of the active area. The external cavity semiconductor laser (ECSL) is fabricated by placing the laser diode in a separate resonator like a conventional gas or solid-state laser. The DBR and ECSL lasers are now discussed.

Distributed Bragg Reflector Lasers

The DBR laser consists of a grating on each side of the active region (Fig. 5); these gratings act as mirrors having a reflectivity that is optimized at one particular wavelength, in addition to narrowing the laser linewidth. At present, DBR lasers are only commercially available having an excitation wavelength of 852 nm.

The emission linewidth is less than 4 MHz, which is suitable for the vast majority of Raman spectroscopic applications and the laser operates in single TEM_{00} mode. The disadvantages of DBR lasers are that the maximum output power is restricted to around 150 mW, so an optical isolator is usually necessary in order to prevent external facet damage caused by external optical feedback, and only an excitation wavelength of 852 nm is currently available.

External-Cavity Semiconductor Lasers (ECSL)

ECSLs provide higher output power (up to ca. 1 W) and a wider range of excitation wavelengths (630 nm to around 850 nm), as well as being less expensive than DBR lasers. They use the semiconductor chip only as the gain medium and employ an external grating, both as the frequency selector and the reflective mirror. Specific excitation wavelengths are becoming widely adopted in Raman microscopy employing single grating spectrograph in conjunction with CCD detection, such as 670, 785, 830, and 852 nm. The ECSLs have emission linewidths as low as a few MHz, but in general they span the range 2 MHz–30 GHz, and can operate in a nearly diffraction-limited transverse mode.

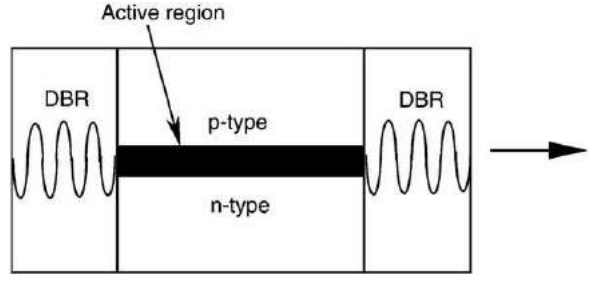


Fig. 5 Diagram of the cavity of the DBR laser. Adapted from Pan M-W, Benner RE and Smith LM (2002) Continuous lasers for Raman spectrometry. In: Chalmers JM and Griffiths PR (eds) *Handbook of Vibrational Spectroscopy*, 1. Chichester, UK: Wiley.

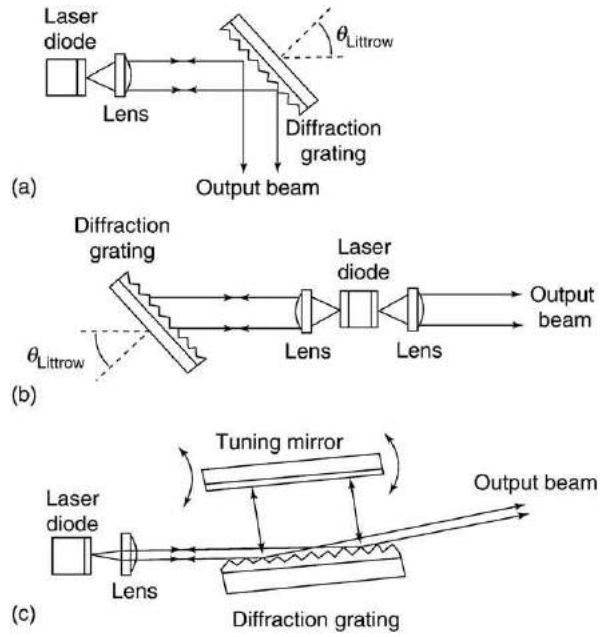


Fig. 6 Diagrams of the cavities of (a) Littrow type-I, (b) Littrow type-II, and (c) Littman external cavity diode lasers. Adapted from Pan M-W, Benner RE and Smith LM (2002) Continuous lasers for Raman spectrometry. In: Chalmers JM and Griffiths PR (eds) *Handbook of Vibrational Spectroscopy*, 1. Chichester, UK: Wiley.

Three different cavity designs employing a diffraction grating have been discussed in the literature: Littrow type-I, Littrow type-II, and Littman configurations. In the Littrow type-I design, the laser diode light is incident on the grating at an angle of incidence equal to $\theta_{Littrow}$ (**Fig. 6(a)**). The diffracted light in first order is re-directed back to the laser diode to provide feedback and the output is the zeroth order (i.e., specularly reflected) light. A disadvantage of this design is that the bandwidth of the grating is relatively large because a single pass geometry is used. In the Littrow type-II design, light from the rear facet of the laser diode is collimated by a lens and directed on to a grating at an angle of incidence equal to $\theta_{Littrow}$ (**Fig. 6(b)**). The diffracted light in first order is re-directed back to the laser diode, in a similar fashion to the Littrow type-I design, in order to provide optical feedback, and the output is the laser light emitted from the front facet. Disadvantages of the Littrow type-II design are the requirement for custom fabrication of antireflection coatings on both facets and the sensitivity to external optical feedback. In the Littman design, the laser diode light is collimated by a lens and directed on to a grating at a grazing angle of incidence. The diffracted light is reflected by a mirror and diffracted back to the laser diode by the grating in order to provide optical feedback (**Fig. 6(c)**). An advantage of this design is that the grating bandwidth is less than half that of the Littrow designs due to the double pass geometry, but a disadvantage is that the external cavity length is longer, resulting in a narrower spacing of the cavity modes. Another advantage of this design is that the tuning is accomplished by rotating the mirror instead of the grating, thus the alignment of the output beam is not altered when tuning. For Littrow type I and II and Littman ECSL designs, filtering of the ASE is required, as it can have an intensity of ca. 0.1 to 1% of the intensity of the laser line. Thus, these designs necessitate the incorporation of a bandpass filter (having a rejection of better than 10^{-4} at the ASE).

Nd:YAG Laser

Under normal conditions, the Nd:YAG laser oscillates on a transition in the near infrared at 1064 nm, and this is the excitation line that is used in most commercial Fourier Transform Raman spectrometers. The gain medium is a crystal of yttrium aluminium garnet ($Y_3Al_5O_{12}$, YAG) doped with ca. 1.0 mol% Nd^{3+} cations that substitute Y^{3+} ions in the cubic YAG lattice.

The Nd:YAG laser is a four-level system (**Fig. 7**), which has high gain and low threshold due to the narrow fluorescent linewidth of the laser transition. Absorption bands of the Nd^{3+} ions around 808 nm conveniently match the energy of commercially available high-power multimode diode lasers that serve as the pump. The Nd^{3+} ions decay nonradiatively to the upper laser level, the $^4F_{3/2}$ state, thereby creating a population inversion. This is because the lower laser level, the $^4I_{11/2}$ state, has no appreciable thermal population at room temperature, since it is >2000 cm^{-1} above the $^4I_{9/2}$ ground state, and the $^4F_{3/2}$ excited state has a relatively long lifetime of 230 μs.

The high thermal conductivity of Nd:YAG enables the laser to operate at high power in either continuous wave (CW) or *Q*-switched modes, and the diode pumped solid state (DPSS) variety of the cw Nd:YAG laser can currently achieve an output power in excess of 20 W on the 1064 nm line. The laser can be designed to operate in the fundamental TEM_{00} mode and the full width at half maximum (FWHM) linewidth of the spontaneous emission of the 1064 nm laser transition is 120 GHz (ca. 0.45 nm). Advantages of the diode pumped Nd:YAG laser are that it is air-cooled and can be operated at 240 V single phase.

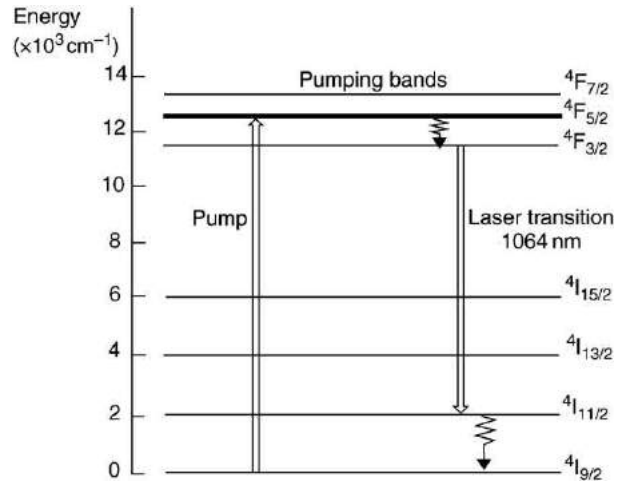


Fig. 7 Energy levels of the Nd:YAG laser.

Also, being all solid state and having a small footprint, it is robust and portable. The lifetime of the laser is dependent on the laser diodes used for pumping, but some Nd:YAG laser designs enable these to be replaced in the field.

The frequency-doubled Nd:YAG laser emitting green light, having a wavelength of 532 nm, is also commonly used nowadays for dispersive Raman spectroscopy, and an output power of 10 W on the 532 nm laser line is provided by commercially available frequency-doubled diode-pumped Nd:YAG lasers.

It is worth mentioning that Nd:YAG lasers have also been fabricated that emit either at 946 or 1330 nm, by suppressing the strong emission at 1064 nm and optimising the optics for the desired wavelength.

Nd:YVO_4 and Nd:YLF Lasers

An intracavity, frequency-doubled DPSS Nd:YVO_4 laser has recently become commercially available, and it has some advantages over the Nd:YAG laser including a larger stimulated emission cross-section and a higher absorption coefficient (along the extraordinary direction of the birefringent crystal). It has the same emission wavelength as the frequency-doubled Nd:YAG laser, i.e., 532 nm, and Nd:YVO_4 is the material of choice for cw end-pumped lasers having around 5 W output power. This is because the diode laser pump beam is tightly focused in the end-pumped system, but a small waist diameter cannot be retained over a distance of more than a few millimeters, consequently a high absorption coefficient and gain are very beneficial.

The gain medium of the Nd:YLF laser consists of a crystal of yttrium lithium fluoride (YLF) that is doped with Nd^{3+} ions on the Y^{3+} cation sites. Unlike the Nd:YAG and Nd:YVO_4 lasers, the emission does not occur at 1064 nm; instead the $^4F_{3/2}$–$^4I_{11/2}$ emission occurs at wavelengths of either 1047 or 1053 nm, depending on the polarization that is selected. The former is due to extraordinary polarized light, whereas the latter is due to ordinary polarization, and either of these emission wavelengths can be selected using an intracavity polarizer. The Nd:YLF laser can offer benefits in *Q*-switched operation when the longer fluorescence lifetime (480 μs) of Nd^{3+} ions in the $^4F_{3/2}$ state enables a higher energy to be stored for the same number of pump laser diodes.

Ti:Sapphire Laser

The Ti:sapphire laser is tunable over the approximate range of 670–1070 nm with the peak of the gain curve at ca. 800 nm. In the gain medium, ca. 0.1% by weight Ti^{3+} is doped into a crystal of sapphire grown by the Czochralski method. The Ti^{3+} ions substitute for Al^{3+} ions in the Al_2O_3 of the sapphire and the laser emission is due to the ^2E–$^2T_2$ transition of the Ti^{3+} cation, which has a $3d^1$ valence electronic configuration (Fig. 8). The laser has a large stimulated emission cross-section but the fluorescence lifetime of the upper laser level (^2E state) is quite short (3.2 μs), thus the laser is usually laser (e.g., by argon ion or frequency-doubled Nd:YAG or Nd:YVO_4 lasers) rather than flashlamp pumped because a very high pump flux is required.

The absorption and emission bands are broad and widely separated, due to the vibronic coupling between the Ti^{3+} host and the Al_2O_3 lattice (Fig. 9). The lower laser level is any one of the vibronic levels of the $^2T_2$ state. Following the laser transition, the Ti^{3+} ions decay from the upper vibronic levels of the $^2T_2$ state down to the lower vibronic levels. Hence, this is a four-level laser.

The broad spontaneous emission linewidth of the ^2E–$^2T_2$ laser transition is around 100 THz, the ouput power can be as high as 50 W, and the laser can be operated in either TEM_{00} or multimodes.

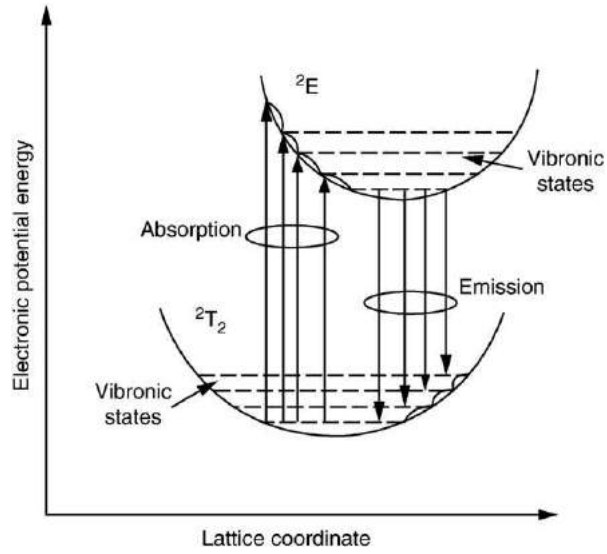


Fig. 8 Energy levels of the Ti:sapphire laser. Adapted from Pan M-W, Benner RE and Smith LM (2002) Continuous lasers for Raman Spectrometry. In: Chalmers JM and Griffiths PR (eds) *Handbook of Vibrational Spectroscopy*, 1. Chichester, UK: Wiley.

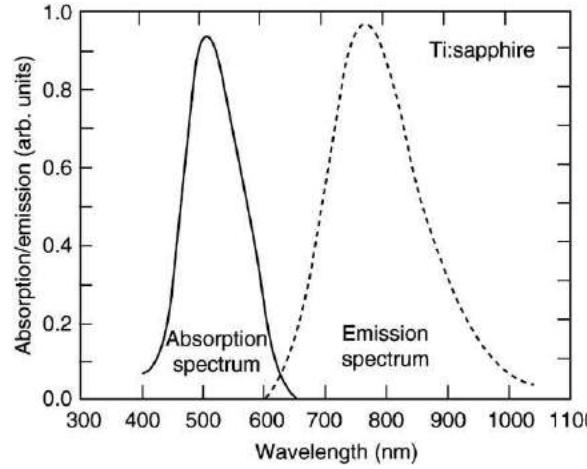


Fig. 9 Absorption and emission spectra of the Ti^{3+} ion in sapphire. Al_2O_3. Adapted from Pan M-W, Benner RE and Smith LM (2002) Continuous lasers for Raman Spectrometry. In: Chalmers JM and Griffiths PR (eds) *Handbook of Vibrational Spectroscopy*, 1. Chichester, UK: Wiley.

UV Lasers

UV laser sources offer numerous advantages over visible laser sources for Raman spectroscopy. A major consideration is that many analytes have absorptions in the near UV, making them amenable to resonance Raman spectroscopic studies. The signal enhancement (up to 10^6), that can be achieved in resonance Raman spectra, results in a large increase in detection sensitivity. Furthermore, some vibrational modes, which normally give rise to weak bands in the off resonance Raman spectrum, can show strong enhancement in the UV excited resonance Raman spectrum. A good example of this is the amide II band of peptides and proteins, which is due to a combination of N–H in plane bending and C–N stretching modes of the peptide linkage. This band is normally weak in the Raman spectrum but strong in the infrared spectrum; however, the amide II band can show strong enhancement in the UV excited resonance Raman spectrum and this can be useful for determining secondary structure in proteins. A further benefit of UV excited Raman spectroscopy is that fluorescence can often be avoided as it tends to occur at a lower energy, outside the Stokes Raman spectral window.

It can be advantageous to use cw rather than pulsed excitation for UV resonance Raman spectroscopy, if the analyte has a long excited state lifetime relative to the pulsewidth of the pulsed laser. This is because the concentration of analyte molecules in the ground state is significantly depleted during pulsed excitation, due to Raman saturation giving a lower signal to noise ratio in the Raman spectrum than for the case of cw excitation. It has been found that pulse energy flux densities should be less than

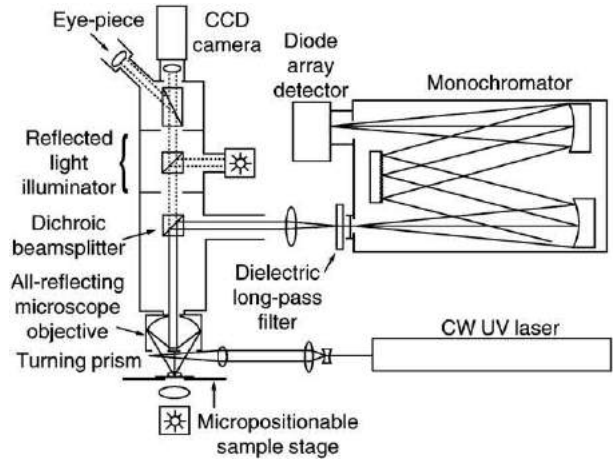


Fig. 10 Diagram of the optical layout of a UV Raman microscope. Adapted from Pajcini V, Munro CH, Bormett RW, Witkowski RE and Asher SA (1997) UV Raman microspectroscopy: Spectral and spatial selectivity and simplicity. *Applied Spectroscopy* 51: 81–86.

1 mJ/cm^2 and flowing or spinning samples should be used, in order to ensure that nonlinear phenomena and saturation effects do not occur.

For UV Raman microscopy, where the laser beam is focused into a small spot of micrometer dimensions on the sample, it is essential to avoid using pulsed lasers delivering high peak powers. This is because such pulsed lasers can cause dielectric breakdown, and even at lower peak powers they can cause nonlinear effects and Raman saturation phenomena. Consequently, with UV Raman microscopy, one is restricted to the use of cw or quasicontinuous laser excitation.

It is only recently that high thoughput UV Raman microspectrometers have been developed. It has been demonstrated for visible Raman spectroscopy that a significantly higher throughput can be achieved by employing a single stage spectrograph with a notch filter instead of a double or triple monochromator. In the UV region, blocking the Rayleigh scattering is a problem because notch filters are not currently available. However, it has been discovered that this can be overcome by introducing two modifications to the design of the optical layout of a visible Raman microscope. First, two novel dielectric longpass filters were used instead of a notch filter in order to reject the elastically scattered light. Second, an all-reflecting Cassegrain microscope objective was used, instead of a lens, in order to block the specular reflection, and thereby further reduce the Rayleigh scattering background. These design modifications have enabled a single stage spectrograph with a holographic grating to be used to disperse the Stokes Raman radiation in a UV Raman microspectrometer (**Fig. 10**).

In the optical layout of **Fig. 10** it can be seen that the laser excitation is not focused by the Cassegrain microscope objective because an epi-illumination configuration is not employed. Instead the laser is focused by a separate lens and directed on to the sample with a turning prism located directly underneath the Cassegrain objective. This minimizes loss of laser beam throughput or scattered light throughput to the spectrometer at the expense of spatial resolution, since the laser light is focused by a longer focal length lens than the microscope objective. An additional difference between the optical layouts of the visible and UV Raman microscopes shown in **Figs. 1** and **10**, respectively, is that pinhole spatial filters are not shown in the latter. However, better axial spatial resolution could be achieved if this UV Raman microscope was made confocal by introducing a pinhole and lens.

Frequency-Doubled Argon Ion Laser

The frequency doubled argon ion laser is a popular choice among UV laser sources. The cw UV laser contains a nonlinear optical beta-barium borate (BBO) crystal, which is located within the laser cavity; this crystal frequency doubles the strong Ar^+ lines to give five UV lines below 260 nm which are listed in **Table 1** for the Coherent Inc. Innova 300 Ar^+ ion laser system. Of these UV lines, the 228.9 nm line is almost ideal for resonance Raman enhancement of tyrosine and trpytophan residues in proteins and the 244.0 nm line is well suited for studying tyrosinate groups. The 244.0 nm line has also been used to excite selectively UV resonance Raman spectra from spatially resolved areas of biological samples within the nucleus of a single cell, from DNA in particular, using low laser powers and short acquisition times to keep sample damage to a minimum under the microscope.

Frequency-Doubled Krypton Ion Laser

Like the cw frequency-doubled argon ion laser mentioned above, the cw frequency-doubled krypton ion laser contains a nonlinear optical BBO crystal, which is located within the laser cavity; this crystal frequency doubles the strong Kr^+ lines to give 206.5 and 234 nm UV lines which are listed in **Table 2** for the Coherent Inc. Innova 400 Kr^+ ion laser system. The 206.5 nm laser line is useful for exciting resonance Raman spectra within the π–π^* amide transition of the peptide backbone of proteins, and the enhanced protein amide vibrational bands can be used to determine the protein secondary structure.

HeCd Laser

The 325 nm HeCd excitation line, in the near ultraviolet, can be used for combined micro-Raman/photoluminescence studies of, for example, semiconductors. These lasers are suitable for low-power applications, typically having output powers in the 1–100 mW range, and they can be designed to operate in TEM_{00} single-mode or multimode.

Quasi-CW Mode-Locked Ti:Sapphire Laser

The second harmonic of the mode-locked Ti:sapphire laser (ca. 350–500 nm), third harmonic (ca. 233–333 nm), and fourth harmonic (ca. 200–250 nm), are all available with conventional nonlinear crystals.

Typically, the Ti:sapphire crystal is pumped by a cw argon ion laser, for example, as the excitation source for a UV Raman microscope. This quasicontinuous laser system produces 2–3 ps pulses at a 76 MHz repetition rate and is continuously tunable over the 200–300 nm range. The typical power output is ca. 50 mW at 250 nm, ca. 10 mW at 240 nm, and 1 mW at 230 nm. In future, it is anticipated that a frequency-doubled Nd:YAG laser will be increasingly favored as the pump, creating an all-solid-state laser source.

Conclusions

Laser systems for Raman microscopy have been described in this article. Although the differences in the terms 'Raman spectroscopy' and 'Raman microscopy' only refer to whether a macro or micro sampling configuration is used, this does have a bearing on the types of laser systems that are employed. This is because the latter imposes a restriction on using low duty pulsed lasers with high pulse peak power, as both spatial and temporal confinement of the laser excitation can lead to dielectric breakdown of the sample, nonlinear phenomena, or Raman saturation. Thus, for this reason, one is limited to cw lasers or quasi-cw lasers for Raman microscopy whereas there is no such restriction for Raman spectroscopy.

The quality of the laser beam, measured by the M^2 value or 'times diffraction limit' influences the ultimate spatial resolution that can be achieved in Raman microscopy.

In the visible region, argon ion, krypton ion, and frequency-doubled Nd:YAG lasers have good beam quality and high power, with helium-neon and helium-cadmium lasers also having good beam quality but lower power; this makes these lasers good light sources for visible Raman microscopy.

For FT Raman microscopy using Nd:YAG excitation, there is a lack of fluorescence interference from a wide variety of samples, but sensitivity and spatial resolution are compromised in comparison to visible dispersive Raman microscopy.

Red and near infrared semiconductor lasers operating in TEM_{00} single mode can have good beam quality, approaching that of gas lasers, but they are power limited. Obviously, when good beam quality is not necessary, for example, for optically pumping other laser sources, high-power diode lasers can be used in multimode operation. Solid state red and near infrared laser sources are becoming very popular for routine Raman microscopic analysis due to their compactness, robustness, and economical power consumption.

Another reason for the popularity of the red and near infrared diode laser sources for Raman microscopy is that fluorescence is less of a problem due to the lower energy of the light, compared to conventional green 514.5 nm excitation. Although there is a penalty, due to the dependency of the scattering efficiency on the ν^4 term, present day silicon-based CCD arrays can exhibit excellent detection quantum efficiency of Stokes shifted Raman radiation, across the whole vibrational spectrum, when using standard 670 and 785 nm diode lasers and, at least across the fingerprint region, when using 830 and 852 nm diode lasers.

For ultimate spatial resolution in the Raman microscopic experiment, short wavelength excitation is advantageous because it is fundamentally possible to focus to a smaller diffraction-limited laser spot. In this regard, it would be preferable to use UV Raman microscopy where fluorescence is also not as problematic as it is in the visible region. However, in spite of recent improvements in UV laser sources such as the intracavity frequency-doubled argon and krypton ion cw lasers, there are current instrumental limitations on the throughput efficiency of the UV Raman microscope imposed by the lack of availability of suitable notch filters and low stray light and high throughput optics.

It is anticipated that in the future all-solid-state, quasi-cw, diode pumped, mode-locked Ti:sapphire lasers operating on their third or fourth harmonics will become increasingly popular for providing tunable UV excitation affording good beam quality. It is now also possible to fabricate hollow cathode metal ion deep UV lasers, 10–15 cm in length, 2–4 cm in diameter, weighing 50–100 g and requiring only 2–3 W of electrical power. These lasers have low beam quality (M^2 equal to ca. 18), but they make possible the development of portable UV fluorescence imaging and Raman microprobes for geobiological exploration of terrestrial and extraterrestrial environments.

Further Reading

Andrews, D.L., 1990. Lasers in Chemistry, 2nd edition Berlin, Heidelberg: Springer-Verlag.

Asher, S.A., Bormett, R.W., Chen, X.G., *et al*., 1993. UV resonance Raman spectroscopy using a new cw laser source: Convenience and experimental simplicity. Applied Spectroscopy 47, 628–633.

Bell, S.E.J., Bourguignon, E.S.O., O'Grady, A., Villaumie, J., Dennis, A.C., 2002. Extracting Raman spectra from highly fluorescent samples with "Scissors" (SSRS, Shifted-Subtracted Raman Spectroscopy). Spectroscopy Europe 14, 17.
Best, S.P., Clark, R.J.H., Withnall, R., 1992. Non-destructive pigment analysis of artefacts by Raman microscopy. Endeavour 16, 66.
Boustany, N.N., Manoharan, R., Dasari, R.R., Feld, M.S., 2000. Ultraviolet resonance Raman spectroscopy of bulk and microscopic human colon tissue. Applied Spectroscopy 54, 24–30.
Delhaye, M., Barbillat, J., Aubard, J., Bridoux, M., Da Silva, E., 1996. Instrumentation. In: Turrell, G., Corset, J. (Eds.), Raman Microscopy: Developments and Applications. London: Academic Press. chap. 3.
Delhaye, M., Dhamelincourt, P., 1975. Raman microprobe and microscope with laser excitation. Journal of Raman Spectroscopy 3, 33.
Derbyshire, A., Withnall, R., 1999. Pigment analysis of portrait miniatures using Raman microscopy. Journal of Raman Sprectroscopy 30, 185.
Gordeyev, S.A., Nikolaeva, G.Y., Prokhorov, K.A., Withnall, R., Dunkin, I.R., Shilton, S.J., 2001. Super-selective polysulfone hollow fiber membranes for gas separation: assessment of molecular orientation by Raman spectroscopy. Laser Physics 11, 82–85.
Hayward, C.L., Best, S.P., Clark, R.J.H., Ross, N.L., Withnall, R., 1994. Polarised single crystal Raman spectroscopy of sinhalite, $MgAlBO_4$. Spectrochimica Acta 50A, 1287–1294.
Hendra, P., Jones, C., Warnes, G., 1991. Fourier Transform Raman Spectroscopy. London: Ellis Horwood.
Hitz, B., 2003. Solid state lasers are gunning for argon ion's place. Photonics Spectra. 54–58.
Holtz, J.S.W., Bormett, R.W., Chi, Z., *et al.*, 1996. Applications of a new 206.5 nm continuous wave laser source: UV Raman determination of protein secondary structure and CVD diamond material properties. Applied Spectroscopy 50, 1459–1468.
http://www.btools.com.
Koechner, W., Bass, M., 2003. Solid-State Lasers. New York: Springer-Verlag.
Landsberg, G., Mandelstam, L., 1928. Eine neue Erschienung bei der Lichtzerstreaung in Krystallen. Naturwiss 16, 557.
Matousek, P., Towrie, M., Stanley, A., Parker, A.W., 1999. Efficient rejection of fluorescence from Raman spectra using picosecond Kerr gating. Applied Spectroscopy 53, 1485.
Nakashima, S., Okumura, H., Yamamoto, T., Shimidzu, R., 2004. Deep-ultraviolet Raman microspectroscopy: Characterization of wide-gap semiconductors. Applied Spectroscopy 58, 224–229.
Pajcini, V., Munro, C.H., Bormett, R.W., Witkowski, R.E., Asher, S.A., 1997. UV Raman microspectroscopy: Spectral and spatial selectivity and simplicity. Applied Spectroscopy 51, 81–86.
Pallister, D.M., Morris, M.D., 1993. In: Morris, M.D. (Ed.), Microscopic and Spectroscopic Imaging of the Chemical State. New York: Marcel Dekker. chap. 1.
Pan, M.-W., Benner, R.E., Johnson, C.W., Smith, L.M., 2000. Near-IR lasers rejuvenate Raman spectroscopy. Optoelectronics World. S5–S10.
Pan, M.-W., Benner, R.E., Smith, L.M., 2002. Continuous lasers for Raman spectrometry. In: Chalmers, J.M., Griffiths, P.R. (Eds.), Handbook of Vibrational Spectroscopy,. Chichester, UK: Wiley.
Raman, C.V., Krishnan, K.S., 1928. A new type of secondary radiation. Nature 121, 501.
Rosasco, G.J., 1980. Raman microprobe spectroscopy. In: Clark, R.J.H., Hester, R.E. (Eds.), Advances in Infrared and Raman Spectroscopy, vol. 7. Chichester: Wiley.
Rosasco, J., Etz, E.S., Cassatt, W.A., 1975. The analysis of discrete fine particles by Raman spectroscopy. Applied Spectroscopy 29, 396.
Schawlow, A.L., Townes, C.H., 1958. Infrared and optical masers. Physics Review 112, 1940.
Smekal, A., 1923. Sur Quantentheorie der Dispersion. Naturwiss 11, 873.
Storrie-Lombardi, M.C., Hug, W.F., McDonald, G.D., Tsapin, A.I., Nealson, K.H., 2001. Hollow cathode ion lasers for deep ultraviolet Raman spectroscopy and fluorescence imaging. Review of Scientific Instruments 72, 4452–4459.
Turrell, G., Delhaye, M., Dhamelincourt, P., 1996. Characteristics of Raman microscopy. In: Turrell, G., Corset, J. (Eds.), Raman Microscopy: Development and Applications 2. London: Academic Press.
Young, M., 1977. Optics and Laser: An Engineering Approach. Berlin, Heidelberg: Springer-Verlag.
Yu, Y.-M., Nam, S., Byungsung, O., *et al.*, 2002. Resonant Raman scattering in ZnS epilayers. Materials Chemistry and Physics 78, 149–153.

Spatial Heterodyne

Mark F Spencer, Albuquerque, NM, United States

Introduction

Many optics and photonics applications benefit from the use of spatial heterodyne, which the open literature often refers to as spatial-heterodyne interferometry, spatial-heterodyne detection, digital-holographic detection, digital holography, or coherent detection. This is said because spatial heterodyne provides us with an estimate of the complex field (i.e., both the amplitude and wrapped phase). As such, this encyclopedia article reviews the fundamentals of spatial heterodyne, so that future research efforts can benefit from the material presented within.

Spatial heterodyne, in practice, results from the interference between a signal and a reference. As described in **Fig. 1**, we can create the signal using the active illumination of an object. Note that the light for this active illumination comes from a master-oscillator (MO) laser. Also note that the surface of the object is often considered to be optically rough compared with the wavelength of the incident light; thus, the receiving aperture of an imaging system collects the resultant speckle (i.e., the constructive and destructive interference which results from the laser-object interaction and propagation to the receiving aperture). To make it all work, we need to then interfere the (signal) light from the speckle with the (reference) light from a local oscillator (LO). This LO, in general, is split off from the MO laser. We can then record the resultant interference pattern or hologram with a focal-plane array (FPA) and digitize it for digital-signal processing. It is this ability to perform digital-signal processing which makes spatial heterodyne practical and such a versatile tool within the optics and photonics research communities.

To ensure that the signal interferes with the reference, it is advantageous to use a MO laser with a coherence length that is greater than twice the distance to the object. For tactical applications, however, this specification becomes difficult to achieve. We can then attempt to path-length match the reference with the signal, for example, using fiber-optic relays and continuous-wave illumination with a coherence length that is greater than twice the depth of the object (McManamon, 2015). With that said, if path-length matching is impractical (and it often is), we can then use pulsed illumination and incur the efficiency losses associated with decreased fringe visibility in our digital holograms. In this regime, it is desirable (although not required) to use a MO laser with a coherence length that is greater than the pulse length (i.e., use transform-limited pulses).

Doppler shifts, fluorescence, and depolarization from the laser-object interaction, in addition to mismatches (in the reference-signal spatial overlap, pulse timing, etc.), extinction, and noise are other sources for efficiency losses that, in practice, degrade fringe visibility. With this in mind, this encyclopedia article uses a scalar formulation (and the assumptions therein) to study three separate recording geometries often used with spatial-heterodyne applications. As described in **Fig. 1**, these recording geometries include the off-axis image plane recording geometry (IPRG), the off-axis pupil plane recording geometry (PPRG), and the on-axis phase shifting recording geometry (PSRG). Each of these recording geometries has its benefits and drawbacks depending on the spatial-heterodyne application of interest. Thus, the goal for the following analysis is to study these recording geometries in detail, so that we can ultimately develop closed-form expressions for their signal-to-noise ratios (SNRs).

In what follows, we will explore the background information needed to investigate the three recording geometries described in **Fig. 1**. In particular, we will review the optics relationships needed to propagate from the object plane to the FPA and the photonics principles needed to record digital holograms with the FPA. Following these two background sections, there are detailed sections on the three recording geometries described in **Fig. 1**. A final section then contains an illustrative comparison of the different recording geometries and an appendix provides example MATLAB® code. The goal throughout is to show that we can process spatial-heterodyne measurements to obtain an estimate of the complex field. Provided the complex field, we can then pursue the spatial-heterodyne application of interest (e.g., wave-front sensing, coherent imaging, etc).

Background Material: Part I

With **Fig. 1** in mind, we need to further define the experimental parameter space. To help orient the reader, **Fig. 2** pictorially describes the various planes of interest within the analysis. The reader should note that the entrance pupil of our proposed imaging system effectively collimates the monochromatic light from the object plane, whereas the exit pupil effectively focuses the monochromatic light to form the image plane at focus.

In the background material that follows, we will develop a relationship that says that the pupil and image planes of our proposed imaging system form a Fourier-conjugate pair. As such, we can use spatial heterodyne to exploit this relationship and obtain an estimate of the desired complex field. Depending on the spatial-heterodyne application of interest (e.g., wavefront sensing, coherent imaging, etc.), this estimate is what makes spatial heterodyne such a versatile tool within the optics and photonics research communities.

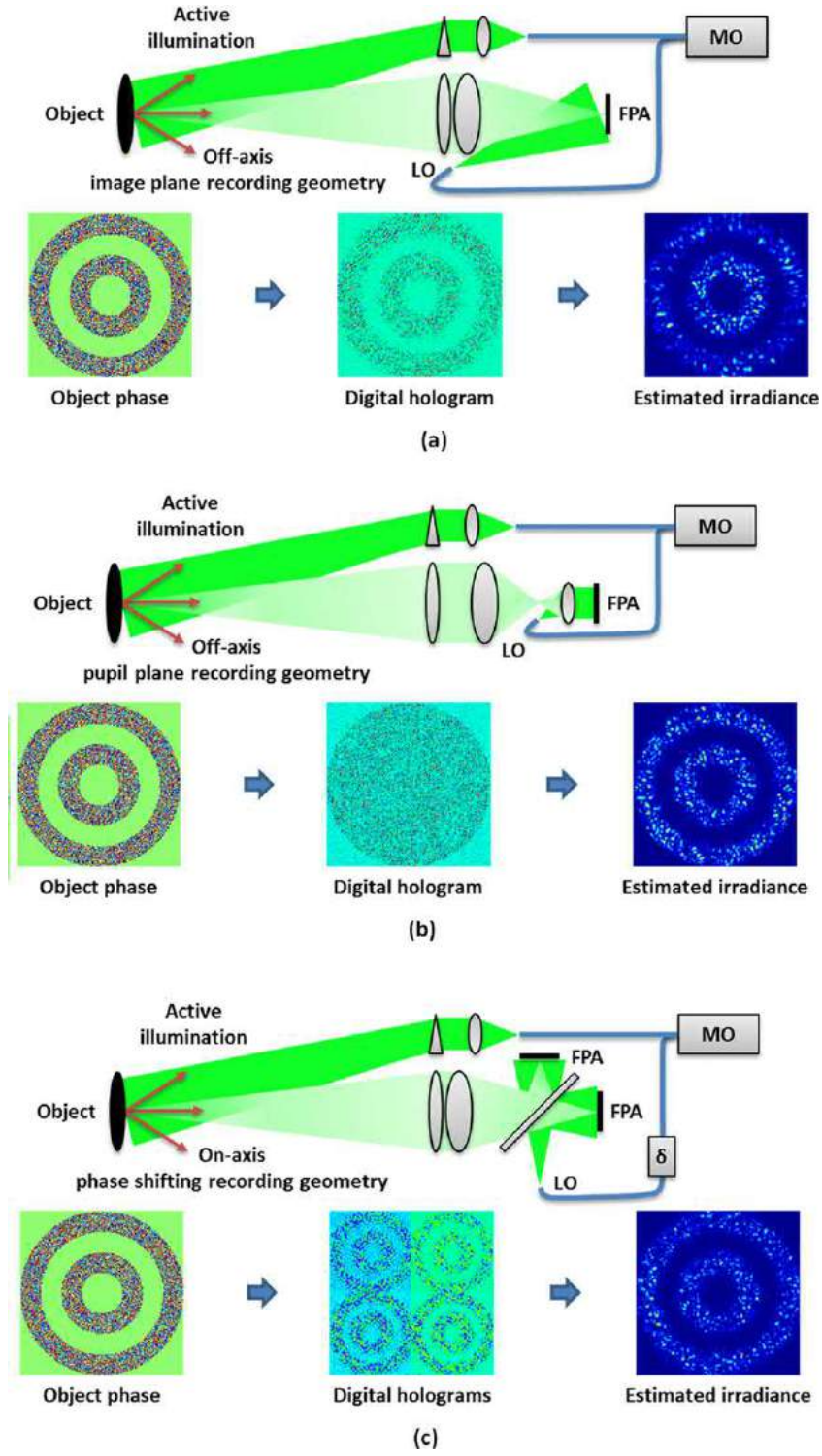


Fig. 1 A description of three spatial-heterodyne recording geometries.

Object Plane

Depending on the spatial-heterodyne application of interest, the object plane of the proposed imaging system takes on different forms. For example, given a point-source object (which is desirable for wavefront sensing),

$$U_O^+(x_0, y_0) = a_S e^{-jkz_P} \delta(x_0)\delta(x_0) \tag{1}$$

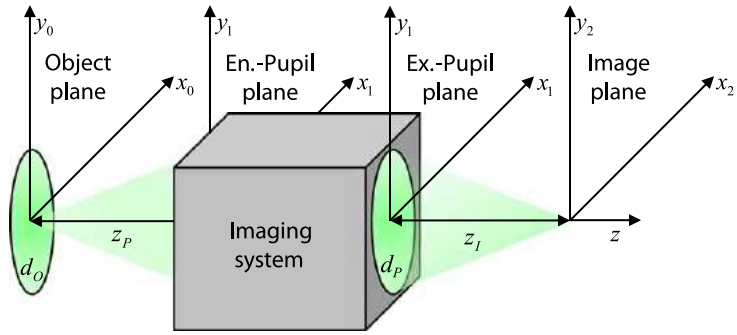


Fig. 2 A description of an imaging system.

where $U_O^+(x_0,y_0)$ is the complex field leaving the object plane, a_s is a real constant, $k=2\pi/\lambda$ is the angular wavenumber, λ is the wavelength, z_P is the distance to the object plane from the pupil plane, and $\delta(x)$ is an impulse function (cf. Eq. (80) in Appendix A). This point-source object gives rise to an ideal-spherical wave which propagates through the various planes of interest in our proposed imaging system (cf. **Fig. 2**). In the image plane, the irradiance associated with our imaged point source becomes a very special quantity known as the point-spread function (PSF) and is a quantity that we can measure with our FPA. In the presence of no aberrations, the PSF is an Airy disk (given a circular entrance/exit pupil), whereas in the presence of isoplanatic phase aberrations, the rings of the Airy disk start to wash out. For spatial-heterodyne applications, the sampling of the PSF with the FPA pixels becomes an important variable within the analysis and is a point that we will explore in more detail in the section on the off-axis IPRG.

For an extended object (which is desirable for coherent imaging),

$$U_O^+(x_0,y_0)=R_O(x_0,y_0)U_O^-(x_0,y_0) \tag{2}$$

where $R_O(x_0,y_0)=U_O^+(x_0,y_0)/U_O^-(x_0,y_0)$ is the object-plane complex reflectance function and $U_O^-(x_0,y_0)$ is the complex field entering the object plane. Let's assume that we actively illuminate the object plane with an uniform-amplitude, on-axis plane wave, so that

$$U_O^-(x_0,y_0)=A_S e^{-jkz_P} \tag{3}$$

where A_s is a complex constant. In addition, let's assume that the surface of the object that we are actively illuminating is optically rough compared to the wavelength of the monochromatic light (Spencer, 2014). If we then assume that the optically rough surface is delta correlated (Goodman, 2007),

$$R_O(x_0,y_0)=O(x_0,y_0)e^{j\varphi_1(x_0,y_0)} \tag{4}$$

where $O(x_0,y_0)$ is the generalized complex object function and $\varphi_k(x,y)$ is the kth realization of uniformly distributed, real-valued random numbers in the interval $[-\pi,\pi]$. In practice, Eqs. (2)–(4) provide us with a phase-screen model for rough-surface scattering that gives the correct speckle-correlation statistics upon propagation from the object plane to the pupil plane. For spatial-heterodyne applications, the sampling of the speckle-correlation statistics with the FPA pixels becomes an important variable within the analysis and is a point that we will explore in more detail in the section on the off-axis PPRG.

Pupil Plane

Using the convolution form of the Fresnel diffraction integral, we can represent the complex field $U_P^-(x_1,y_1)$ entering the pupil plane as

$$U_P^-(x_1,y_1)=\frac{e^{jkz_P}}{j\lambda z_P}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}U_O^+(x_1,y_1)\exp\left(j\frac{k}{2z_P}\left[(x_1-x_0)^2+(y_1-y_0)^2\right]\right)dx_0dy_0 \tag{5}$$

where again, $k=2\pi/\lambda$ is the angular wavenumber, λ is the wavelength, z_P is the distance to the pupil plane from the object plane, and $U_O^+(x_0,y_0)$ is the complex field leaving the object plane (cf. Eqs. (1) and (2) above).

As mentioned previously, the entrance pupil of the proposed imaging system effectively collimates the monochromatic light from the object plane, whereas the exit pupil effectively focuses the monochromatic light to form the image plane at focus (cf. **Fig. 2**). With this in mind, we can represent the complex field $U_P^+(x_1,y_1)$ leaving the pupil plane as

$$U_P^+(x_1,y_1)=T_P(x_1,y_1)U_P^-(x_1,y_1) \tag{6}$$

where

$$T_P(x_1,y_1)=\frac{U_P^+(x_1,y_1)}{U_P^-(x_1,y_1)}=\exp\left[-j\frac{k}{2z_P}\left(x_1^2+y_1^2\right)\right]\exp\left[-j\frac{k}{2z_I}\left(x_1^2+y_1^2\right)\right]P(x_1,y_1) \tag{7}$$

is the pupil-plane complex transmittance function, z_I is the distance to the image plane from the pupil plane, and

$$P(x_1,y_1)=\mathrm{cyl}\left(\frac{\sqrt{x_1^2+y_1^2}}{d_P}\right)e^{j\phi(x_1,y_1)} \tag{8}$$

is the generalized complex pupil function. In Eq. (8), $\mathrm{cyl}\left(\sqrt{x^2+y^2}\right)$ is a cylinder function (cf. Eq. (81) in Appendix A), d_P is the exit-pupil diameter, and $\phi(x_1,y_1)$ is the isoplanatic phase function, which represents all of the linear, shift-invariant phase aberrations present within the imaging system (Gaskill, 1978).

Substituting Eq. (7) into Eq. (6), we arrive at the following relationship:

$$U_P^+(x_1,y_1)=\exp\left[-j\frac{k}{2z_I}\left(x_1^2+y_1^2\right)\right]U_P(x_1,y_1) \tag{9}$$

Here,

$$U_P(x_1,y_1)=\exp\left[-j\frac{k}{2z_P}\left(x_1^2+y_1^2\right)\right]P(x_1,y_1)U_P^-(x_1,y_1) \tag{10}$$

is the pupil-plane complex field (i.e., the complex field that exists in the pupil plane of the proposed imaging system). It is important to note that $U_P(x_1,y_1)$ is being multiplied by a positive-thin-lens complex transmittance function in Eq. (9). This multiplication will allow us to form an image at focus upon propagation from the pupil plane to the image plane – our going-in disposition.

Image Plane

Provided Eq. (9), we can now account for the image-plane complex field $U_I(x_2,y_2)$ (i.e., the complex field that exists in the image plane of the proposed imaging system). In particular,

$$U_I(x_2,y_2)=\frac{e^{jkz_I}}{j\lambda z_I}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}U_P^+(x_1,y_1)\exp\left(j\frac{k}{2z_I}\left[(x_2-x_1)^2+(y_2-y_1)^2\right]\right)dx_1dy_1 \tag{11}$$

which again makes use of the convolution form of the Fresnel diffraction integral. Substituting Eq. (9) into Eq. (11), we arrive at the following relationship:

$$U_I(x_2,y_2)=\frac{e^{jkz_I}}{j\lambda z_I}\exp\left[j\frac{k}{2z_I}\left(x_2^2+y_2^2\right)\right]\mathcal{F}\{U_P(x_1,y_1)\}_{\nu_x=\frac{x_2}{\lambda z_I},\nu_y=\frac{y_2}{\lambda z_I}}=\frac{e^{jkz_I}}{j\lambda z_I}\exp\left[j\frac{k}{2z_I}\left(x_2^2+y_2^2\right)\right]\tilde{U}_P\left(\frac{x_2}{\lambda z_I},\frac{y_2}{\lambda z_I}\right) \tag{12}$$

This relationship says that the pupil plane and image plane form a Fourier-conjugate pair. Moving forward we can use spatial heterodyne to exploit this relationship and obtain an estimate, $\hat{U}_P(x_1,y_1)$ or $\hat{U}_I(x_2,y_2)$, of the desired complex field, $U_P(x_1,y_1)$ or $U_I(x_2,y_2)$, that exist in the pupil or image plane of our proposed imaging system.

Background Material: Part II

In the background material that follows, we will derive a relationship for the hologram photoelectron density that, in practice, a FPA records via spatial-heterodyne measurements. Provided **Fig. 3**, it is important to note that the spatial-heterodyne recording geometry is what ultimately determines where we place the FPA within the analysis. For example, in the off-axis IPRG we must place the FPA in the image plane of the proposed imaging system, whereas in the off-axis PPRG we must place the FPA in the pupil plane (or a plane that is conjugate to the pupil plane). The on-axis PSRG is unique in the sense that we do not have to go to the Fourier plane in order to obtain an estimate. Therefore, we can place the FPA in either the image or pupil plane of our proposed imaging system. In subsequent sections, we will explore the differences and similarities between the various spatial-heterodyne recording geometries. With that said, it is important to acknowledge their subtle differences up front, so that the following background material becomes more tractable.

Hologram Irradiance

In units of Watts per square meter (W-m^{-2}), we can determine the real-valued hologram irradiance $i_H(x,y)$ that is incident on the FPA as

$$i_H(x,y)=|U_S(x,y)+U_R(x,y)|^2=|U_S(x,y)|^2+|U_R(x,y)|^2+U_S(x,y)U_R^*(x,y)+U_R(x,y)U_S^*(x,y) \tag{13}$$

Here, $U_S(x,y)$ and $U_R(x,y)$ are, respectively, the signal and reference complex fields that are incident on the FPA and the superscript * denotes complex conjugate. It is important to note that in Eq. (13) we have made the assumption that the spatial-heterodyne

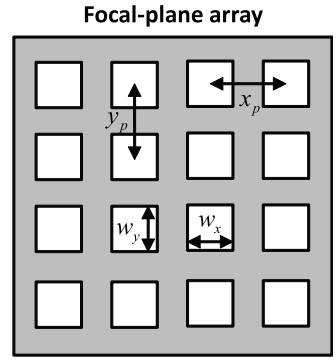


Fig. 3 Nominal layout for a focal-plane array or FPA.

measurements are fast enough so that $U_S(x,y)$ and $U_R(x,y)$ do not change significantly over the integration time of the FPA. Consequently, there is no need to include a time-dependent variable within the analysis.

For all intents and purposes, the FPA will convert the hologram irradiance $i_H(x,y)$, which is in an analog form, into a form that is suitable for digital-signal processing. Following the approach taken by Gaskill (1978), let's assume that "digitization" is to take place at sampling intervals of x_p and y_p, which are, respectively, the x- and y-axis pixel pitches of the FPA (cf. **Fig. 3**). At any particular FPA pixel, we can then obtain an estimate, $\hat{\imath}_H(x,y)$, of the hologram irradiance, $i_H(x,y)$, by computing its average value over the active area of the FPA pixel, which is centered at $x=\mathrm{n}x_p$ and $y=\mathrm{m}y_p$, where $\mathrm{n}=1,2,\ldots,\mathrm{N}$ and $\mathrm{m}=1,2,\ldots,\mathrm{M}$. Specifically,

$$\hat{\imath}_H(\mathrm{n}x_p,\mathrm{m}y_p)=\frac{1}{w_xw_y}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}i_H(x',y')\mathrm{rect}\left(\frac{x'-\mathrm{n}x_p}{w_x}\right)\mathrm{rect}\left(\frac{y'-\mathrm{m}y_p}{w_y}\right)dx'dy' \tag{14}$$

where w_x and w_y are, respectively, the x- and y-axis pixel widths of the FPA, and $\mathrm{rect}(x)$ is a rectangle function (cf. Eq. (78) in Appendix A).

Hologram Mean Number of Photoelectrons

At any particular FPA pixel, the real-valued hologram mean number of photoelectrons $\overline{m}_H(\mathrm{n}x_p,\mathrm{m}y_p)$ results from random photoevents (i.e., when the FPA pixel creates an electron–hole pair or liberates a photoelectron (Saleh and Teich, 2007)). Therefore, over the integration time of the FPA,

$$\overline{m}_H(\mathrm{n}x_p,\mathrm{m}y_p)=\frac{\eta\tau}{h\nu}w_xw_y\hat{\imath}_H(\mathrm{n}x_p,\mathrm{m}y_p) \tag{15}$$

where η is the quantum efficiency, τ is the integration time, $h\nu$ is the quantized photon energy, and w_xw_y is the active area of the FPA pixel. It is important to note that in Eq. (15) we have made the assumption that the measurements are without noise.

Following the approach taken by Tippie (2012), noise is additive. As such,

$$\overline{m}_H^{+}(\mathrm{n}x_p,\mathrm{m}y_p)=\mathcal{P}\{\overline{m}_H(\mathrm{n}x_p,\mathrm{m}y_p)\}+\sigma_r n_1(\mathrm{n}x_p,\mathrm{m}y_p) \tag{16}$$

where $\mathcal{P}\{\circ\}$ denotes a Poisson-noise operator, σ_r is the read-noise standard deviation, and $n_k(x,y)$ is the kth realization of real-valued, zero-mean, unit-variance Gaussian random numbers. For most spatial-heterodyne applications, $\overline{m}_H(\mathrm{n}x_p,\mathrm{m}y_p)\gg 1$ because of the use of an LO in creating the reference complex field $U_R(x,y)$. Thus, to a very good approximation, we can rewrite Eq. (16) as

$$\overline{m}_H^{+}(\mathrm{n}x_p,\mathrm{m}y_p)\approx\overline{m}_H(\mathrm{n}x_p,\mathrm{m}y_p)+\sigma_s(\mathrm{n}x_p,\mathrm{m}y_p)n_2(\mathrm{n}x_p,\mathrm{m}y_p)+\sigma_r n_1(\mathrm{n}x_p,\mathrm{m}y_p) \tag{17}$$

where $\sigma_s(\mathrm{n}x_p,\mathrm{m}y_p)$ is the shot-noise standard deviation. This approximation says that when $\overline{m}_H(\mathrm{n}x_p,\mathrm{m}y_p)\gg 1$, our Poisson-distributed random process becomes a Gaussian-distributed random process with mean $\overline{m}_H(\mathrm{n}x_p,\mathrm{m}y_p)$ and variance $\sigma_s^2(\mathrm{n}x_p,\mathrm{m}y_p)$. Note that for Poisson-distributed random processes, the variance is equal to the mean (Dereniak and Boreman, 1996). In turn,

$$\sigma_s^2(\mathrm{n}x_p,\mathrm{m}y_p)=\overline{m}_S(\mathrm{n}x_p,\mathrm{m}y_p)+\overline{m}_R(\mathrm{n}x_p,\mathrm{m}y_p)+\overline{m}_B(\mathrm{n}x_p,\mathrm{m}y_p) \tag{18}$$

where $\overline{m}_S(\mathrm{n}x_p,\mathrm{m}y_p)$, $\overline{m}_R(\mathrm{n}x_p,\mathrm{m}y_p)$, and $\overline{m}_B(\mathrm{n}x_p,\mathrm{m}y_p)$ are, respectively, the mean-number of photoelectrons associated with the signal, reference, and background illumination that is incident on the FPA. Also note that in Eq. (17), we have made the assumption that pixel to pixel both the shot and read noise result from statistically independent (i.e., delta correlated) random processes (Frieden, 2001). This assumption is a sound one since the sources for the shot noise (e.g., the object, LO, and sun) and read noise (i.e., the read out integrated circuit (ROIC) of the FPA) are physically separated from one another.

Hologram Photoelectron Density

At any particular FPA pixel, a real-valued hologram digital number $d_H(\mathrm{n}x_p,\mathrm{m}y_p)$ results via an analog-to-digital (A/D) conversion (Janesick, 2007). From Eqs. (15)–(18), it then follows that

$$d_H^+(\mathrm{n}x_p,\mathrm{m}y_p) = g_{\mathrm{A/D}}\overline{m}_H^+(\mathrm{n}x_p,\mathrm{m}y_p) + \sigma_q n_3(\mathrm{n}x_p,\mathrm{m}y_p) \tag{19}$$

where $g_{\mathrm{A/D}}$ is the A/D gain factor in units of digital number per photoelectron (DN-pe^{-1}) and σ_q is the quantization-noise standard deviation. In practice,

$$\sigma_q^2 = \frac{1}{12} \tag{20}$$

and the factor of 12 in the denominator comes from rounding to the nearest integer. Before moving on in the analysis, it is important to note that the effects of dark-current, $1/f$, leakage, fano, and fixed-pattern noise are not included in Eq. (19). Under some circumstances, these noise effects might become important (Dereniak and Boreman, 1996; Saleh and Teich, 2007; Janesick, 2007); however, they are beyond the scope of the present analysis.

As previously stated, for most spatial-heterodyne applications, $\overline{m}_H(\mathrm{n}x_s,\mathrm{m}y_s) \gg 1$ because of the use of an LO in creating the reference complex field $U_R(x,y)$ (cf. Eqs. (13)–(15)). For example, given an ideal reference with uniform irradiance, where $|U_R(x,y)|^2 = |A_R|^2$, we can determine the reference mean number of photoelectrons $\overline{m}_R(\mathrm{n}x_p,\mathrm{m}y_p) = \overline{m}_R$ from the following relationship:

$$\overline{m}_R = \frac{\eta\tau}{h\nu} w_x w_y |A_R|^2 \tag{21}$$

Provided Eq. (21), we can then dictate that $\overline{m}_R \gg \overline{m}_S(\mathrm{n}x_p,\mathrm{m}y_p)$, $\overline{m}_R \gg \overline{m}_B(\mathrm{n}x_p,\mathrm{m}y_p)$, and $\overline{m}_R \gg g_{\mathrm{A/D}}^{-1}\sigma_q$ within the constraints of the FPA pixel well depth ℓ (i.e., $\ell \geq \overline{m}_H(\mathrm{n}x_p,\mathrm{m}y_p)$, so that the FPA pixels do not saturate). As such, $\sigma_s^2(\mathrm{n}x_p,\mathrm{m}y_p) \approx \overline{m}_R$ (cf. Eq. (18)), and Eq. (19) simplifies, such that

$$d_H^+(\mathrm{n}x_p,\mathrm{m}y_p) \approx g_{\mathrm{A/D}}\left[\overline{m}_H(\mathrm{n}x_p,\mathrm{m}y_p) + \sqrt{\overline{m}_R}n_2(\mathrm{n}x_p,\mathrm{m}y_p) + \sigma_r n_1(\mathrm{n}x_p,\mathrm{m}y_p)\right] \tag{22}$$

Dividing both sides of Eq. (22) by the A/D gain factor $g_{\mathrm{A/D}}$, we obtain the following relationship:

$$\overline{m}_H^+(\mathrm{n}x_p,\mathrm{m}y_p) \approx \overline{m}_H(\mathrm{n}x_p,\mathrm{m}y_p) + \sigma_n n_4(\mathrm{n}x_p,\mathrm{m}y_p) \tag{23}$$

where

$$\sigma_n^2 \approx \overline{m}_R + \sigma_r^2 \tag{24}$$

is the noise variance. This relationship says that we only have to account for photoelectrons within the analysis.

Neglecting the effects of pixel-edge diffusion in the FPA (Poon and Liu, 2014), the hologram photoelectron density $\rho_H(x_2,y_2)$, in units of pe-m^{-2}, is simply a sampled version of the analog form of Eq. (23). This declaration leads us to the following relationship:

$$\rho_H^+(x,y) = [\overline{m}_H(x,y) + \sigma_n n_4(x,y)]\frac{1}{x_p}\mathrm{comb}\left(\frac{x}{x_p}\right)\frac{1}{y_p}\mathrm{comb}\left(\frac{y}{y_p}\right)\mathrm{rect}\left(\frac{x}{\mathrm{N}x_p}\right)\mathrm{rect}\left(\frac{y}{\mathrm{M}y_p}\right) \tag{25}$$

where

$$\overline{m}_H(x,y) = \frac{\eta\tau}{h\nu}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} i_H(x',y')\mathrm{rect}\left(\frac{x'-x}{w_x}\right)\mathrm{rect}\left(\frac{y'-y}{w_y}\right)dx'dy' = \frac{\eta\tau}{h\nu} i_H(x,y) ** \mathrm{rect}\left(\frac{x}{w_x}\right)\mathrm{rect}\left(\frac{y}{w_y}\right) \tag{26}$$

is the analog form of Eq. (15) and comb(x) is a comb function (cf. Eq. (79) in Appendix A). The reader should note that in Eq. (26), ** denotes a 2-D convolution, as defined in Eq. (83) in Appendix B. Moving forward we will use Eq. (25) to explore the differences and similarities between the various spatial-heterodyne recording geometries.

Off-Axis Image Plane Recording Geometry

The goal for the following analysis is to model spatial heterodyne in the off-axis IPRG. With that said, we can represent the signal complex field $U_S(x,y)$ that is incident on the FPA as

$$U_S(x,y) = U_I(x_2,y_2) \tag{27}$$

where again, $U_I(x_2,y_2)$ is the image-plane complex field (cf. Eq. (12)). In addition, if by choice we inject an off-axis LO, then by choice we can represent the reference complex field $U_R(x,y)$ that is incident FPA as

$$U_R(x,y) = U_R(x_2,y_2) = A_R e^{jkz_I}\exp\left[j\frac{k}{2z_I}(x_2^2+y_2^2)\right]\exp\left(j2\pi x_R\frac{x_2}{\lambda z_I}\right)\exp\left(j2\pi y_R\frac{y_2}{\lambda z_I}\right) \tag{28}$$

where here, A_R is a complex constant and (x_R,y_R) are the coordinates of the off-axis LO. It is important to note that the relationship given in Eq. (28) is the Fresnel approximation to a tilted spherical wave.

Fourier Plane

From Eqs. (1)–(28), we can gain access to an estimate, $\hat{U}_P(x_1,y_1)$, of the pupil-plane complex field (cf. Eq. (10)), $U_P(x_1,y_1)$, by going to the Fourier plane. To go to the Fourier plane, we first let $x=x_2=\lambda z_I v_x$ and $y=y_2=\lambda z_I v_y$. Next, we apply a 2-D inverse Fourier transformation, as defined in Eq. (85) in Appendix B, to the relationship contained in Eq. (25), so that

$$\mathcal{F}^{-1}\left\{\rho_H^+\left(\lambda z_I v_x,\lambda z_I v_y\right)\right\}_{x_1,y_1}=\frac{1}{\lambda^2 z_I^2}\tilde{\rho}_H^+\left(\frac{-x_1}{\lambda z_I},\frac{-y_1}{\lambda z_I}\right)=\left[\frac{\eta\tau}{h\nu}w_x\mathrm{sinc}\left(\frac{w_x x_1}{\lambda z_I}\right)w_y\mathrm{sinc}\left(\frac{w_y y_1}{\lambda z_I}\right)\mathcal{F}^{-1}\left\{i_H\left(\lambda z_I v_x,\lambda z_I v_y\right)\right\}_{x_1,y_1}\right.$$
$$\left.+\sigma_n\mathcal{F}^{-1}\left\{n_4\left(\lambda z_I v_x,\lambda z_I v_y\right)\right\}_{x_1,y_1}\right]**\frac{1}{\lambda^2 z_I^2}\mathrm{comb}\left(\frac{x_p x_1}{\lambda z_I}\right)\mathrm{comb}\left(\frac{y_p y_1}{\lambda z_I}\right)**\frac{\mathrm{N}x_p}{\lambda z_I}\mathrm{sinc}\left(\frac{\mathrm{N}x_p x_1}{\lambda z_I}\right)\frac{\mathrm{M}y_p}{\lambda z_I}\mathrm{sinc}\left(\frac{\mathrm{M}y_p y_1}{\lambda z_I}\right) \tag{29}$$

where $\mathrm{sinc}(x)$ is a sinc function (cf. Eq. (82) in Appendix A). Taking a look at the remaining 2-D inverse Fourier transformations in Eq. (29), we obtain the following relationship with respect to the hologram irradiance $i_H(x,y)$ (cf. Eqs. (12), (13), and (28)):

$$\mathcal{F}^{-1}\left\{i_H\left(\lambda z_I v_x,\lambda z_I v_y\right)\right\}_{x_1,y_1}=\frac{1}{\lambda^2 z_I^2}U_P(x_1,y_1)**U_P^*(-x_1,-y_1)$$
$$+\left|A_R\right|^2\delta(x_1)\delta(y_1)+\frac{A_R^*}{j\lambda z_I}U_P(x_1-x_R,y_1-y_R)-\frac{A_R}{j\lambda z_I}U_P^*(x_1+x_R,y_1+y_R) \tag{30}$$

This relationship contains the desired pupil-plane complex field $U_P(x_1,y_1)$.

To see that this last statement is true, let's analyze the various terms contained in the right-hand side of Eq. (30). The first term is an amplitude-scaled, 2-D autocorrelation. This term is centered on axis and is physically twice the circumference of the exit-pupil diameter d_P. The second term is also centered on axis and contains separable impulse functions (cf. Eq. (80) in Appendix A). These impulse functions are at the strength of the ideal reference with uniform irradiance (i.e., $|A_R|^2$). The last two terms form a complex-conjugate pair and contain the pupil-plane complex field $U_P(x_1,y_1)$, both scaled in amplitude and shifted off axis.

Substituting Eq. (30) into Eq. (29), we obtain the following relationship:

$$\tilde{\rho}_H^+\left(\frac{-x_1}{\lambda z_I},\frac{-y_1}{\lambda z_I}\right)=\left\{\frac{\eta\tau}{h\nu}w_x\mathrm{sinc}\left(\frac{w_x x_1}{\lambda z_I}\right)w_y\mathrm{sinc}\left(\frac{w_y y_1}{\lambda z_I}\right)\left[\frac{1}{\lambda^2 z_I^2}U_P(x_1,y_1)**U_P^*(-x_1,-y_1)+\left|A_R\right|^2\delta(x_1)\delta(y_1)\right.\right.$$
$$\left.\left.+\frac{A_R^*}{j\lambda z_I}U_P(x_1-x_R,y_1-y_R)-\frac{A_R}{j\lambda z_I}U_P^*(x_1+x_R,y_1+y_R)\right]+\sigma_n\mathcal{F}^{-1}\left\{n_4\left(\lambda z_I v_x,\lambda z_I v_y\right)\right\}_{x_1,y_1}\right\}**\mathrm{comb}\left(\frac{x_p x_1}{\lambda z_I}\right)\mathrm{comb}\left(\frac{y_p y_1}{\lambda z_I}\right)$$
$$**\frac{\mathrm{N}x_p}{\lambda z_I}\mathrm{sinc}\left(\frac{\mathrm{N}x_p x_1}{\lambda z_I}\right)\frac{\mathrm{M}y_p}{\lambda z_I}\mathrm{sinc}\left(\frac{\mathrm{M}y_p y_1}{\lambda z_I}\right) \tag{31}$$

which is in units of pe. This relationship is remarkably physical, as the sampling theorem dictates that a sampled function becomes periodic upon finding its spectrum (Gaskill, 1978). In turn, the 2-D convolution with the separable comb functions causes the terms contained within the squiggly brackets in Eq. (31) to repeat at intervals of $\lambda z_I/x_p$ and $\lambda z_I/y_p$ along the x and y axes, respectively. This periodicity occurs because of the convolution-sifting property of the impulse function (Gaskill, 1978). Lastly, the final 2-D convolution with the separable narrow sinc functions serves to smooth out these repeated terms.

Sampling Quotient

To help simplify the analysis to a case that we can easily simulate using N × N computational grids, let's assume that the FPA has adjacent square pixels, so that $x_p=y_p=w_x=w_y=p$. In so doing, we can rewrite Eq. (31) in terms of the image-plane sampling quotient q_I, where

$$q_I=\frac{\lambda z_I}{p d_P} \tag{32}$$

For all intents and purposes, q_I is a measure for the number of FPA pixels across the half width of the PSF. Recall that for linear, shift-invariant imaging systems, the PSF is the irradiance associated with an imaged point source (Gaskill, 1978). In turn, the relationship given in Eq. (32) allows us to vary the sampling with the FPA pixels.

Using Eq. (32), we can rewrite Eq. (31) in terms of the image-plane sampling quotient q_I, such that

$$\tilde{\rho}_H^+\left(\frac{-x_1}{\lambda z_I},\frac{-y_1}{\lambda z_I}\right)=\left\{\frac{\eta\tau}{h\nu}p^2\mathrm{sinc}\left(\frac{x_1}{q_I d_P}\right)\mathrm{sinc}\left(\frac{y_1}{q_I d_P}\right)\left[\frac{1}{\lambda^2 z_I^2}U_P(x_1,y_1)**U_P^*(-x_1,-y_1)+\left|A_R\right|^2\delta(x_1)\delta(y_1)\right.\right.$$
$$\left.\left.+\frac{A_R^*}{j\lambda z_I}U_P(x_1-x_R,y_1-y_R)-\frac{A_R}{j\lambda z_I}U_P^*(x_1+x_R,y_1+y_R)\right]+\sigma_n\mathcal{F}^{-1}\left\{n_4\left(\lambda z_I v_x,\lambda z_I v_y\right)\right\}_{x_1,y_1}\right\}**\mathrm{comb}\left(\frac{x_1}{q_I d_P}\right)\mathrm{comb}\left(\frac{y_1}{q_I d_P}\right)$$
$$**\frac{\mathrm{N}^2}{q_I^2 d_P^2}\mathrm{sinc}\left(\frac{\mathrm{N}x_1}{q_I d_P}\right)\mathrm{sinc}\left(\frac{\mathrm{N}y_1}{q_I d_P}\right) \tag{33}$$

Here, $q_I d_P=\lambda z_I/p$ is the side length of the N × N computational grid in the Fourier plane. Note that as $N\to\infty$ (cf. Eq. (80) in Appendix A), we can make use of the convolution-sifting property of the impulse function and neglect the final 2-D convolution in Eq. (33). Accordingly, for large N the smoothing caused by the final 2-D convolution in Eq. (33) becomes minimized; however, for small N the smoothing becomes more pronounced. Let's assume that $x_R=y_R=q_I d_P/4$, so that the last two terms within the square brackets in

Eq. (33) shift diagonally. When $q_I \geq 4$, the last two terms no longer overlap with the first two terms which are centered on axis. Correspondingly, when $2 \leq q_I < 4$, the last two terms are still resolvable within the side length of the $N \times N$ computational grid but overlap with the first term. This latter case allows for us to obtain more samples across the exit-pupil diameter d_P which in turn minimizes the smoothing caused by the final 2-D convolution in Eq. (33) but increases the noise sampling. Also note that this functional overlap becomes negligible when the amplitude of the reference $|A_R|$ is dominate (in comparison to the other amplitude terms).

Estimate

Provided Eqs. (32) and (33), we must shift the Fourier-plane data and apply a window function to obtain an estimate, $\hat{U}_P(x_1, y_1)$, of the pupil-plane complex field, $U_P(x_1, y_1)$. Specifically,

$$\hat{U}_P(x_1, y_1) = w(x_1, y_1)\, \tilde{\rho}_H^{+}\left(\frac{-x_1 - x_R}{\lambda z_I}, \frac{-y_1 - y_R}{\lambda z_I}\right) \tag{34}$$

where $w(x_1, y_1)$ is the window function. In using Eq. (34), we must satisfy Nyquist sampling with the FPA pixels (Gaskill, 1978), so that the repeated terms within Eq. (33) do not overlap and cause significant aliasing. As such, the Nyquist rate is $q_I d_P = \lambda z_I / p$ and the Nyquist interval is $1/(q_I d_P) = p/(\lambda z_I)$ when $x_R = y_R = q_I d_P / 4$.

Moving forward let's assume that we satisfy Nyquist sampling. In addition, let's assume that $q_I \geq 2$, $|A_R|$ is dominant, and

$$w(x_1, y_1) = \mathrm{cyl}\left(\frac{\sqrt{x_1^2 + y_1^2}}{d_P}\right) \tag{35}$$

In so doing, Eq. (34) simplifies, such that

$$\begin{aligned}\hat{U}_P(x_1, y_1) \approx q_I^2 d_P^2 \Bigg[&\frac{\eta\tau}{h\nu} p^2 \mathrm{sinc}\left(\frac{x_1 + x_R}{q_I d_P}\right) \mathrm{sinc}\left(\frac{y_1 + y_R}{q_I d_P}\right) \frac{A_R^*}{j\lambda z_I} U_P(x_1, y_1) + \sigma_n w(x_1, y_1) \mathcal{F}^{-1}\left\{ n_4\left(\lambda z_I \nu_x, \lambda z_I \nu_y\right)\right\}_{x_1 + x_R, y_1 + y_R} \Bigg] \\ &** \delta(x_1)\delta(y_1) ** \frac{\mathrm{N}^2}{q_I^2 d_P^2} \mathrm{sinc}\left(\frac{\mathrm{N}x_1}{q_I d_P}\right) \mathrm{sinc}\left(\frac{\mathrm{N}y_1}{q_I d_P}\right)\end{aligned} \tag{36}$$

The reader should note that as $N \to \infty$ (cf. Eq. (80) in Appendix A), we can again make use of the convolution-sifting property of the impulse function and neglect the final 2-D convolutions in Eq. (36). Accordingly, for large N the smoothing caused by the final 2-D convolution in Eq. (36) becomes minimized (and the noise remains delta correlated); however, for small N the smoothing becomes more pronounced (and the noise is no longer delta correlated).

Assuming that N is large, the estimate $\hat{U}_P(x_1, y_1)$ simplifies (cf. Eq. (36)), such that

$$\hat{U}_P(x_1, y_1) \approx \frac{\hat{\eta}_I(x_1, y_1)\tau}{h\nu} p^2 A_R^* U_P(x_1, y_1) + \frac{\sigma_I}{\sqrt{2}} N_1(x_1, y_1) \tag{37}$$

where $\hat{\eta}_I(x_1, y_1)$ is the estimation efficiency in the off-axis IPRG, σ_I is the compressed-noise standard deviation in the off-axis IPRG, and $N_k(x,y)$ is the kth realization of complex-circular Gaussian random numbers with zero mean and unit variance for both the real and imaginary parts (hence the factor of $\sqrt{2}$ in the denominator). It is important to note that by Parseval's theorem (Gaskill, 1978), the total noise power in the windowed Fourier plane is equal to α_I times the original noise power in the image plane, where α_I is the ratio of the pupil area to the Fourier-plane area. Therefore,

$$\sigma_I^2 = \alpha_I \sigma_n^2 = \frac{\pi (d_P/2)^2}{(q_I d_P)^2} \sigma_n^2 = \frac{\pi}{4 q_I^2} \sigma_n^2 \tag{38}$$

which says that the noise variance σ_n^2 (cf. Eq. (24)) is compressed by a factor of α_I when performing spatial heterodyne in the off-axis IPRG.

Signal-to-noise ratio

We can determine the SNR S/N_{IPRG} for the off-axis IPRG as

$$S/N_{\mathrm{IPRG}} = \frac{\mathscr{E}\left\{ |\hat{U}_P(x_1, y_1)|^2 \right\}}{\mathcal{V}\left\{ \hat{U}_P(x_1, y_1) \right\}} \tag{39}$$

where $\mathscr{E}\{\circ\}$ denotes an expected-value operator and $\mathcal{V}\{\circ\}$ denotes a variance operator. Provided a strong reference, $|A_R|^2 \gg |A_S|^2$ and $|U_S(x,y)|^2 \approx |A_S|^2$. As such, we can determine the signal mean number of photoelectrons $\overline{m}_S(\mathrm{n}x_p, \mathrm{m}y_p) \approx \overline{m}_S$ from the following relationship:

$$\overline{m}_S = \frac{\eta\tau}{h\nu} p^2 |A_S|^2 \tag{40}$$

From Eqs. (37) and (38), it then follows that

$$\mathscr{E}\left\{ |\hat{U}_P(x_1, y_1)|^2 \right\} \approx \overline{m}_R \overline{m}_S \tag{41}$$

and

$$\mathcal{V}\left\{ \hat{U}_P(x_1, y_1) \right\} \approx \sigma_I^2 \tag{42}$$

where $\overline{m}_R$ is the reference mean number of photoelectrons (cf. Eq. (21)). Substituting Eqs. (41) and (42) into Eq. (39), we obtain the following closed-form expression (cf. Eq. (24)):

$$S/N_{\text{IPRG}} \approx \frac{4q_I^2}{\pi}\frac{\overline{m}_R\overline{m}_S}{\overline{m}_R+\sigma_r^2} \tag{43}$$

In writing this closed-form expression, we must assume that the estimation efficiency in the off-axis IPRG is equal to the quantum efficiency (i.e., $\hat{\eta}_I(x_1,y_1)=\eta$) for simplicity in the analysis.

Shot-noise limit

Before moving on in the analysis, it is important to note that if $\overline{m}_R \gg \sigma_r^2$, then the SNR S/N_{IPRG} for the off-axis IPRG simplifies (cf. Eq. (43)), such that

$$S/N_{\text{IPRG}} \approx \frac{4q_I^2}{\pi}\overline{m}_S \tag{44}$$

The open literature often refers to this approximation as the shot-noise limit or as obtaining quantum-limited detection. This limit depends on the constraints of the FPA pixel well depth ℓ. In general, $\ell \geq \overline{m}_H(\text{n}x_s, \text{m}y_s)$ (cf. Eq. (15)), so that the FPA pixels do not saturate.

Off-Axis Pupil Plane Recording Geometry

The goal for the following analysis is to model spatial heterodyne in the off-axis PPRG. In turn, we can represent the signal complex field $U_S(x,y)$ that is incident on the FPA as

$$U_S(x,y) = U_P(x_1,y_1) \tag{45}$$

where again, $U_P(x_1,y_1)$ is the pupil-plane complex field (cf. Eq. (10)). Furthermore, if by choice we inject an off-axis LO, then by choice we can represent the reference complex field $U_R(x,y)$ that is incident FPA as

$$U_R(x,y) = U_R(x_1,y_1) = A_R\exp\left(-j2\pi\frac{x_R}{\lambda z_P}x_1\right)\exp\left(-j2\pi\frac{y_R}{\lambda z_P}y_1\right) \tag{46}$$

where here, A_R is a complex constant and (x_R,y_R) are the coordinates of the off-axis LO. It is important to note that the relationship given in Eq. (46) is a tilted plane wave.

Fourier Plane

From Eqs. (1)–(26) and Eqs. (45) and (46), we can gain access to an estimate, $\hat{U}_I(x_1,y_1)$, of the image-plane complex field, $U_I(x_1,y_1)$ (cf. Eq. (10)), by going to the Fourier plane. To go to the Fourier plane, we first let $x=x_1$ and $y=y_1$. Then, we apply a 2-D Fourier transformation, as defined in Eq. (84) in Appendix B, to the relationship contained in Eq. (25), so that

$$\begin{aligned}\mathcal{F}\{\rho_H^+(x_1,y_1)\}_{v_x=\frac{x_2}{\lambda z_P},v_y=\frac{y_2}{\lambda z_P}} &= \tilde{\rho}_H^+\left(\frac{x_2}{\lambda z_P},\frac{y_2}{\lambda z_P}\right) = \left[\frac{\eta\tau}{h\nu}w_x\text{sinc}\left(\frac{w_x x_2}{\lambda z_P}\right)w_y\text{sinc}\left(\frac{w_y y_2}{\lambda z_P}\right)\mathcal{F}\{i_H(x_1,y_1)\}_{v_x=\frac{x_2}{\lambda z_P},v_y=\frac{y_2}{\lambda z_P}}\right.\\ &\left.+\sigma_n\mathcal{F}\{n_4(x_1,y_1)\}_{v_x=\frac{x_2}{\lambda z_P},v_y=\frac{y_2}{\lambda z_P}}\right] ** \frac{1}{\lambda^2 z_P^2}\text{comb}\left(\frac{x_p x_2}{\lambda z_P}\right)\text{comb}\left(\frac{y_p y_2}{\lambda z_P}\right) ** \frac{\text{N}x_p}{\lambda z_P}\text{sinc}\left(\frac{\text{N}x_p x_2}{\lambda z_P}\right)\frac{\text{M}y_p}{\lambda z_P}\text{sinc}\left(\frac{\text{M}y_p y_2}{\lambda z_P}\right)\end{aligned} \tag{47}$$

Taking a look at the remaining 2-D Fourier transformations in Eq. (47), we obtain the following relationship with respect to the hologram irradiance $i_H(x,y)$ (cf. Eqs. (10), (13), and (46)):

$$\begin{aligned}\mathcal{F}\{i_H(x_1,y_1)\}_{v_x=\frac{x_2}{\lambda z_P},v_y=\frac{y_2}{\lambda z_P}} &= \frac{1}{\lambda^2 z_P^2}\tilde{U}_P\left(\frac{x_2}{\lambda z_P},\frac{y_2}{\lambda z_P}\right) ** \tilde{U}_P^*\left(\frac{-x_2}{\lambda z_P},\frac{-y_2}{\lambda z_P}\right) + \lambda^2 z_P^2|A_R|^2\delta(x_2)\delta(y_2) + A_R\tilde{U}_P\left(\frac{x_2-x_R}{\lambda z_P},\frac{y_2-y_R}{\lambda z_P}\right)\\ &+ A_R\tilde{U}_P^*\left(\frac{x_2+x_R}{\lambda z_P},\frac{y_2+y_R}{\lambda z_P}\right)\end{aligned} \tag{48}$$

This relationship contains a 2-D Fourier transformation of the pupil-plane complex field $U_P(x_1,y_1)$ which is proportional to the image-plane complex field $U_I(x_2,y_2)$ (cf. Eq. (12)).

Moving forward let's analyze the various terms contained in the right-hand side of Eq. (48). The first term is a 2-D auto-correlation. This term is centered on axis and is physically twice the circumference of the object diameter d_O. The second term is also centered on axis and contains separable impulse functions (cf. Eq. (80) in Appendix A). These impulse functions are at the strength of the ideal reference with uniform irradiance (i.e., $|A_R|^2$). The last two terms form a complex-conjugate pair and contain a 2-D Fourier transformation of the pupil-plane complex field $U_P(x_1,y_1)$, both scaled in amplitude and shifted off axis.

Substituting Eq. (48) into Eq. (47), we obtain the following relationship:

$$\tilde{\rho}_H^+\left(\frac{x_2}{\lambda z_P},\frac{y_2}{\lambda z_P}\right)=\left\{\frac{\eta\tau}{h\nu}w_x\mathrm{sinc}\left(\frac{w_x x_2}{\lambda z_P}\right)w_y\mathrm{sinc}\left(\frac{w_y y_2}{\lambda z_P}\right)\left[\frac{1}{\lambda^2 z_P^2}\tilde{U}_P\left(\frac{x_2}{\lambda z_P},\frac{y_2}{\lambda z_P}\right)**\tilde{U}_P^*\left(\frac{-x_2}{\lambda z_P},\frac{-y_2}{\lambda z_P}\right)+\lambda^2 z_P^2|A_R|^2\delta(x_2)\delta(y_2)\right.\right.$$
$$\left.\left.+A_R\tilde{U}_P\left(\frac{x_2-x_R}{\lambda z_P},\frac{y_2-y_R}{\lambda z_P}\right)+A_R\tilde{U}_P^*\left(\frac{x_2+x_R}{\lambda z_P},\frac{y_2+y_R}{\lambda z_P}\right)\right]+\sigma_n\mathcal{F}\{n_4(x_1,y_1)\}_{v_x=\frac{x_2}{\lambda z_P},v_y=\frac{y_2}{\lambda z_P}}\right\}$$
$$**\frac{1}{\lambda^2 z_P^2}\mathrm{comb}\left(\frac{x_p x_2}{\lambda z_P}\right)\mathrm{comb}\left(\frac{y_p y_2}{\lambda z_P}\right)**\frac{\mathrm{N}x_p}{\lambda z_P}\mathrm{sinc}\left(\frac{\mathrm{N}x_p x_2}{\lambda z_P}\right)\frac{\mathrm{M}y_p}{\lambda z_P}\mathrm{sinc}\left(\frac{\mathrm{M}y_p y_2}{\lambda z_P}\right) \tag{49}$$

In units of pe, this relationship is physically relevant. Note that the sampling theorem dictates that a sampled function becomes periodic upon finding its spectrum (Gaskill, 1978). In turn, the 2-D convolution with the separable comb functions and the convolution-sifting property of the impulse function causes the terms contained within the squiggly brackets in Eq. (49) to repeat at intervals of $\lambda z_P/x_p$ and $\lambda z_P/y_p$ along the x and y axes, respectively. Also note that the final 2-D convolution with the separable narrow sinc functions serves to smooth out these repeated terms.

Sampling Quotient

To help simplify the analysis to a case that we can easily simulate using N × N computational grids, let's again assume that the FPA has adjacent square pixels, so that $x_p=y_p=w_x=w_y=p$. In so doing, we can rewrite Eq. (49) in terms of the pupil-plane sampling quotient q_P, where

$$q_P=\frac{\lambda z_P}{p d_O} \tag{50}$$

For all intents and purposes, q_P is a measure for the number of FPA pixels across the half width of the speckle-correlation radius (Goodman, 2007; Spencer, 2014). Recall that for monochromatic light, the speckle-correlation radius is proportional to the size of the speckles. In turn, the relationship given in Eq. (50) allows us to vary the sampling with the FPA pixels.

Using Eq. (50), we can rewrite Eq. (49) in terms of the pupil-plane sampling quotient q_P, so that

$$\tilde{\rho}_H^+\left(\frac{x_2}{\lambda z_P},\frac{y_2}{\lambda z_P}\right)=\frac{1}{p^2}\left\{\frac{\eta\tau}{h\nu}p^2\mathrm{sinc}\left(\frac{x_2}{q_P d_O}\right)\mathrm{sinc}\left(\frac{y_2}{q_P d_O}\right)\left[\frac{1}{\lambda^2 z_P^2}\tilde{U}_P\left(\frac{x_2}{\lambda z_P},\frac{y_2}{\lambda z_P}\right)**\tilde{U}_P^*\left(\frac{-x_2}{\lambda z_P},\frac{-y_2}{\lambda z_P}\right)+\lambda^2 z_P^2|A_R|^2\delta(x_2)\delta(y_2)\right.\right.$$
$$\left.\left.+A_R^*\tilde{U}_P\left(\frac{x_2-x_R}{\lambda z_P},\frac{y_2-y_R}{\lambda z_P}\right)+A_R\tilde{U}_P^*\left(\frac{x_2+x_R}{\lambda z_P},\frac{y_2+y_R}{\lambda z_P}\right)\right]+\sigma_n\mathcal{F}\{n_4(x_1,y_1)\}_{v_x=\frac{x_2}{\lambda z_P},v_y=\frac{y_2}{\lambda z_P}}\right\}$$
$$**\frac{1}{q_P^2 d_O^2}\mathrm{comb}\left(\frac{x_2}{q_P d_O}\right)\mathrm{comb}\left(\frac{y_2}{q_P d_O}\right)**\frac{\mathrm{N}^2}{q_P^2 d_O^2}\mathrm{sinc}\left(\frac{\mathrm{N}x_2}{q_P d_O}\right)\mathrm{sinc}\left(\frac{\mathrm{N}y_2}{q_P d_O}\right) \tag{51}$$

Here, $q_P d_O=\lambda z_P/p$ is the side length of the N × N computational grid in the Fourier plane. Note that as $N\to\infty$ (cf. Eq. (80) in the Appendix), we can make use of the convolution-sifting property of the impulse function and neglect the final 2-D convolution in Eq. (51). Accordingly, for large N the smoothing caused by the final 2-D convolution in Eq. (51) becomes minimized; however, for small N the smoothing becomes more pronounced. Let's assume that $x_R=y_R=q_P d_O/4$, so that the last two terms within the square brackets in Eq. (51) shift diagonally. When $q_p\geq 4$, the last two terms no longer overlap with the first two terms which are centered on axis. Correspondingly, when $2\leq q_p<4$, the last two terms are still resolvable within the side length of the N × N computational grid but overlap with the first term. This latter case allows for us to obtain more samples across the object diameter d_O which in turn minimizes the smoothing caused by the final 2-D convolution in Eq. (51) but increases the noise sampling. Also note that this functional overlap becomes negligible when the amplitude of the reference $|A_R|$ is dominate (in comparison to the other amplitude terms).

Estimate

Provided Eqs. (50) and (51), we must shift the Fourier-plane data and apply a window function to obtain an estimate, $\hat{U}_I(x_2,y_2)$, of the image-plane complex field, $U_I(x_2,y_2)$. In particular,

$$\hat{U}_I(x_2,y_2)=w(x_2,y_2)\tilde{\rho}_H^+\left(\frac{x_2+x_R}{\lambda z_P},\frac{y_2+y_R}{\lambda z_P}\right) \tag{52}$$

where $w(x_2,y_2)$ is the window function. In using Eq. (52), we must satisfy Nyquist sampling with the FPA pixels (Gaskill, 1978), so that the repeated terms within Eq. (51) do not overlap and cause significant aliasing. Thus, the Nyquist rate is $q_P d_O=\lambda z_P/p$ and the Nyquist interval is $1/(q_P d_O)=p/(\lambda z_P)$ when $x_R=y_R=q_P d_O/4$.

Assuming that we satisfy Nyquist sampling, let's assume that $q_P\geq 2$, $|A_R|$ is dominant, and

$$w(x_2,y_2)=\mathrm{cyl}\left(\frac{\sqrt{x_2^2+y_2^2}}{d_O}\right) \tag{53}$$

In turn, Eq. (52) simplifies, such that

$$\hat{U}_I(x_2,y_2)\approx\frac{1}{p^2}\left[\frac{\eta\tau}{h\nu}p^2\mathrm{sinc}\left(\frac{x_2+x_R}{q_Pd_O}\right)\mathrm{sinc}\left(\frac{y_2+y_R}{q_Pd_O}\right)A_R^*\tilde{U}_P\left(\frac{x_2}{\lambda z_P},\frac{y_2}{\lambda z_P}\right)+\sigma_n w(x_1,y_1)\mathcal{F}\{n_4(x_1,y_1)\}_{\nu_x=\frac{x_2}{\lambda z_P},\nu_y=\frac{y_2}{\lambda z_P}}\right]**\delta(x_2)\delta(y_2)$$
$$**\frac{\mathrm{N}^2}{q_P^2d_O^2}\mathrm{sinc}\left(\frac{\mathrm{N}x_2}{q_Pd_O}\right)\mathrm{sinc}\left(\frac{\mathrm{N}y_2}{q_Pd_O}\right) \tag{54}$$

The reader should note that as $\mathrm{N}\to\infty$ (cf. Eq. (80) in Appendix A), we can again make use of the convolution-sifting property of the impulse function and neglect the final 2-D convolutions in Eq. (54). Accordingly, for large N the smoothing caused by the final 2-D convolution in Eq. (54) becomes minimized (and the noise remains delta correlated); however, for small N the smoothing becomes more pronounced (and the noise is no longer delta correlated).

Assuming that N is large, the estimate $\hat{U}_I(x_1,y_1)$ simplifies (cf. Eq. (54)), such that

$$\hat{U}_I(x_2,y_2)\approx\frac{\hat{\eta}_P(x_2,y_2)\tau}{h\nu}A_R^*\tilde{U}_P\left(\frac{x_2}{\lambda z_P},\frac{y_2}{\lambda z_P}\right)+\frac{\sigma_P}{\sqrt{2}}N_1(x_2,y_2) \tag{55}$$

where here, $\hat{\eta}_P(x_1,y_1)$ is the estimation efficiency in the off-axis PPRG, σ_P is the compressed-noise standard deviation in the off-axis PPRG, and $N_k(x,y)$ is again the kth realization of complex-circular Gaussian random numbers with zero mean and unit variance for both the real and imaginary parts (hence the factor of $\sqrt{2}$ in the denominator). It is important to note that by Parseval's theorem (Gaskill, 1978), the total noise power in the windowed Fourier plane is equal to α_P times the original noise power in the pupil plane, where α_P is the ratio of the image area to the Fourier-plane area. Therefore,

$$\sigma_P^2=\alpha_P\sigma_n^2=\frac{\pi(d_O/2)^2}{(q_Pd_O)^2}\sigma_n^2=\frac{\pi}{4q_P^2}\sigma_n^2 \tag{56}$$

which says that the noise variance σ_n^2 (cf. Eq. (24)) is compressed by a factor of α_P when performing spatial heterodyne in the off-axis PPRG.

Signal-to-noise ratio

We can determine the SNR S/N_{PPRG} for the off-axis PPRG as

$$S/N_{\mathrm{PPRG}}=\frac{\mathscr{E}\{|\hat{U}_I(x_2,y_2)|^2\}}{\mathcal{V}\{\hat{U}_I(x_2,y_2)\}} \tag{57}$$

where again, $\mathscr{E}\{\circ\}$ denotes an expected-value operator and $\mathcal{V}\{\circ\}$ denotes a variance operator. From Eqs. (55) and (56),

$$\mathscr{E}\{|\hat{U}_I(x_2,y_2)|^2\}\approx\overline{m}_R\overline{m}_S \tag{58}$$

and

$$\mathcal{V}\{\hat{U}_I(x_2,y_2)\}\approx\sigma_P^2 \tag{59}$$

Here, $\overline{m}_R$ and $\overline{m}_S$ are the reference and signal mean number of photoelectrons, respectively, assuming a strong reference (cf. Eqs. (21) and (40)). Substituting Eqs. (58) and (59) into Eq. (57), we obtain the following closed-form expression (cf. Eq. (24)):

$$S/N_{\mathrm{PPRG}}\approx\frac{4q_P^2}{\pi}\frac{\overline{m}_R\overline{m}_S}{\overline{m}_R+\sigma_r^2} \tag{60}$$

In writing this closed-form expression, we must assume that the estimation efficiency in the off-axis PPRG is equal to the quantum efficiency (i.e., $\hat{\eta}_P(x_1,y_1)=\eta$) for simplicity in the analysis.

Shot-noise limit

Before moving on in the analysis, it is important to note that if $\overline{m}_R\gg\sigma_r^2$, then the SNR S/N_{PPRG} for the off-axis PPRG simplifies (cf. Eq. (60)), such that

$$S/N_{\mathrm{PPRG}}\approx\frac{4q_P^2}{\pi}\overline{m}_S \tag{61}$$

Again, the open literature often refers to this approximation as the shot-noise limit or as obtaining quantum-limited detection. This limit depends on the constraints of the FPA pixel well depth ℓ. In general, $\ell\geq\overline{m}_H(\mathrm{n}x_s,\mathrm{m}y_s)$ (cf. Eq. (15)), so that the FPA pixels do not saturate.

On-Axis Phase Shifting Recording Geometry

The goal for the following analysis is to model spatial heterodyne in the on-axis PSRG. For this purpose, we can represent the signal complex field $U_S(x,y)$ that is incident on the FPA as

$$U_S(x,y)=U_P(x_1,y_1) \tag{62}$$

or

$$U_S(x,y) = U_I(x_2,y_2) \tag{63}$$

where again, $U_P(x_1,y_1)$ is the pupil-plane complex field (cf. Eq. (10)) and $U_I(x_2,y_2)$ is the image-plane complex field (cf. Eq. (12)). Moreover, if by choice we inject an on-axis LO, then by choice we can represent the reference complex field $U_R(x,y)$ that is incident FPA as

$$U_R^{(\delta)}(x,y) = U_R^{(\delta)}(x_1,y_1) = U_R^{(\delta)}(x_2,y_2) = A_R e^{-j\delta} \tag{64}$$

where here, A_R is a complex constant and δ is a real constant. It is important to note that the relationship given in Eq. (64) is an ideal reference with a uniform irradiance and a piston-phase shift (i.e., $|A_R|^2$ and δ, respectively).

Estimate

Provided Eqs. (62)–(64), Eq. (13) becomes

$$I_H^{(\delta)}(x,y) = |U_S(x,y)|^2 + |A_R|^2 + U_S(x,y)A_R^* e^{j\delta} + A_R e^{-j\delta} U_S^*(x,y) \tag{65}$$

This relationship says that via (simultaneous or sequential) piston-phase shifts, we can acquire four hologram-irradiance measurements with the FPA (Poon and Liu, 2014) to obtain an estimate, $\hat{U}_P(x_1,y_1)$ or $\hat{U}_I(x_2,y_2)$, of the desired complex field, $U_P(x_1,y_1)$ or $U_I(x_2,y_2)$. For example, if $\delta = 0$, $\pi/2$, π, and $3\pi/2$, then

$$\begin{aligned}
I_H^{(0)}(x,y) &= |U_S(x,y)|^2 + |A_R|^2 + U_S(x,y)A_R^* + A_R U_S^*(x,y) \\
I_H^{(\pi/2)}(x,y) &= |U_S(x,y)|^2 + |A_R|^2 + jU_S(x,y)A_R^* - jA_R U_S^*(x,y) \\
I_H^{(\pi)}(x,y) &= |U_S(x,y)|^2 + |A_R|^2 - U_S(x,y)A_R^* - A_R U_S^*(x,y) \\
I_H^{(3\pi/2)}(x,y) &= |U_S(x,y)|^2 + |A_R|^2 - jU_S(x,y)A_R^* + jA_R U_S^*(x,y)
\end{aligned} \tag{66}$$

To remove the common terms (i.e., $|U_S(x,y)|^2 + |A_R|^2$), we can perform the following subtractions:

$$\begin{aligned}
I_H^{(0)}(x,y) - I_H^{(\pi)}(x,y) &= 2U_S(x,y)A_R^* + 2A_R U_S^*(x,y) \\
I_H^{(\pi/2)}(x,y) - I_H^{(3\pi/2)}(x,y) &= j2U_S(x,y)A_R^* - j2A_R U_S^*(x,y)
\end{aligned} \tag{67}$$

so that

$$\left[I_H^{(0)}(x,y) - I_H^{(\pi)}(x,y)\right] + j\left[I_H^{(3\pi/2)}(x,y) - I_H^{(\pi/2)}(x,y)\right] = 4A_R^* U_S(x,y) \tag{68}$$

From Eqs. (1)–(26) and Eq. (68), our estimate, in units of pe, becomes

$$\begin{aligned}
\hat{U}_S(x,y) &\approx \frac{4}{S}\frac{\eta\tau}{h\nu} A_R^* U_S(x,y) ** \mathrm{rect}\left(\frac{x}{p}\right)\mathrm{rect}\left(\frac{y}{p}\right) + 2\sigma_n N_2(x,y) \\
&\approx \frac{4}{S}\frac{\hat{\eta}(x,y)\tau}{h\nu} p^2 A_R^* U_S(x,y) + 2\sigma_n N_2(x,y)
\end{aligned} \tag{69}$$

with the assumption that the FPA has adjacent square pixels, so that $x_p = y_p = w_x = w_y = p$. In Eq. (69), S is the total number of amplitude splits required to make the four hologram-irradiance measurements with the FPA (Rhoadarmer and Barchers, 2002). It is important to note that since we have four hologram-irradiance measurements with the FPA, the noise variance σ_n^2 (cf. Eq. (24)) is multiplied by a factor of 4 when performing spatial heterodyne in the on-axis PSRG.

Signal-to-noise ratio

We can determine the SNR S/N_{PSRG} for the on-axis PSRG as

$$S/N_{\mathrm{PSRG}} = \frac{\mathscr{E}\{|\hat{U}_S(x,y)|^2\}}{\mathscr{V}\{\hat{U}_S(x,y)\}} \tag{70}$$

where again, $\mathscr{E}\{\circ\}$ denotes an expected-value operator and $\mathscr{V}\{\circ\}$ denotes a variance operator. From Eq. (69),

$$\mathscr{E}\{|\hat{U}_I(x_1,y_1)|^2\} \approx \frac{16}{S^2}\overline{m}_R\overline{m}_S \tag{71}$$

and

$$\mathscr{V}\{\hat{U}_I(x_1,y_1)\} \approx 4\sigma_n^2 \tag{72}$$

where here, $\overline{m}_R$ and $\overline{m}_S$ are the reference and signal mean number of photoelectrons, respectively, assuming a strong reference (cf. Eqs. (21) and (40)). Substituting Eqs. (71) and (72) into Eq. (70), we obtain the following closed-form expression (cf. Eq. (24)):

$$S/N_{\mathrm{PSRG}} \approx \frac{4}{S^2}\frac{\overline{m}_R\overline{m}_S}{\overline{m}_R + \sigma_r^2} \tag{73}$$

In writing this closed-form expression, we must assume that the estimation efficiency in the on-axis PSRG is equal to the quantum efficiency (i.e., $\hat{\eta}(x,y) = \eta$) for simplicity in the analysis.

Shot-noise limit

Before moving on in the analysis, it is important to note that if $\overline{m}_R \gg \sigma_r^2$, then the SNR S/N_{PSRG} for the on-axis PSRG simplifies (cf. Eq. (73)), such that

$$S/N_{\text{PSRG}} \approx \frac{4}{S^2}\overline{m}_S \tag{74}$$

Again, the open literature often refers to this approximation as the shot-noise limit or obtaining quantum-limited detection. This limit depends on the constraints of the FPA pixel well depth ℓ. In general, $\ell \geq \overline{m}_H(\mathrm{n}x_s, \mathrm{m}y_s)$ (cf. Eq. (15)), so that the FPA pixels do not saturate.

Comparison of the Different Recording Geometries

In this section, we will explore the differences between the various recording geometries (studied in the previous sections) from a visual standpoint. For this purpose, let's assume that we have a point-source object (cf. Eq. (1)), so that the laser-object interaction gives rise to an ideal spherical wave upon propagation. Let's also assume that we have isoplanatic phase aberrations present (i.e., those that exist (or approximately exist) within the pupil plane of our imaging system (cf. **Fig. 2**)). As such (cf. Eqs. (8) and (10)),

$$\frac{U_P(x_1, y_1)}{A_s} = P(x_1, y_1) \tag{75}$$

and we can neglect the effects of speckle due to an optically rough object. Before moving on in the analysis, it is important to note that example MATLAB® code for the off-axis IPRG, the off-axis PPRG, and the on-axis PSRG is included in Appendix C. The readers interested in extended objects (in addition to point-source objects) will be able to visualize the associated physics (cf. **Fig. 1**) by executing these provided scripts.

To quantify performance, we will use the field-estimated Strehl ratio S_F, such that

$$S_F = \frac{|\mathscr{E}\left\{U_S(x,y)\hat{U}_S^*(x,y)\right\}|^2}{\mathscr{E}\{|U_S(x,y)|^2\}\mathscr{E}\{|\hat{U}_S(x,y)|^2\}} \tag{76}$$

where $U_S(x,y)$ and $\hat{U}_S(x,y)$ are the "truth" and "estimated" signal complex fields, respectively, and again, $\mathscr{E}\{\circ\}$ is the expected-value operator. This performance metric bears some resemblance to a Strehl ratio, which in practice provides a normalized measure for performance (Spencer *et al.*, 2016). In Eq. (76), if $U_S(x,y) = \hat{U}_S(x,y)$, then $S_F = 1$. Else if $U_S(x,y) \neq \hat{U}_S(x,y)$, then $S_F < 1$. Thus, Eq. (76) is in line with the general understanding of a Strehl ratio, and provides a normalized measure for field-estimated performance. The reader should note that Eq. (76) ultimately stems from the following definition of the Strehl ratio:

$$S = \frac{|\mathscr{E}\{U_P(x_1, y_1)\}|^2}{\mathscr{E}\{|U_P(x_1, y_1)|^2\}} \tag{77}$$

Here, we have made use of the fact that the expected value of a pupil-plane field quantity is equivalent to the on-axis DC term of the 2D Fourier transformation of that pupil-plane field quantity.

Off-Axis Image Plane Recording Geometry

Fig. 4 shows the results for the off-axis IPRG. In the left-hand column, $q_I = 2$, whereas in the right-hand column, $q_I = 4$ (cf. Eq. (32)). Note that the computational grids were setup so that the image-plane grid length was set equal to the pupil-plane grid length (which was also set equal to the exit-pupil diameter $d = 0.3$ m) using 256×256 grid points. This setup required that $z_I = q_I d^2/(\mathrm{N}\lambda)$ when solving the Fresnel integral numerically (Spencer *et al.*, 2016). Also note that

- **Fig. 4(a)** and **(b)** show the 1 RMS isoplanatic phase aberrations that existed in the simulated pupil plane;
- **Fig. 4(c)** and **(d)** show the zero-mean digital holograms recorded with the off-axis IPRG;
- **Fig. 4(e)** and **(f)** show the amplitudes associated with the simulated Fourier plane; and
- **Fig. 4(g)** and **(h)** show the wrapped phases associated with the estimated complex field.

Here, $S_F = 0.943$ and $S_F = 0.984$ for the left-hand and right-hand columns of **Fig. 4**, respectively. To calculate the field-estimated Strehl ratio S_F (cf. Eq. (76)), the simulated Fourier plane was zero padded to the appropriate size, so that the windowed Fourier plane (i.e., the estimated complex field) contained 256×256 grid points. This zero padding ensured that the estimated complex field had the same number of grid points as the simulated pupil plane. Given that $\overline{m}_S = 4$ pe, $\overline{m}_R = 0.75\ell$, and $\ell = 100 \times 10^3$ pe, the following SNRs were also calculated for the left-hand and right-hand columns of **Fig. 4** (cf. Eqs. (43) and (44)): $S/N = 9$ and $S/N = 18$.

Off-Axis Pupil Plane Recording Geometry

Fig. 5 shows the results for the off-axis PPRG. Here, $q_P = 2$ in the left-hand column and $q_P = 4$ in the right-hand column (cf. Eq. (50)). The computational grids were setup with 256×256 grid points and the pupil-plane grid length was set equal to the exit-pupil diameter $d = 0.3$ m. It is important to note that

- **Fig. 5(a)** and **(b)** show the 1 RMS isoplanatic phase aberrations that existed in the simulated pupil plane;
- **Fig. 5(c)** and **(d)** show the zero-mean digital holograms recorded with the off-axis PPRG;

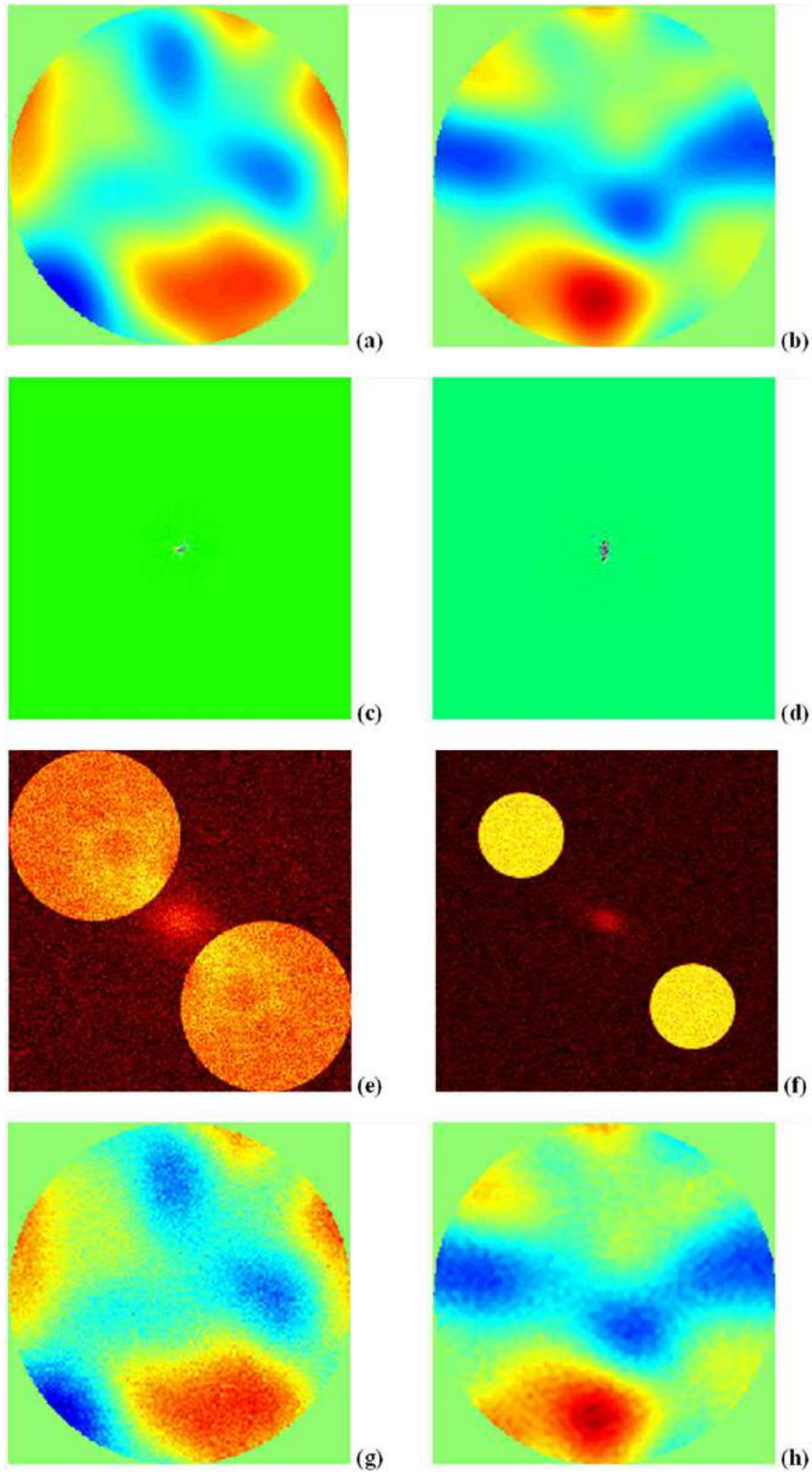


Fig. 4 Results for the off-axis image plane recording geometry or IPRG.

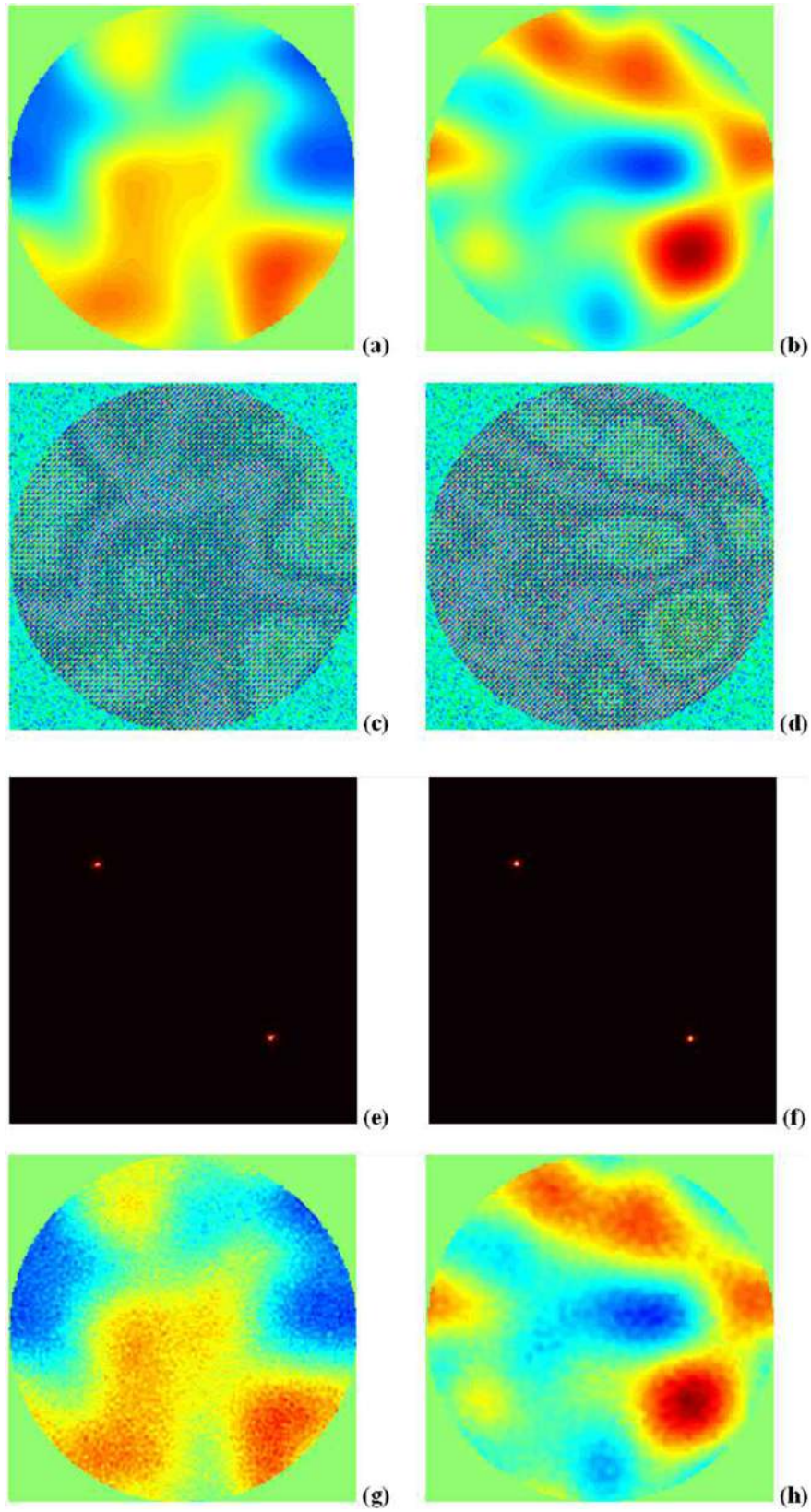


Fig. 5 Results for the off-axis pupil plane recording geometry or PPRG.

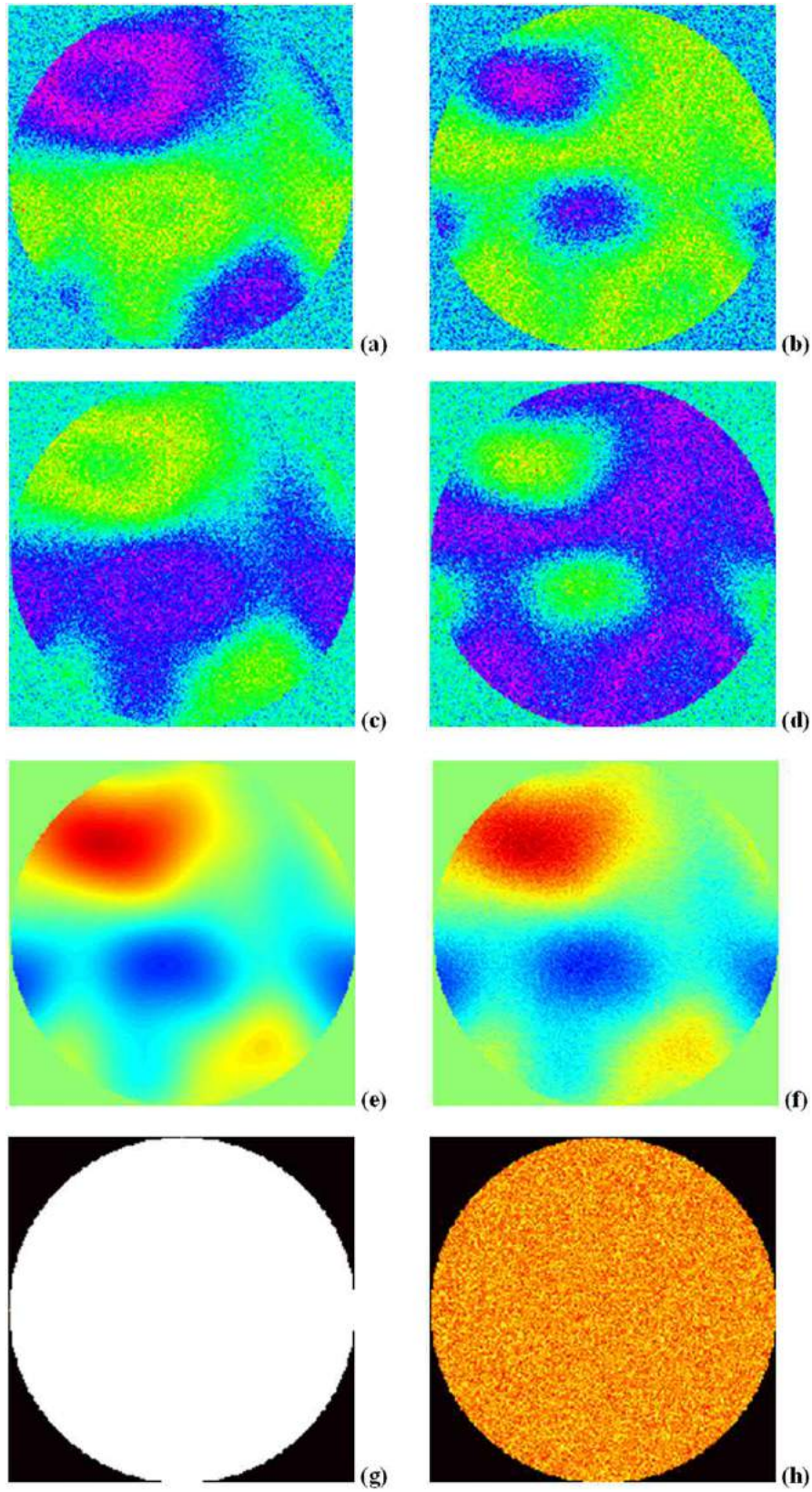


Fig. 6 Results for the on-axis phase shifting recording geometry or PSRG.

- Fig. 5(e) and (f) show the amplitudes associated with the simulated Fourier plane; and
- Fig. 5(g) and (h) show the wrapped phases associated with the estimated complex field.

To calculate the estimated complex field, the windowed Fourier plane was zero padded to 256×256 grid points and was numerically inverse Fourier transformed. This zero padding ensured that the estimated complex field had the same number of grid points as the simulated pupil plane for the purposes of computing field-estimated Strehl ratio S_F (cf. Eq. (76)), where here, $S_F = 0.955$ and $S_F = 0.986$ for the left-hand and right-hand columns of Fig. 5, respectively. Given that $\overline{m}_S = 4$ pe, $\overline{m}_R = 0.75\ell$, and $\ell = 100 \times 10^3$ pe, the following SNRs were also calculated for the left-hand and right-hand columns of Fig. 5 (cf. Eqs. (60) and (61)): $S/N = 9$ and $S/N = 18$.

On-Axis Phase Shifting Recording Geometry

Fig. 6 shows the results for the on-axis PSRG. Once again, the computational grids were setup with 256×256 grid points and the pupil-plane grid length was set equal to the exit-pupil diameter $d = 0.3$ m. Fig. 6(a)–(d) show the four phase-shifted digital holograms, where $\delta = 0$, $\pi/2$, π, and $3\pi/2$, respectively. Here, $S = 1$, so that $\overline{m}_S = 4/S$ pe, $\overline{m}_R = 0.75\ell$, and $\ell = 100 \times 10^3$ pe. Given these parameters, the following SNR was calculated for the results shown in Fig. 6 (cf. Eqs. (60) and (61)): $S/N = 14$. It is important to note that Fig. 6(e) and (g) show the wrapped phase and amplitude for the truth complex field, whereas Fig. 6(f) and (h) show the wrapped phase and amplitude for the estimated complex field. As such, the field-estimated Strehl ratio S_F (cf. Eq. (76)), was calculated as $S_F = 0.947$.

Conclusion

Spatial heterodyne, in general, offers a distinct way forward to combat low SNRs that often occur when performing optics and photonics applications. In using spatial heterodyne, we can set the strength of the reference so that it boosts the signal above the read-noise floor of the FPA. As such, we can approach a shot-noise limited detection regime. This last statement is of course dependent on the parameters of the FPA, such as the pixel well depth. With that said, this encyclopedia article provides the toolset needed for future research efforts to use spatial heterodyne in their optics and photonics applications.

Appendix A

This appendix defines the rectangle, comb, impulse, cylinder, and sinc functions as

$$\mathrm{rect}(x) = \begin{cases} 0, & |x| > 0.5 \\ 0.5, & x = 0.5 \\ 1, & |x| < 0.5 \end{cases} \tag{78}$$

$$\frac{1}{|w|}\mathrm{comb}\left(\frac{x}{w}\right) = \sum_{n=-\infty}^{\infty} \delta(x - nw) \tag{79}$$

$$\delta(x) = \lim_{w \to 0} \frac{1}{|w|} p\left(\frac{x}{w}\right) \tag{80}$$

(where $p(x)$ is a pulse-like function (e.g., the rectangle function)),

$$\mathrm{cyl}(\rho) = \begin{cases} 1, & 0 \le \rho < 0.5 \\ 0.5, & \rho = 0.5 \\ 0, & \rho > 0.5 \end{cases} \tag{81}$$

(where $\rho = \sqrt{x^2 + y^2}$), and

$$\mathrm{sinc}(x) = \frac{\sin(\pi x)}{\pi x} \tag{82}$$

respectively.

Appendix B

This appendix defines 2-D convolution as

$$V(x,y) * *W(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} V(x',y')W(x-x',y-y')dx'dy' \tag{83}$$

(where x' and y' are dummy variables of integration), a 2-D Fourier transform as

$$\mathcal{F}\{V(x,y)\}_{v_x,v_y} = \tilde{V}(v_x,v_y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} V(x,y)e^{-j2\pi(xv_x+yv_y)}dxdy \tag{84}$$

and a 2-D inverse Fourier transformation as

$$\mathcal{F}^{-1}\left\{\tilde{V}(v_x,v_y)\right\}_{x,y} = V(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \tilde{V}(v_x,v_y)e^{j2\pi(xv_x+yv_y)}dv_xdv_y \tag{85}$$

Appendix C

This appendix provides the MATLAB® code needed to explore the off-axis IPRG, the off-axis PPRG, and the on-axis PSRG.

Listing 1. MATLAB® code for the cylinder function.

```
function z = cyl(x,y)
% function z = cyl(x,y)
r = sqrt(x.^2+y.^2);
z = double(r < 0.5);
z(r == 0.5) = 0.5;
return
```

Listing 2. MATLAB® code for the off-axis image plane recording geometry or IPRG.

```
% script_IPRG_final
clear all
close all
clc
%
N = 256;                              % number of grid points
d = 0.3;                              % exit-pupil diamter [m]
delta = d/N;                          % grid spacing [m]
wvl = 1e-6;                           % wavelength [m]
k = 2*pi/wvl;                         % angular wavenumber [rad/m]
q_I = 2.0;                            % IPRG quotient
ifov = wvl/d/q_I;                     % pixel field of view
z_I = d/(N*ifov);                     % propagation length [m] (FPA side = d)
sigma_r = 100;                        % read-noise standard deviation [pe]
l = 100e3;                            % FPA well depth [pe]
m_R = 0.75*l;                         % reference mean number of pe's [pe]
m_S = 4;                              % signal mean number of pe's [pe]
%
% Setup the pupil plane
%
[x1,y1] = meshgrid((-N/2 : N/2-1)*delta);
r1 = hypot(x1,y1);
%
phi = randn(N);                       % phase aberration [rad]
a = N/10*delta;                       % 1/e width [m]
g = exp(-r1.^2/(2*a^2));              % Gaussian function
phi = ...                             % smoothed phase aberrtaion [rad]
    real(ifft2(fft2(phi).*ifft2(ifftshift(g))));
```

```
phi = ...                               % rms = 1 [rad]
    (phi-mean(phi(:)))/std(phi(:));
P = ...                                 % gen. complex pupil function
    cyl(x1/d,y1/d).*exp(1i*phi);
%
% Let's assume a point-source object
%
U_P = P;                                % pup.-plane complex field [sqrt(W)/m]
%
figure(1); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_P)); axis image;  axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_P),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
U_Pp = ...                              % leaving complex field [sqrt(W)/m]
    exp(-1j*k/(2*z_I)*(x1.^2 + y1.^2)).*U_P;
%
figure(2); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_Pp)); axis image;  axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_Pp),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
% Propagate from the pupil plane to the image plane
%
[x2,y2] = meshgrid((-N/2:N/2-1)*delta);
%
M = 2*N+rem(N,2);                       % double the number of grid points
U_Ppg = zeros(M);                       % pad with a guard band of 2
nn = round(N/2)+(1:N);                  % original coordinates
U_Ppg(nn,nn) = U_Pp;                    % embed in padded grid
% transform the embedded field
U_Ppgt = delta^2*fftshift(fft2(fftshift(U_Ppg)));
% create the propagation kernel
const = k/(2*z_I)*delta^2;
kernel = exp(1j*const*(ceil(-M/2):floor((M-1)/2))'.^2);
% transform the kernel
kernelt = delta*fftshift(fft(fftshift(kernel)));
% multiply and transform back with appropriate scale factors
U_Ppgt = (kernelt*kernelt.').*U_Ppgt;
U_Ip = 1/(1j*wvl*z_I)*exp(1j*k*z_I) ...
    *(1/delta)^2*ifftshift(ifft2(ifftshift(U_Ppgt)));
% use original coordinates
U_I = U_Ip(nn,nn);                      % img.-plane complex field [sqrt(W)/m]
U_S = ...                               % signal complex field [sqrt(pe)]
    sqrt(m_S)*U_I/sqrt(mean(abs(U_I(:)).^2));
display(['Sig. mean number of photoectrons: ' ...
    num2str(mean(abs(U_S(:)).^2)) ' [pe] (point-source object)']);
%
figure(3); clf;
ax1=subplot(1,2,1);
imagesc(abs(U_S)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_S),[-pi,pi]); axis image; axis off,colormap(ax2,jet);
%
x_R = d*q_I/4;                          % x-axis reference origin
y_R = d*q_I/4;                          % y-axis reference origin
U_R = ...                               % reference complex field [sqrt(W)/m]
    1/1j*exp(1j*k*z_I)*exp(1j*k*(x2.^2+y2.^2)/(2*z_I)) ...
```

```
    .*exp(1i*2*pi*y_R*(y2/(wvl*z_I))).*exp(1i*2*pi*x_R*(x2/(wvl*z_I)));
U_R = ...                           % reference complex field [sqrt(pe)]
    sqrt(m_R)*U_R/sqrt(mean(abs(U_R(:)).^2));
display(['Ref. mean number of photoectrons: ' ...
    num2str(mean(abs(U_R(:)).^2)) ' [pe]']);
%
figure(4); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_R)); axis image; axis off; colormap(ax1,hot);
ax2 = subplot(1,2,2);
imagesc(angle(U_R),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
% Measure the image plane on a FPA
%
m_H = abs(U_S + U_R).^2;            % holgram mean number of pe's [pe]
display(['Hol. mean number of photoectrons: ' ...
    num2str(mean(m_H(:))) ' [pe] (point-source object)']);
m_Hn = ...                          % add noise
    poissrnd(m_H)+sigma_r*randn(size(m_H));
% use the follwoing if you do not have the Statistics Toolbox
% m_Hn = ...                          % add noise
%    m_H+sqrt(m_S+m_R)*randn(size(m_H))+sigma_r*randn(size(m_H));
display(['Hol. mean number of photoectrons: ' ...
    num2str(mean(m_Hn(:))) ' [pe] (point-source object with noise)']);
%
m_H = m_H - mean(m_H(:));           % remove DC
m_Hn = m_Hn - mean(m_Hn(:));        % remove DC
%
figure(5); clf;
ax1 = subplot(1,2,1);
imagesc(m_H); axis image; axis off; colormap(ax1,jet);
ax2 = subplot(1,2,2);
imagesc(m_Hn); axis image; axis off; colormap(ax2,jet);
%
M = round(q_I*N);                   % desired number of grid points
m_Hng = zeros(M);                   % pad with a gaurdband
nn = round(M/2-N/2)+(1:N);          % original coordinates
m_Hng(nn,nn) = m_Hn;                % embed in padded grid
%
% Go to the Fourier plane to obtain an estimate
%
m_Hngt = ...                        % Fourier plane
    (1/delta)^2*ifftshift(ifft2(ifftshift(m_Hng)));
%
figure(6); clf;
ax1 = subplot(1,2,1);
imagesc(abs(m_Hngt)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(m_Hngt),[-pi,pi]); axis image; axis off; colormap(ax2,hsv);
%
nn = round(3*M/4-N/2)+(1:N);        % estimate coordinates
U_Pe = ...                          % window Fourier plane
    cyl(x1/d,y1/d).*m_Hngt(nn,nn);
%
figure(7); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_Pe)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_Pe),[-pi,pi]); axis image; axis off; colormap(ax2,hsv);
%
S_F = ...                           % calculate field-estimated Strehl
```

```
    abs(mean(U_P(:).*conj(U_Pe(:)))).^2 ...
    /(mean(abs(U_P(:)).^2).*mean(abs(U_Pe(:)).^2));
display(['Field-estimated Strehl Ratio    : ' num2str(S_F)]);
%
% Setup the object plane
%
[x0,y0] = meshgrid((-N/2 : N/2-1)*delta);
%
% Let's assume an extended object
%
dO = d;                              % diameter of the object [m]
O = ...                              % gen. complex object function
    cyl(x0/dO,y0/dO) ...
    - cyl(x0/(3*dO/4),y0/(3*dO/4)) ...
    - cyl(x0/(dO/2),y0/(dO/2)) ...
    + cyl(x0/(dO/4),y0/(dO/4));
phz = -pi + 2*pi*rand(N);            % unif. dist. random phase (-pi,pi]
R_O = O.*exp(1j*phz);                % obj.-plane complex reflectance.
%
U_I = ...                            % img.-plane complex field [sqrt(W)/m]
    ifftshift(ifft2(ifftshift( fftshift(fft2(fftshift(R_O))) ...
    .* fftshift(fft2(fftshift((U_I)))) )));
U_S = ...                            % signal complex field [sqrt(pe)]
    sqrt(m_S)*U_I/sqrt(mean(abs(U_I(:)).^2));
display(['Sig. mean number of photoectrons: ' ...
    num2str(mean(abs(U_S(:)).^2)) ' [pe] (extended object)']);
%
figure(8); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_S)); axis image; axis off; colormap(ax1,hot);
ax2 = subplot(1,2,2);
imagesc(angle(U_S),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
x_R = d*q_I/4;                       % x-axis reference origin
y_R = d*q_I/4;                       % y-axis reference origin
U_R = ...                            % reference complex field [sqrt(W)/m]
    1/1j*exp(1j*k*z_I)*exp(1j*k*(x2.^2+y2.^2)/(2*z_I)) ...
    .*exp(1i*2*pi*y_R*(y2/(wvl*z_I))).*exp(1i*2*pi*x_R*(x2/(wvl*z_I)));
U_R = ...                            % reference complex field [sqrt(pe)]
    sqrt(m_R)*U_R/sqrt(mean(abs(U_R(:)).^2));
display(['Ref. mean number of photoectrons: ' ...
    num2str(mean(abs(U_R(:)).^2)) ' [pe]']);
%
figure(9); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_R)); axis image; axis off; colormap(ax1,hot);
ax2 = subplot(1,2,2);
imagesc(angle(U_R),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
% Measure the image plane on a FPA
%
m_H = abs(U_S + U_R).^2;             % holgram mean number of pe's [pe]
display(['Hol. mean number of photoectrons: ' ...
    num2str(mean(m_H(:))) ' [pe] (extended object)']);
m_Hn = ...                           % add noise
    poissrnd(m_H)+sigma_r*randn(size(m_H));
% use the follwoing if you do not have the Statistics Toolbox
% m_Hn = ...                           % add noise
%     m_H+sqrt(m_S+m_R)*randn(size(m_H))+sigma_r*randn(size(m_H));
display(['Hol. mean number of photoectrons: ' ...
```

```
    num2str(mean(m_Hn(:))) ' [pe] (extended object with noise)']);
%
m_H = m_H - mean(m_H(:));           % remove DC
m_Hn = m_Hn - mean(m_Hn(:));        % remove DC
%
figure(10); clf;
ax1 = subplot(1,2,1);
imagesc(m_H); axis image; axis off; colormap(ax1,hsv);
ax2 = subplot(1,2,2);
imagesc(m_Hn); axis image; axis off; colormap(ax2,hsv);
%
% Go to the Fourier plane to obtain an estimate
%
m_Hnt = ...                         % Fourier plane
    (1/delta)^2*ifftshift(ifft2(ifftshift(m_Hn)));
%
figure(11); clf;
ax1 = subplot(1,2,1);
imagesc(abs(m_Hnt)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(m_Hnt),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
nn = ...                            % estimate coordinates
    round(N/4-round(N/q_I)/2)+(1:round(N/q_I));
M = length(nn);                     % estimate number of grid points
[xf,yf] = meshgrid((-M/2 : M/2-1)*d*q_I/N);
U_Pe = ...                          % window Fourier plane
    cyl(xf/d,yf/d).*m_Hnt(nn,nn);
%
figure(12); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_Pe)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_Pe),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
nn = round(N/2-M/2)+(1:M);          % desired coordinates
U_Peg = zeros(N);                   % pad with a gaurdband
U_Peg(nn,nn) = U_Pe;                % embed in padded grid
U_Ie = ...
    (d*q_I/N)^2 * fftshift(fft2(fftshift(U_Peg)));
%
figure(13),clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_Ie)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(abs(R_O)); axis image; axis off; colormap(ax2,hot);
%
SNR = ...                           % calculate SNR
    4*q_I*m_R*m_S/(pi*(m_R + sigma_r^2));
display(['PPRG signal-to-noise ratio      : ' num2str(SNR)]);
```

Listing 3. MATLAB® code for the off-axis pupil plane recording geometry or PPRG.

```
% script_PPRG_final
clear all
close all
clc
%
```

```
N = 256;                                % number of grid points
d = 0.3;                                % exit-pupil diamter [m]
delta = d/N;                            % grid spacing [m]
wvl = 1e-6;                             % wavelength [m]
k = 2*pi/wvl;                           % angular wavenumber [rad/m]
q_P = 2.0;                              % PPRG quotient
ifov = wvl/d/q_P;                       % pixel field of view
z_P = d/(N*ifov);                       % propagation length [m] (FPA side = d)
sigma_r = 100;                          % read-noise standard deviation [pe]
l = 100e3;                              % FPA well depth [pe]
m_R = 0.75*l;                           % reference mean number of pe's [pe]
m_S = 4;                                % signal mean number of pe's [pe]
%
% Setup the object plane
%
[x0,y0] = meshgrid((-N/2 : N/2-1)*delta);
%
% Let's assume an extended object
%
dO = d;                                 % diameter of the object [m]
O = ...                                 % gen. complex object function
    cyl(x0/dO,y0/dO) ...
    - cyl(x0/(3*dO/4),y0/(3*dO/4)) ...
    - cyl(x0/(dO/2),y0/(dO/2)) ...
    + cyl(x0/(dO/4),y0/(dO/4));
phz = -pi + 2*pi*rand(N);               % unif. dist. random phase (-pi,pi]
R_O = O.*exp(1j*phz);                   % obj.-plane complex reflectance.
U_O = ...                               % obj.-plane complex field [sqrt(W)/m]
    R_O*exp(-1j*k*z_P);
%
figure(1); clf;
ax1=subplot(1,2,1);
imagesc(abs(U_O)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_O),[-pi,pi]); axis image; axis off,colormap(ax2,jet);
%
% Propagate from the object plane to the pupil plane
%
[x1,y1] = meshgrid((-N/2:N/2-1)*delta);
r1 = hypot(x1,y1);
%
M = 2*N+rem(N,2);                       % double the number of grid points
U_Og = zeros(M);                        % pad with a guard band of 2
nn = round(N/2)+(1:N);                  % original coordinates
U_Og(nn,nn) = U_O;                      % embed in padded grid
% transform the embedded field
U_Ogt = delta^2*fftshift(fft2(fftshift(U_Og)));
% create the propagation kernel
const = k/(2*z_P)*delta^2;
kernel = exp(1j*const*(ceil(-M/2):floor((M-1)/2))'.^2);
% transform the kernel
kernelt = delta*fftshift(fft(fftshift(kernel)));
% multiply and transform back with appropriate scale factors
U_Ogt = (kernelt*kernelt.').*U_Ogt;
U_Pg = k/(1j*2*pi*z_P)*exp(1j*k*z_P) ...
    *(1/delta)^2*ifftshift(ifft2(ifftshift(U_Ogt)));
% use original coordinates
U_Pm = U_Pg(nn,nn);                     % entering complex field [sqrt(W)/m]
%
figure(2); clf;
ax1=subplot(1,2,1);
```

```
imagesc(abs(U_Pm)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_Pm),[-pi,pi]); axis image; axis off,colormap(ax2,jet);
%
% Setup the pupil plane
%
phi = randn(N);                         % phase aberration [rad]
a = N/10*delta;                         % 1/e width [m]
g = exp(-r1.^2/(2*a^2));                % Gaussian function
phi = ...                               % smoothed phase aberrtaion [rad]
    real(ifft2(fft2(phi).*ifft2(ifftshift(g))));
phi = ...                               % rms = 1 [rad]
    (phi-mean(phi(:)))/std(phi(:));
P = ...                                 % gen. complex pupil function
    cyl(x1/d,y1/d).*exp(1i*phi);
U_P = ...                               % pup.-plane complex field [sqrt(W)/m]
    exp(-1j*k/(2*z_P)*(x1.^2 + y1.^2)).*P.*U_Pm;
U_S = ...                               % signal complex field [sqrt(pe)]
    sqrt(m_S)*U_P/sqrt(mean(abs(U_P(:)).^2));
display(['Sig. mean number of photoectrons: ' ...
    num2str(mean(abs(U_S(:)).^2)) ' [pe] (extended object)']);
%
figure(3); clf;
ax1=subplot(1,2,1);
imagesc(abs(U_S)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_S),[-pi,pi]); axis image; axis off,colormap(ax2,jet);
%
x_R = d*q_P/4;                          % x-axis reference origin
y_R = d*q_P/4;                          % y-axis reference origin
U_R = ...                               % reference complex field [sqrt(W)/m]
    exp(-1i*2*pi*y_R*(y1/(wvl*z_P))).*exp(-1i*2*pi*x_R*(x1/(wvl*z_P)));
U_R = ...                               % reference complex field [sqrt(pe)]
    sqrt(m_R)*U_R/sqrt(mean(abs(U_R(:)).^2));
display(['Ref. mean number of photoectrons: ' ...
    num2str(mean(abs(U_R(:)).^2)) ' [pe]']);
%
figure(4); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_R)); axis image; axis off; colormap(ax1,hot);
ax2 = subplot(1,2,2);
imagesc(angle(U_R),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
% Measure the pupil plane on a FPA
%
m_H = abs(U_S + U_R).^2;                % holgram mean number of pe's [pe]
display(['Hol. mean number of photoectrons: ' ...
    num2str(mean(m_H(:))) ' [pe] (extended object)']);
m_Hn = ...                             % add noise
    poissrnd(m_H)+sigma_r*randn(size(m_H));
% use the follwoing if you do not have the Statistics Toolbox
% m_Hn = ...                             % add noise
%     m_H+sqrt(m_S+m_R)*randn(size(m_H))+sigma_r*randn(size(m_H));
display(['Hol. mean number of photoectrons: ' ...
    num2str(mean(m_Hn(:))) ' [pe] (extended object with noise)']);
%
m_H = m_H - mean(m_H(:));               % remove DC
m_Hn = m_Hn - mean(m_Hn(:));            % remove DC
%
figure(5); clf;
ax1 = subplot(1,2,1);
imagesc(m_H); axis image; axis off; colormap(ax1,hsv);
```

```
ax2 = subplot(1,2,2);
imagesc(m_Hn); axis image; axis off; colormap(ax2,hsv);
%
M = round(q_P*N);                    % desired number of grid points
m_Hng = zeros(M);                    % pad with a gaurdband
nn = round(M/2-N/2)+(1:N);           % original coordinates
m_Hng(nn,nn) = m_Hn;                 % embed in padded grid
%
% Go to the Fourier plane to obtain an estimate
%
m_Hngt = ...                         % Fourier plane
    delta^2*fftshift(fft2(fftshift(m_Hng)));
%
figure(6); clf;
ax1 = subplot(1,2,1);
imagesc(abs(m_Hngt)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(m_Hngt),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
nn = round(3*M/4-N/2)+(1:N);         % estimate coordinates
U_Ie = ...                           % window Fourier plane
    cyl(x1/d,y1/d).*m_Hngt(nn,nn);
%
figure(7); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_Ie)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_Ie),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
SNR = ...                            % calculate SNR
    4*q_P*m_R*m_S/(pi*(m_R + sigma_r^2));
display(['PPRG signal-to-noise ratio      : ' num2str(SNR)]);
%
% Let's assume a point-source object
%
U_P = P;                             % pup.-plane complex field [sqrt(W)/m]
U_S = ...                            % signal complex field [sqrt(pe)]
    sqrt(m_S)*U_P/sqrt(mean(abs(U_P(:)).^2));
display(['Sig. mean number of photoectrons: ' ...
    num2str(mean(abs(U_S(:)).^2)) ' [pe] (point-source object)']);
%
figure(8); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_S)); axis image; axis off; colormap(ax1,hot);
ax2 = subplot(1,2,2);
imagesc(angle(U_S),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
x_R = d*q_P/4;                       % x-axis reference origin
y_R = d*q_P/4;                       % y-axis reference origin
U_R = ...                            % reference complex field [sqrt(W)/m]
    exp(-1i*2*pi*y_R*(y1/(wvl*z_P))).*exp(-1i*2*pi*x_R*(x1/(wvl*z_P)));
U_R = ...                            % reference complex field [sqrt(pe)]
    sqrt(m_R)*U_R/sqrt(mean(abs(U_R(:)).^2));
display(['Ref. mean number of photoectrons: ' ...
    num2str(mean(abs(U_R(:)).^2)) ' [pe]']);
%
figure(9); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_R)); axis image; axis off; colormap(ax1,hot);
ax2 = subplot(1,2,2);
imagesc(angle(U_R),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
```

```
% Measure the pupil plane on a FPA
%
m_H = abs(U_S + U_R).^2;            % holgram mean number of pe's [pe]
display(['Hol. mean number of photoectrons: ' ...
    num2str(mean(m_H(:))) ' [pe] (point-source object)']);
m_Hn = ...                          % add noise
    poissrnd(m_H)+sigma_r*randn(size(m_H));
% use the follwoing if you do not have the Statistics Toolbox
% m_Hn = ...                          % add noise
%     m_H+sqrt(m_S+m_R)*randn(size(m_H))+sigma_r*randn(size(m_H));
display(['Hol. mean number of photoectrons: ' ...
    num2str(mean(m_Hn(:))) ' [pe] (point-source object with noise)']);
%
m_H = m_H - mean(m_H(:));           % remove DC
m_Hn = m_Hn - mean(m_Hn(:));        % remove DC
%
figure(10); clf;
ax1 = subplot(1,2,1);
imagesc(m_H); axis image; axis off; colormap(ax1,hsv);
ax2 = subplot(1,2,2);
imagesc(m_Hn); axis image; axis off; colormap(ax2,hsv);
%
% Go to the Fourier plane to obtain an estimate
%
m_Hnt = ...                         % Fourier plane
    delta^2*fftshift(fft2(fftshift(m_Hn)));
%
figure(11); clf;
ax1 = subplot(1,2,1);
imagesc(abs(m_Hnt)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(m_Hnt),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
nn = ...                            % estimate coordinates
    round(3*N/4-round(N/q_P)/2)+(1:round(N/q_P));
M = length(nn);                     % estimate number of grid points
[xf,yf] = meshgrid((-M/2 : M/2-1)*d*q_P/N);
U_Ie = ...                          % window Fourier plane
    cyl(xf/d,yf/d).*m_Hnt(nn,nn);
%
figure(12); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_Ie)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_Ie),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
nn = round(N/2-M/2)+(1:M);          % desired coordinates
U_Ieg = zeros(N);                   % pad with a gaurdband
U_Ieg(nn,nn) = U_Ie;                % embed in padded grid
U_Pe = ...
    (d*q_P)^2*fftshift(ifft2(fftshift(U_Ieg))).*cyl(x1/d,y1/d);
%
figure(13),clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_Pe)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_Pe),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
S_F = ...                           % calculate field-estimated Strehl
    abs(mean(U_P(:).*conj(U_Pe(:)))).^2 ...
    /(mean(abs(U_P(:)).^2).*mean(abs(U_Pe(:)).^2));
display(['Field-estimated Strehl ratio    : ' num2str(S_F)]);
```

Listing 4. MATLAB® code for the on-axis phase shifting recording geometry or PSRG.

```
% script_PSRG_final
%
clear all
close all
clc
%
N = 256;                              % number of grid points
d = 0.3;                              % exit-pupil diamter [m]
delta = d/N;                          % grid spacing [m]
wvl = 1e-6;                           % wavelength [m]
k = 2*pi/wvl;                         % angular wavenumber [rad/m]
q_I = 4.0;                            % IPRG quotient
ifov = wvl/d/q_I;                     % pixel field of view
z_I = d/(N*ifov);                     % propagation length [m] (FPA side = d)
sigma_r = 100;                        % read-noise standard deviation [pe]
l = 100e3;                            % FPA well depth [pe]
m_R = 0.75*l;                         % reference mean number of pe's [pe]
S = 1;                                % number of amplitude splits
m_S = 4/S;                            % signal mean number of pe's [pe]
%
% Setup the pupil plane
%
[x1,y1] = meshgrid((-N/2 : N/2-1)*delta);
r1 = hypot(x1,y1);
%
phi = randn(N);                       % phase aberration [rad]
a = N/10*delta;                       % 1/e width [m]
g = exp(-r1.^2/(2*a^2));              % Gaussian function
phi = ...                             % smoothed phase aberrtaion [rad]
    real(ifft2(fft2(phi).*ifft2(ifftshift(g))));
phi = ...                             % rms = 1 [rad]
    (phi-mean(phi(:)))/std(phi(:));
P = ...                               % gen. complex pupil function
    cyl(x1/d,y1/d).*exp(1i*phi);
%
% Let's assume a point-source object
%
U_P = P;                              % pup.-plane complex field [sqrt(W)/m]
U_S = ...                             % signal complex field [sqrt(pe)]
    sqrt(m_S)*U_P/sqrt(mean(abs(U_P(:)).^2));
display(['Sig. mean number of photoectrons: ' ...
    num2str(mean(abs(U_S(:)).^2)) ' [pe] (point-source object)']);
%
figure(1); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_S)); axis image;  axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_S),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
U_R = 1;                              % reference complex field [sqrt(W)/m]
U_R1 = ...                            % reference complex field [sqrt(pe)]
    sqrt(m_R)/sqrt(mean(abs(U_R(:)).^2));
display(['Ref. mean number of photoectrons: ' ...
    num2str(mean(abs(U_R1(:)).^2)) ' [pe] (1)']);
U_R2 = ...                            % reference complex field [sqrt(pe)]
    sqrt(m_R)*exp(-1i*pi/2)/sqrt(mean(abs(U_R(:)).^2));
display(['Ref. mean number of photoectrons: ' ...
    num2str(mean(abs(U_R2(:)).^2)) ' [pe] (2)']);
U_R3 = ...                            % reference complex field [sqrt(pe)]
```

```
    sqrt(m_R)*exp(-1i*pi)/sqrt(mean(abs(U_R(:)).^2));
display(['Ref. mean number of photoectrons: ' ...
    num2str(mean(abs(U_R3(:)).^2)) ' [pe] (3)']);
U_R4 = ...                          % reference complex field [sqrt(pe)]
    sqrt(m_R)*exp(-1i*3*pi/2)/sqrt(mean(abs(U_R(:)).^2));
display(['Ref. mean number of photoectrons: ' ...
    num2str(mean(abs(U_R4(:)).^2)) ' [pe] (4)']);
%
figure(2); clf;
ax1 = subplot(2,2,1);
imagesc(angle(U_R1),[-pi,pi]); axis image; axis off; colormap(ax1,hsv);
ax2 = subplot(2,2,2);
imagesc(angle(U_R2),[-pi,pi]); axis image; axis off; colormap(ax2,hsv);
ax3 = subplot(2,2,3);
imagesc(angle(U_R3),[-pi,pi]); axis image; axis off; colormap(ax3,hsv);
ax4 = subplot(2,2,4);
imagesc(angle(U_R4),[-pi,pi]); axis image; axis off; colormap(ax4,hsv);
%
% Measure the pupil plane on a FPA
%
m_H1 = abs(U_S + U_R1).^2;          % holgram mean number of pe's [pe]
display(['Hol. mean number of photoectrons: ' ...
    num2str(mean(m_H1(:))) ' [pe] (point-source object 1)']);
m_H2 = abs(U_S + U_R2).^2;          % holgram mean number of pe's [pe]
display(['Hol. mean number of photoectrons: ' ...
    num2str(mean(m_H2(:))) ' [pe] (point-source object 2)']);
m_H3 = abs(U_S + U_R3).^2;          % holgram mean number of pe's [pe]
display(['Hol. mean number of photoectrons: ' ...
    num2str(mean(m_H3(:))) ' [pe] (point-source object 3)']);
m_H4 = abs(U_S + U_R4).^2;          % holgram mean number of pe's [pe]
display(['Hol. mean number of photoectrons: ' ...
    num2str(mean(m_H4(:))) ' [pe] (point-source object 4)']);
%
figure(3); clf;
ax1 = subplot(2,2,1);
imagesc(m_H1); axis image; axis off; colormap(ax1,hsv);
ax2 = subplot(2,2,2);
imagesc(m_H2); axis image; axis off; colormap(ax2,hsv);
ax3 = subplot(2,2,3);
imagesc(m_H3); axis image; axis off; colormap(ax3,hsv);
ax4 = subplot(2,2,4);
imagesc(m_H4); axis image; axis off; colormap(ax4,hsv);
%
m_H1n = ...                         % add noise
    poissrnd(m_H1)+sigma_r*randn(size(m_H1));
display(['Hol. mean number of photoectrons: ' ...
    num2str(mean(m_H1n(:))) ' [pe] (point-source object with noise 1)']);
m_H2n = ...                         % add noise
    poissrnd(m_H2)+sigma_r*randn(size(m_H2));
display(['Hol. mean number of photoectrons: ' ...
    num2str(mean(m_H2n(:))) ' [pe] (point-source object with noise 2)']);
m_H3n = ...                         % add noise
    poissrnd(m_H3)+sigma_r*randn(size(m_H3));
display(['Hol. mean number of photoectrons: ' ...
    num2str(mean(m_H3n(:))) ' [pe] (point-source object with noise 3)']);
m_H4n = ...                         % add noise
    poissrnd(m_H4)+sigma_r*randn(size(m_H4));
display(['Hol. mean number of photoectrons: ' ...
    num2str(mean(m_H4n(:))) ' [pe] (point-source object with noise 4)']);
%
```

```
figure(4); clf;
ax1 = subplot(2,2,1);
imagesc(m_H1n); axis image; axis off; colormap(ax1,hsv);
ax2 = subplot(2,2,2);
imagesc(m_H2n); axis image; axis off; colormap(ax2,hsv);
ax3 = subplot(2,2,3);
imagesc(m_H3n); axis image; axis off; colormap(ax3,hsv);
ax4 = subplot(2,2,4);
imagesc(m_H4n); axis image; axis off; colormap(ax4,hsv);
%
U_Se = ((m_H1n-m_H3n) + 1j*(m_H4n-m_H2n)).*cyl(x1/d,y1/d);
%
figure(5); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_Se)); axis image; axis off; colormap(ax1,hot);
ax2 = subplot(1,2,2);
imagesc(angle(U_Se),[-pi,pi]); axis image; axis off; colormap(ax2,hsv);
%
S_F = ...                              % calculate field-estimated Strehl
    abs(mean(U_S(:).*conj(U_Se(:)))).^2 ...
    /(mean(abs(U_S(:)).^2).*mean(abs(U_Se(:)).^2));
display(['Field-estimated Strehl Ratio    : ' num2str(S_F)]);
%
% Setup the object plane
%
[x0,y0] = meshgrid((-N/2 : N/2-1)*delta);
%
% Let's assume an extended object
%
dO = d;                                % diameter of the object [m]
O = ...                                % gen. complex object function
    cyl(x0/dO,y0/dO) ...
    - cyl(x0/(3*dO/4),y0/(3*dO/4)) ...
    - cyl(x0/(dO/2),y0/(dO/2)) ...
    + cyl(x0/(dO/4),y0/(dO/4));
phz = -pi + 2*pi*rand(N);              % unif. dist. random phase (-pi,pi]
R_O = O.*exp(1j*phz);                  % obj.-plane complex reflectance.
%
figure(6); clf;
ax1 = subplot(1,2,1);
imagesc(abs(R_O)); axis image;  axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(R_O),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
U_Pp = ...                             % leaving complex field [sqrt(W)/m]
    exp(-1j*k/(2*z_I)*(x1.^2 + y1.^2)).*U_P;
%
figure(7); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_Pp)); axis image;  axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_Pp),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
% Propagate from the pupil plane to the image plane
%
[x2,y2] = meshgrid((-N/2:N/2-1)*delta);
%
M = 2*N+rem(N,2);                      % double the number of grid points
U_Ppg = zeros(M);                      % pad with a guard band of 2
nn = round(N/2)+(1:N);                 % original coordinates
U_Ppg(nn,nn) = U_Pp;                   % embed in padded grid
```

```
% transform the embedded field
U_Ppgt = delta^2*fftshift(fft2(fftshift(U_Ppg)));
% create the propagation kernel
const = k/(2*z_I)*delta^2;
kernel = exp(1j*const*(ceil(-M/2):floor((M-1)/2))'.^2);
% transform the kernel
kernelt = delta*fftshift(fft(fftshift(kernel)));
% multiply and transform back with appropriate scale factors
U_Ppgt = (kernelt*kernelt.').*U_Ppgt;
U_Ip = 1/(1j*wvl*z_I)*exp(1j*k*z_I) ...
    *(1/delta)^2*ifftshift(ifft2(ifftshift(U_Ppgt)));
% use original coordinates
U_I = U_Ip(nn,nn);                  % img.-plane complex field [sqrt(W)/m]
%
figure(8); clf;
ax1=subplot(1,2,1);
imagesc(abs(U_I)); axis image; axis off; colormap(ax1,hot);
ax2=subplot(1,2,2);
imagesc(angle(U_I),[-pi,pi]); axis image; axis off,colormap(ax2,jet);
%
U_I = ...                           % img.-plane complex field [sqrt(W)/m]
    ifftshift(ifft2(ifftshift( fftshift(fft2(fftshift(R_O))) ...
    .* fftshift(fft2(fftshift((U_I)))) )));
U_S = ...                           % signal complex field [sqrt(pe)]
    sqrt(m_S)*U_I/sqrt(mean(abs(U_I(:)).^2));
display(['Sig. mean number of photoectrons: ' ...
    num2str(mean(abs(U_S(:)).^2)) ' [pe] (extended object)']);
%
figure(9); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_S)); axis image; axis off; colormap(ax1,hot);
ax2 = subplot(1,2,2);
imagesc(angle(U_S),[-pi,pi]); axis image; axis off; colormap(ax2,jet);
%
U_R = 1;                            % reference complex field [sqrt(W)/m]
U_R1 = ...                          % reference complex field [sqrt(pe)]
    sqrt(m_R)/sqrt(mean(abs(U_R(:)).^2));
display(['Ref. mean number of photoectrons: ' ...
    num2str(mean(abs(U_R1(:)).^2)) ' [pe] (1)']);
U_R2 = ...                          % reference complex field [sqrt(pe)]
    sqrt(m_R)*exp(-1i*pi/2)/sqrt(mean(abs(U_R(:)).^2));
display(['Ref. mean number of photoectrons: ' ...
    num2str(mean(abs(U_R2(:)).^2)) ' [pe] (2)']);
U_R3 = ...                          % reference complex field [sqrt(pe)]
    sqrt(m_R)*exp(-1i*pi)/sqrt(mean(abs(U_R(:)).^2));
display(['Ref. mean number of photoectrons: ' ...
    num2str(mean(abs(U_R3(:)).^2)) ' [pe] (3)']);
U_R4 = ...                         % reference complex field [sqrt(pe)]
    sqrt(m_R)*exp(-1i*3*pi/2)/sqrt(mean(abs(U_R(:)).^2));
display(['Ref. mean number of photoectrons: ' ...
    num2str(mean(abs(U_R4(:)).^2)) ' [pe] (4)']);
%
figure(10); clf;
ax1 = subplot(2,2,1);
imagesc(angle(U_R1),[-pi,pi]); axis image; axis off; colormap(ax1,hsv);
ax2 = subplot(2,2,2);
imagesc(angle(U_R2),[-pi,pi]); axis image; axis off; colormap(ax2,hsv);
ax3 = subplot(2,2,3);
imagesc(angle(U_R3),[-pi,pi]); axis image; axis off; colormap(ax3,hsv);
ax4 = subplot(2,2,4);
imagesc(angle(U_R4),[-pi,pi]); axis image; axis off; colormap(ax4,hsv);
%
```

```
% Measure the image plane on a FPA
%
m_H1 = abs(U_S + U_R1).^2;          % holgram mean number of pe's [pe]
display(['Hol. mean number of photoectrons: ' ...
    num2str(mean(m_H1(:))) ' [pe] (extended object 1)']);
m_H2 = abs(U_S + U_R2).^2;          % holgram mean number of pe's [pe]
display(['Hol. mean number of photoectrons: ' ...
    num2str(mean(m_H2(:))) ' [pe] (extended object 2)']);
m_H3 = abs(U_S + U_R3).^2;          % holgram mean number of pe's [pe]
display(['Hol. mean number of photoectrons: ' ...
    num2str(mean(m_H3(:))) ' [pe] (extended object 3)']);
m_H4 = abs(U_S + U_R4).^2;          % holgram mean number of pe's [pe]
display(['Hol. mean number of photoectrons: ' ...
    num2str(mean(m_H4(:))) ' [pe] (extended object 4)']);
%
figure(11); clf;
ax1 = subplot(2,2,1);
imagesc(m_H1); axis image; axis off; colormap(ax1,hsv);
ax2 = subplot(2,2,2);
imagesc(m_H2); axis image; axis off; colormap(ax2,hsv);
ax3 = subplot(2,2,3);
imagesc(m_H3); axis image; axis off; colormap(ax3,hsv);
ax4 = subplot(2,2,4);
imagesc(m_H4); axis image; axis off; colormap(ax4,hsv);
%
m_H1n = ...                         % add noise
    poissrnd(m_H1)+sigma_r*randn(size(m_H1));
display(['Hol. mean number of photoectrons: ' ...
    num2str(mean(m_H1n(:))) ' [pe] (extended object with noise 1)']);
m_H2n = ...                         % add noise
    poissrnd(m_H2)+sigma_r*randn(size(m_H2));
display(['Hol. mean number of photoectrons: ' ...
    num2str(mean(m_H2n(:))) ' [pe] (extended object with noise 2)']);
m_H3n = ...                         % add noise
    poissrnd(m_H3)+sigma_r*randn(size(m_H3));
display(['Hol. mean number of photoectrons: ' ...
    num2str(mean(m_H3n(:))) ' [pe] (extended object with noise 3)']);
m_H4n = ...                         % add noise
    poissrnd(m_H4)+sigma_r*randn(size(m_H4));
display(['Hol. mean number of photoectrons: ' ...
    num2str(mean(m_H4n(:))) ' [pe] (extended object with noise 4)']);
%
figure(12); clf;
ax1 = subplot(2,2,1);
imagesc(m_H1n); axis image; axis off; colormap(ax1,hsv);
ax2 = subplot(2,2,2);
imagesc(m_H2n); axis image; axis off; colormap(ax2,hsv);
ax3 = subplot(2,2,3);
imagesc(m_H3n); axis image; axis off; colormap(ax3,hsv);
ax4 = subplot(2,2,4);
imagesc(m_H4n); axis image; axis off; colormap(ax4,hsv);
%
U_Se = (m_H1n-m_H3n) + 1j*(m_H4n-m_H2n);
%
figure(13); clf;
ax1 = subplot(1,2,1);
imagesc(abs(U_Se)); axis image; axis off; colormap(ax1,hot);
ax2 = subplot(1,2,2);
imagesc(angle(U_Se),[-pi,pi]); axis image; axis off; colormap(ax2,hsv);
%
SNR = ...                           % calculate SNR
    4*m_R*m_S/(S^2*(m_R + sigma_r^2));
display(['PSRG signal-to-noise ratio      : ' num2str(SNR)]);
```

See also: Heterodyning

References

Dereniak, E.L., Boreman, G.D., 1996. Infrared Detectors and Systems. New York: John Wiley and Sons.
Frieden, B.R., 2001. Probability, Statistical Optics, and Data Testing, third ed. New York: Springer-Verlag.
Gaskill, J.D., 1978. Linear Systems, Fourier Transforms, and Optics. New York: John Wiley and Sons.
Goodman, J.W., 2007. Speckle Phenomena in Optics Theory and Application. Englewood: Roberts and Company.
Janesick, J.R., 2007. Photon Transfer. Bellingham: SPIE Press.
McManamon, P.F., 2015. Field Guide to Lidar. Bellingham: SPIE Press.
Poon, T.-C., Liu, J.-P., 2014. Introduction to Modern Digital Holography. New York: Cambridge University Press.
Rhoadarmer, T.A., Barchers, J.D., 2002. Noise Analysis for complex field etimation using a self-referencing interferometer wave front sensor. s.l. SPIE.
Saleh, B.E.A., Teich, M.C., 2007. Fundamentals of Photonics, second ed. New York: John Wiley and Sons.
Spencer, M.F., 2014. The Scattering of Partially Coherent Electromagnetic Beam Illumination from Statistically Rough Surfaces. PhD Dissertation, Air Force Institute of Technology, Volume ADA603227.
Spencer, M.F., Raynor, R.A., Banet, M.T., Marker, D.K., 2016. Deep-turbulence wavefront sensing using digital-holographic detection in the off-axis image plane recording geometry. Optical Engineering 56 (3), 031213.
Tippie, A.E., 2012. Aberration Correction in Digital Holography. PhD Dissertation, University of Rochester, Volume AS38.6635.

Further Reading

Steinbock, M.J., Hyde IV, M.W., Schmidt, J.D., 2014. LSPV + 7, a branch-point-tolerant reconstructor for strong turbulence adaptive optics. Applied Optics 53 (18), 3821–3831.

3D Metrics for Airborne Topographic Lidar

Shea T Hagstrom and Myron Z Brown, The Johns Hopkins University Applied Physics Laboratory, Laurel, MD, United States

Introduction

The development and widespread use of airborne topographic lidar systems is one of the most significant changes within the remote sensing field in recent decades. Lidar is now commonly used for production of high-resolution foundation data such as Digital Elevation Models (DEMs) and bare-earth Digital Terrain Models (DTMs) as well as for mapping power lines and discovering ancient ruins under jungle canopy. Community standards for 3D lidar metrics have emerged to set minimum expectations for lidar data quality and to enable objective and quantitative evaluation and comparison of systems and capabilities.

Lidar data quality depends on sensor technology (e.g., well established linear-mode lidar and more recently emerging photon counting lidar), system design, calibration, operating environment, and post-processing such as registration of overlapping swaths and noise filtering. Metrics for assessment of 3D lidar point clouds must be general enough to account for all of these factors. Where possible, metrics should also be general enough to apply to 3D data derived from other sources such as multiple view imagery (Bosch *et al.*, 2016) and synthetic aperture radar (Zhu and Bamler, 2012), though we do not explicitly address these in this article. As lidar system designs continue to evolve, and as lidar technology continues to be applied in new ways, metrics for ensuring acceptable data quality are advancing as well to keep up. This article reviews lidar metrics currently in common use, highlights emerging challenges, and discusses recent work to begin to address those challenges.

Current Lidar Metric Standards

Minimum expectations for topographic mapping lidar metric performance are captured in recent community standards for lidar collection and government specifications for delivered lidar products. The United States Geological Survey (USGS) National Geospatial Program (NGP) released the Lidar Base Specification in 2014 to define minimum metric requirements for the vertical accuracy, data density, and completeness of lidar products for the National interagency 3D Elevation Program (3DEP). Similarly, the American Society for Photogrammetry and Remote Sensing (ASPRS) released the ASPRS Positional Accuracy Standards for Digital Geospatial Data in 2015 to specify vertical and horizontal accuracy requirements for data providers. While other organizations and agencies also independently define metric standards, these documents describe the current state of the industry in metric performance characterization of lidar products. The following discussion draws heavily from these documents to summarize commonly used lidar metrics to assess data density, completeness, and accuracy.

Density and Completeness

Data density is perhaps the most intuitive metric associated with a lidar point cloud, as it indicates what sizes of objects in a scene can be expected to be well represented. The terms nominal pulse spacing (NPS) and nominal pulse density (NPD) are commonly used to indicate lidar point spacing and density respectively on a measured ground surface within the area covered by data of a single flight line or scan. The lidar base specification also defines the terms aggregate nominal pulse spacing (ANPS) and aggregate nominal pulse density (ANPD) to indicate point spacing and density more generally for point clouds that include points from multiple flight lines. Furthermore, to properly characterize point spacing and density for lidar systems incorporating focal plane arrays, even more general terms would be required. The intent of any of these terms is to capture the equivalent of ground sample distance (GSD) in remote sensing digital photography, though lidar point spacing is generally irregular. These metrics are two-dimensional and intended primarily for conventional topographic lidar data collection. In the section on emerging capabilities, a more general three-dimensional metric for data density is discussed that is suitable for lidar collections consisting of multiple off-nadir viewpoints to increase coverage of objects in the scene.

In addition to nominal point density, the completeness and spatial distribution of points are also important in determining the utility of a point cloud. Gaps in lidar coverage are commonly termed data voids. The lidar base specification defines a data void as any area of missing points that is at least as large as $4 \times \text{ANPS}^2$ and indicates that these data voids are unacceptable in delivered lidar products unless caused by water bodies or low near infrared surface reflectivity. Regularity of the spatial distribution of points is also required. This is measured by producing a two-dimensional occupancy grid with cell size $2 \times \text{ANPS}$ and determining if at least 90% of cells are occupied. These metrics are also intended primarily for conventional topographic lidar. A more general discussion of three-dimensional completeness is provided in the section on emerging capabilities.

Absolute and Relative Vertical Accuracy

Absolute vertical accuracy of lidar elevation values is typically estimated by comparison with check point elevation values that have been independently measured in the lidar coverage area. ASPRS recommends that check points be at least three times more accurate than the required accuracy of the product and located in areas with low and uniform slope to avoid interpolation errors.

Highly accurate check points are typically obtained by Global Positioning System (GPS) survey. Control points used for calibration of a lidar system are also typically obtained by GPS survey and should not be included as check points. Absolute vertical accuracy is reported separately in non-vegetated and vegetated terrain and is commonly reported as root mean squared error (RMSE) for non-vegetated terrain, assuming a normal error distribution, or percentile confidence level for vegetated terrain.

Relative vertical accuracy refers to precision or repeatability and internal consistency of elevation measurements in a lidar product. Vertical precision is estimated as the variation in elevation values measured on a flat hard surface such as a building roof. Internal consistency is estimated as the variation in elevation values measured on non-vegetated terrain in overlapping swaths. ASPRS and USGS define standards for repeatability based on the maximum measured difference and for consistency based on both root mean squared difference (RMSD) and maximum difference. Sources of relative vertical error include range precision of the lidar instrument and registration error among overlapping lidar swaths, among others.

Absolute and Relative Horizontal Accuracy

Absolute horizontal accuracy of a lidar product is estimated by comparison with horizontal check points that have been independently measured in the lidar coverage area. Unlike vertical check points described above, horizontal check points must be located at points in the lidar coverage area for which the horizontal position may be very accurately measured independently and which are clearly identifiable in the lidar product itself. These check points are typically obtained by GPS survey and the locations clearly documented with photography. ASPRS defines horizontal accuracy standards for RMSE of X and Y coordinates independently, RMSE of radial accuracy, and percentile confidence level of radial accuracy.

Relative horizontal accuracy refers to point to point horizontal distance measurement accuracy within a geospatial product and may be estimated similarly to absolute horizontal accuracy using check points. ASPRS does not define an explicit standard for relative horizontal accuracy in lidar datasets, and absolute horizontal accuracy standards for lidar collection are sufficiently stringent that independently measuring relative horizontal accuracy would have limited utility. However, in situations for which ground control and GPS accuracy may be limited, such as remote lidar collection for humanitarian purposes, explicitly measuring relative horizontal error may be desirable. Methods for beginning to address these issues are discussed in the following section on emerging capabilities.

Emerging Capabilities and Challenges

Current lidar metrics standards have been developed for use with mature commercial linear-mode lidar technology, fixed nadir viewing angles, and accurate ground survey. New lidar technology, system designs, and use cases may require additional or revised metrics to accurately characterize error. For instance, more emphasis on measuring relative horizontal accuracy may be desirable for recent photon counting lidar systems which require post-processing for photon aggregation and noise filtering that can improve or degrade resolution. Similarly, comprehensive modeling of all error sources may be more desirable for lidar systems that collect from higher altitudes and with significantly off-nadir look angles such that some errors often considered negligible become important. Off-nadir data collection also enables enhanced foliage penetration and true three-dimensional mapping capabilities. Current metrics and conventional wisdom for topographic lidar data collection are insufficient to fully characterize these new collection capabilities. The following sections discuss preliminary efforts within the lidar research community to begin to address some of these recent challenges.

Defining a Lidar Error Model

Inaccuracies in lidar point cloud geospatial coordinates are primarily influenced by uncertainties in sensor position, sensor orientation, and range measurements as described by Habib *et al.* (2010). These uncertainties are commonly characterized for lidar system error budgets and sensor calibration processing; however, they generally are not readily available to end users of lidar data. While current standards may ensure that products meet minimum performance specifications, additional knowledge of specific uncertainties is desirable for an end user who wishes to merge or compare multiple datasets.

Rodarmel *et al.* (2015) have proposed a Universal Lidar Error Model (ULEM) that captures sensor-space parameter uncertainties and explains how to mathematically propagate errors into ground plane covariance predictions. The capability of ULEM to predict absolute horizontal and vertical accuracy of air-to-ground lidar data was demonstrated using check points surveyed with GPS combined with standard metric analysis methods. Horizontal or Circular Error at the 90th percentile (CE90) and vertical or linear error at the 90th percentile (LE90) were computed, consistent with National Geospatial-Intelligence Agency (NGA) practices, and approximately 90% of the estimated check point accuracies were determined to be within predicted bounds, indicating the potential utility of this model for practical use. **Fig. 1** shows the concept of the ULEM and how the error in horizontal and vertical accuracy leads to position uncertainty following an ellipsoidal shape.

An ongoing challenge in the adoption of a lidar error model is incorporating metadata required to support such models into standard lidar file formats. ULEM metadata variable length records (VLRs) have been developed that are compliant with the ASPRS-developed LAS specification. Similarly, ULEM metadata is defined in an optional header in the BPF3 specification adopted by the National Center for Geospatial Intelligence Standards in 2015. At present, ULEM is used within specialized government

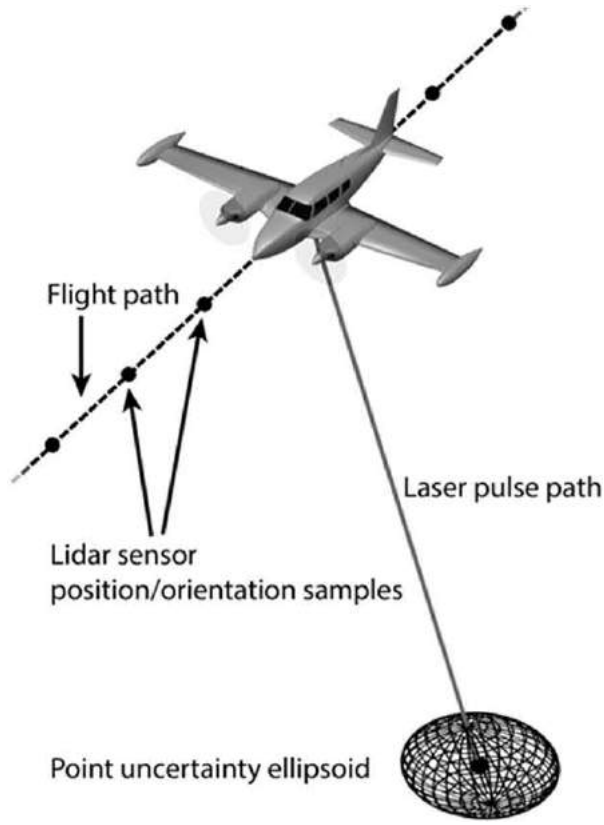


Fig. 1 Universal Lidar Error Model (ULEM) sensor-space error modeling.

communities. With the widespread proliferation of lidar and other 3D data from a variety of sources, development of common standards for metadata used to support error modeling and error propagation is becoming increasingly important.

Measuring Point Cloud Resolution

Horizontal resolution of a lidar point cloud refers to the minimum distance between two adjacent objects that are still resolvable as unique objects. Resolution is often assumed to be equivalent to NPS or GSD, but this is often not the case. For a conventional lidar system, horizontal resolution is largely determined by flight altitude and laser beam divergence. For a system with a focal plane array detector, resolution is also determined by pixel pitch. For photon-counting lidar, the resolution of the point cloud may be degraded by the coincidence processing algorithms used to filter noise. For any point cloud generated by aggregating multiple spatially coincident datasets, resolution may be degraded by sensor calibration error or by registration error. Measuring horizontal resolution explicitly can be an important factor in determining a sensor's capability to produce an interpretable 3D image.

Stevens *et al.* (2011) report methods for estimating resolution achievable by a lidar sensor. Among other methods, horizontal resolution was determined by calculating the contrast transfer function (CTF) based on lidar point densities both on and between pairs of closely spaced reflective panels from an array of panels with varying width and spacing. Visual inspection of panel resolvability in the point clouds was shown to be consistent with the Rayleigh criterion, indicating the utility of this approach. This method for determining resolution requires significant oversampling for reliable calculation and unique targets in the lidar coverage area. While this method provides a useful tool for lidar system calibration and characterization, it is not practical for general use to assess product quality. Methods for calculating the edge spread function (ESF) and line spread function (LSF) were also proposed for estimating horizontal precision using linear features such as building edges which would not require unique calibration targets in the lidar coverage area. However, for reliable calculation these methods also require significant oversampling, and the analytical assumptions associated with these methods do not hold for the broad range of modern lidar system designs. While resolution remains difficult to explicitly measure except in controlled environments, new methods for measuring relative horizontal accuracy are discussed in the next section that may be sufficient to characterize the horizontal precision of lidar point clouds.

Accounting for Relative Horizontal Accuracy

The ASPRS lidar calibration/validation Working Group led by USGS recognizes that current standards are not sufficient to capture the relative horizontal accuracy of lidar data. To address this shortcoming, Sampath *et al.* (2014) define data quality measures

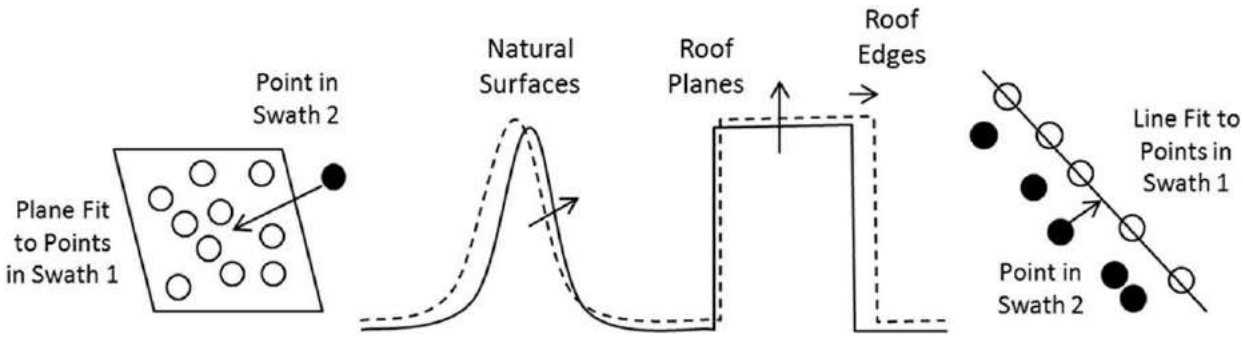


Fig. 2 Relative horizontal and vertical errors with overlapping lidar swaths capture point-to-plane differences for sloped surfaces and flat roof planes and point-to-line differences for roof edges.

(DQM) for determining inter-swath goodness of fit (**Fig. 2**) based on quality control methods proposed by Habib *et al.* (2010). These measures have been proposed to verify the quality of lidar system calibration and other sources of relative horizontal error in the data. For natural sloped terrain surfaces, a DQM measures the point to plane distance for conjugate points in overlapping swaths. Since a lidar system is unlikely to measure the same exact point in two swaths, a tangential plane is fit to each point in one swath and the point to plane distance is computed rather than point to point distance. For roof planes, conjugate planes are extracted from two swaths, and the distance from the centroid point from one swath to the plane from the other swath is measured. For roof edges, lines are fit to the conjugate edges and the distance from the centroid point in one line is measured to the conjugate line. These metrics do not rely on additional ground control or unique targets in the scene and so can be practically used for routine assessment of relative accuracy; however, they require multiple overlapping swaths which may not always be available. The working group is currently evaluating prototype tools that implement the DQMs before making formal recommendations for their use. Metrics like these will become increasingly important as more complex lidar system designs requiring more careful calibration and post-processing become commercially available.

Characterizing Foliage Penetration Performance

As reviewed by Wulder *et al.* (2012), multiple-return lidar provides a useful capability for measuring tree height, volume, and biomass for forestry inventory. Federal Emergency Management Agency (FEMA) Lidar Specifications for Flood Hazard Mapping (2009) suggest that high point densities may enable satisfactory data collection under foliage canopy. The USGS lidar base specification requires a minimum ANPD of two pulses per square meter (referred to by USGS as quality level 2) for lidar to be suitable for the 3DEP. For forestry applications requiring canopy penetration, 4–8 pulses per square meter (quality level 1 in the base specification) is commonly recommended in the literature. However, ANPD is insufficient to fully characterize foliage penetration performance.

For high obscuration environments (e.g., multiple layers of foliage canopy), Burton (2010) examines the value of line of sight angle diversity to improve the likelihood of canopy penetration. Both linear mode (Roth *et al.*, 2007) and photon counting (Vaidyanathan *et al.*, 2007) lidar systems have been developed that incorporate a gimbal to enable multiple off-nadir line of sight angles that better exploit the nonuniform gap structure in foliage canopy and improve foliage penetration performance. Current rules of thumb based on average point density also do not account for saturation in modern single photon detection lidar systems as described by Henriksson (2005) or the influence of filtering in coincidence processing as discussed in Stevens *et al.* (2011). To better account for the increasing variety of lidar system designs, explicit measurement of point density on the ground under foliage canopy is desirable to accurately characterize performance and determine data utility.

Defining Density and Completeness in 3D

Current two-dimensional point density metrics are useful for evaluating coverage and finding voids in relatively flat regions, but unfortunately do not work in areas with significant 3D structure because the sample density and completeness of vertical surfaces is ignored. No standard definition of 3D point density exists yet, making evaluation of coverage difficult. In addition, the presence of obscurants means that complete coverage requires sensing from multiple viewpoints, a consequence that is not captured by 2D metrics.

Prior works such as those of Popescu and Zhao (2008); Pyysalo and Sarjakoski (2008) have evaluated 3D point density by using a grid of volumetric cells, known as voxels, and counting the number of points within each cell to derive the density. However, this makes the assumption that the entire volume is evenly sampled by the lidar which is typically not true due to many effects including lidar scan patterns and obscurations. A lack of points within a volume may be due to either an absence of objects or a lidar system being unable to sense certain objects. Ambiguity between those two scenarios makes evaluating 3D completeness and voids difficult, if not impossible, using existing point-based methods.

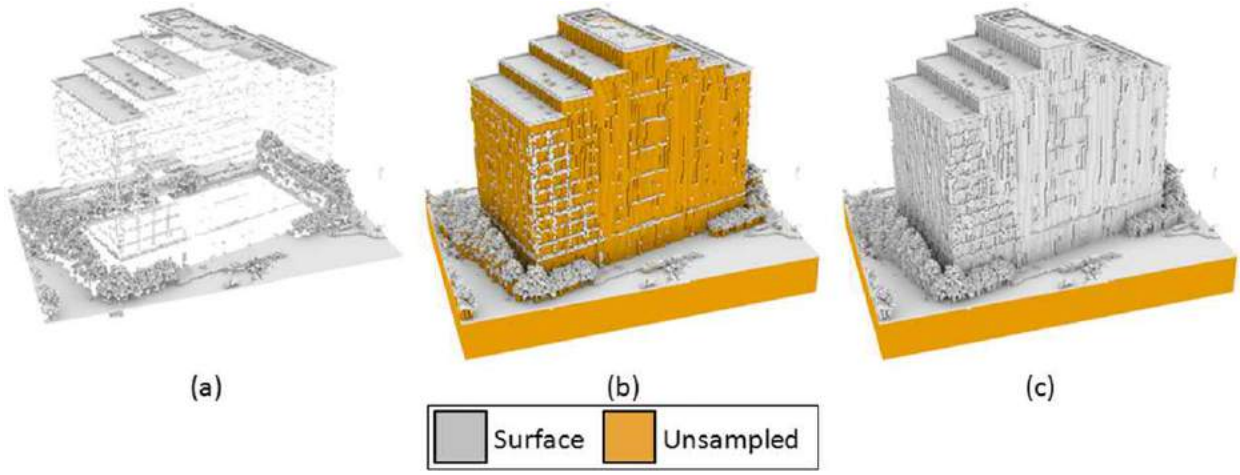


Fig. 3 Estimating 3D completeness of a dataset. The true surface of the model cannot be well-estimated from the sparse lidar points in (a). The orange unsampled areas in (b), identified using lidar path tracing, estimate the positions of missing surfaces. Given that a complete model should appear as (c), the completeness can be estimated by taking the ratio of the measured surface area to the estimated total surface area.

Several research groups have approached the problem of evaluating density by tracing the path of laser pulses through the dataset volume. The line segment path between the system and each point in the cloud point is computed using the coincident lidar instrument position records. Hagstrom (2014) defines the sampling density for a given voxel as the number of laser pulses intersecting that voxel divided by the voxel's geometric volume, regardless of whether the voxel contains any points itself. This is similar to the ASPRS two-dimensional definitions, where the requirement of using only first returns is effectively counting pulses rather than points.

A byproduct of this volumetric path tracing is that every voxel can be classified into one of three states: full due to containing points, known-empty due to containing no points but having pulses pass through, or unsampled due to being unreachable by the lidar (**Fig. 3**). This type of classification and knowledge of the unsampled regions has been used to improve methods such as object detection by Yapo *et al.* (2008) and change detection by Haas (2006).

Hagstrom (2014) build on this volumetric analysis idea to estimate completeness of a 3D dataset. Making the assumption that opaque objects always consist of an unsampled interior and known-empty exterior, a completely sampled object would always have at least one layer of full voxels between these two sections. Interfaces between adjacent full and known-empty voxels indicate a known surface, while interfaces between unsampled and known-empty voxels can only appear where surface points are missing. Taking the ratio of the missing surface area to the total surface area gives an estimate of the unsampled surface fraction (USF). While not yet adopted as an official metric, their work shows that 3D completeness can be estimated and that volumetric lidar representations can assist with analysis.

Evaluating 3D density and completeness remains an open problem but one which is increasingly important as lidar becomes more common and datasets increase in density. Though not yet standardized, volume-based metrics are one solution to evaluating 3D dataset quality, and show promise in this regard.

Conclusions

Common standards for metric assessment of lidar data are required to ensure that data quality is sufficient to meet end user needs and to enable objective and quantitative evaluation and comparison of lidar systems and capabilities as well as emerging non-lidar sources of 3D data. As lidar system designs continue to evolve and as lidar data continues to be used in new ways, community standards are also evolving to ensure that metrics continue to accurately characterize data quality. This article has reviewed current industry standards for metric evaluation of lidar data and discussed a few important emerging challenges and preliminary efforts toward addressing them.

See also: Multi-Dimensional Laser Radars

References

Bosch, M., Kurtz, Z., Hagstrom, S., Brown, M., 2016. A multiple view stereo benchmark for satellite imagery. In: Proceedings of the IEEE Applied Imagery Pattern Recognition Workshop.

Burton, R., 2010. Foliage penetration obscuration probability density function analysis from overhead canopy photos for gimbaled linear-mode and geiger-mode airborne lidar. In: Proceedings of the SPIE Laser Radar Technology and Applications XV.
Haas, G., 2006. Three-dimensional change detection with the use of an evidence grid. ARL-TR-3916, Army Research Laboratory.
Habib, A., Kersting, A.P., Bang, K.I., Lee, D.-C., 2010. Alternative methodologies for the internal quality control of parallel LiDAR strips. IEEE Transactions on Geoscience and Remote Sensing 48 (1), 221–236.
Hagstrom, S., 2014. Voxel-based LIDAR analysis and applications. PhD Thesis, Rochester Institute of Technology.
Henriksson, M., 2005. Detection probabilities for photon-counting avalanche photodiodes applied to a laser radar system. Applied Optics 44 (24), 5140–5147.
Lidar Specifications for Flood Hazard Mapping. 2009. Appendix 4B: Airborne Light Detection and Ranging Systems, Federal Emergency Management Agency.
Popescu, S.C., Zhao, K., 2008. A voxel-based lidar method for estimating crown base height for deciduous and pine trees. Remote Sensing of Environment 112 (3), 767–781.
Pyysalo, U., Sarjakoski, T., 2008. Voxel approach to landscape modelling. International Archives of the Photogrammetry, Remote Sensing, and Spatial Information Sciences. 37. Part B4.
Rodarmel, C., Lee, M., Gilbert, J., *et al*., 2015. The universal lidar error model. Photogrammetric Engineering & Remote Sensing 81 (7), 543–556.
Roth, M., Hunnell, J., Murphy, K., Scheck, A., 2007. High-resolution foliage penetration with gimbaled lidar. In: Proceedings of the SPIE Laser Radar Technology and Applications XII.
Sampath, A., Heidemann, H.K., Stensaas, G.L., Christopherson, J.B., 2014. ASPRS research on quantifying the geometric quality of lidar data. Photogrammetric Engineering and Remote Sensing 80 (3), 201–205.
Stevens, J.R., Lopez, N.A., Burton, R.R., 2011. Quantitative data quality metrics for 3D laser radar systems. In: Proceedings of the SPIE Laser Radar Technology and Applications XVI.
Vaidyanathan, M., Blask, S., Higgins, T., *et al*., 2007. Jigsaw phase III: A miniaturized airborne 3-D imaging laser radar with photon-counting sensitivity for foliage penetration. In: Proceedings of the SPIE 6550, Laser Radar Technology and Applications XII.
Wulder, M.A., White, J.C., Nelson, R.F., *et al*., 2012. Lidar sampling for large-area forest characterization: A review. Remote Sensing of Environment 121.
Yapo, T.C., Stewart, C.V., Radke, R.J., 2008. A probabilistic representation of LiDAR range data for efficient 3D object detection. In: Proceedings of the IEEE Computer Vision and Pattern Recognition Workshops.
Zhu, X.X., Bamler, R., 2012. Demonstration of super-resolution for tomographic SAR imaging in urban environment. IEEE Transactions on Geoscience and Remote Sensing 50 (8), 3150–3157.

Further Reading

ASPRS Positional Accuracy Standards for Digital Geospatial Data, 2015. The American Society for Photogrammetry & Remote Sensing, vol. 81, no. 3, pp. A1–A26.
Baltsavias, E.P., 1999. Airborne laser scanning: Basic relations and formulas. ISPRS Journal of Photogrammetry & Remote Sensing 54, 199–214.
Lidar Base Specification, 2014. Version 1.2, National Geospatial Program, USGS.
Salvaggio, K., Salvaggio, C., Hagstrom, S., 2014. A voxel based approach for imaging voids in three-dimensional point clouds. In: Proceedings of the SPIE Geospatial InfoFusion and Video Analytics IV; and Motion Imagery for ISR and Situational Awareness II.

Multiple Input, Multiple Output, MIMO, Active Electro-Optical Sensing

Paul F McManamon, Exciting Technology LLC, Dayton, OH, United States
Jeffrey R Kraczek, University of Dayton: Electro-Optics, Dayton, OH, United States

Introduction

Weight and space constraints are frequently placed on active EO systems, including lidar systems, but high resolution imaging in cross range requires large optics due to diffraction. Large optics are frequently impractical to use because an imaging system that uses them has to be deeper than a small optical imager to maintain the same f#. This contributes to large monolithic optical systems being heavy and expensive. In order to achieve high resolution at lighter weight and smaller size, methods have been developed to synthesize large apertures either by motion of a single smaller aperture or by an array of smaller apertures (Krause *et al.*, 2011; McManamon and Thompson, 2003).

Using multiple transmitter apertures and multiple receive apertures allows design flexibility not present in monolithic apertures. Synthetic aperture radar is comprised of a single moving aperture, so the transmit and receive apertures both move (Skolnik, 1980, 1990; Soumekh, 1999; Richards, 2005). A synthetic aperture lidar, SAL, uses motion of the single aperture to synthesize a larger effective aperture. For both SAR and SAL the synthetic aperture is almost twice as large as the actual flown distance, due to the angle of incidence being equal to the angle of reflection. This has been experimentally demonstrated over decades in the microwave region, and recently in the optical regime, when SAL, was demonstrated (Krause *et al.*, 2011; Beck *et al.*, 2005). Duncan provides a cross range resolution equation for spotlight-mode SAL:

$$\delta = \frac{R\lambda}{2L + D} \tag{1}$$

where D is the size of the aperture, R is the distance to the object, λ is the wavelength, and L is the distance moved (Duncan and Dierking, 2009). As a result, the effective size of a synthetic aperture based on motion from a diffraction point of view is:

$$D_{eff} = 2 * L + D_{real} \tag{2}$$

For RF systems the size of the real aperture is neglected, making the effective aperture twice as large as a monolithic aperture of width equal to the distance flown. The real aperture in RF systems is often on the order of a meter compared to the distance moved, which is multiple kilometers. In an EO system, the real aperture can be a significant fraction of the flight distance, so the D_{real} term is retained. For example, a SAL might have a real beam size of 15 cm and a flight distance of 1 m.

In order to synthesize a large aperture the returned field, including both amplitude and phase, must be measured or estimated. Optical carrier frequencies are much higher than can be measured directly by detectors. For example, a 1.5 μm wavelength has a carrier frequency of 200 Thz, orders of magnitude higher than even the highest bandwidth detectors can measure. In order to measure phase as well as amplitude we will need to use a local oscillator, LO, to beat against the return signal. We can choose to have a local oscillator that aligns with the return signal, but uses a different frequency than the transmitter. The two return signals beat against each other to create an intermediate frequency, IF, return. If the LO and the transmitted signal are stable, then the IF return will have the same phase difference as the carrier return. This is called temporal heterodyne, and allows us to measure phase. Alternately we can use spatial heterodyne, sometimes called by the broader term of digital holography, to measure the return phase of the signal. For spatial heterodyne we use an LO that is the same frequency as the transmitter signal, but is offset in angle from the return. This creates a spatial beat, generating fringes across the receive plane, which allows us to measure spatial variations in phase. Both temporal and spatial heterodyne can use either pupil plane or image plane imaging to measure the return fields. When you measure the field, it is possible to convert from image to pupil plane and back, as well as to digitally adjust focus or correct for aberrations.

Conceptual Discussion of Multiple Input, Multiple Output Lidar

Instead of using motion to synthesize a larger effective aperture, such as done in Synthetic Aperture Lidar (SAL), Multiple Input, Multiple Output (MIMO) active EO sensing uses multiple physical sub-apertures to synthesize a larger effective aperture. An array of receive-only sub-apertures can synthesize an effective aperture as large as the receive array if the field across the array can be measured, or estimated. With an array of transmit and receive sub-apertures, even more flexibility is obtained, so long as it is possible on receive to identify which transmitter each photon initially came from. MIMO can be used in both dimensions, or in only one dimension. It can be combined with SAL, with SAL synthesizing a larger pupil plane aperture in one dimension and MIMO in the other dimension.

One effect from multiple transmit and receive sub-apertures is increased angular resolution. This is similar to the angular resolution increase from motion based synthetic aperture sensors. Instead of motion, an array of n sub-apertures that both transmit and receive can be used. Alternately we can have separate arrays for transmitting and for receiving. For 9 sub-apertures in a row, with both transmit and receive from each aperture, the array will have a diffraction limit consistent with a monolithic

aperture that is 1.89 times as large in diameter as the array. This is because 8 sub-apertures are equivalent to the distance moved, L, in Eq. (1) while 1 sub-aperture is equivalent to the real aperture, D_{real}. If transmission occurs from one sub-aperture in the middle of an array, the receive aperture array is effectively in its normal location. If the transmit beam is up one sub-aperture, it is as though the receive aperture were moved down one sub-aperture. If all of the sub-apertures transmit the result is something like what is shown in **Fig. 1**, where the lighter color linear arrays indicate the perceived location of the linear receive array, depending on which transmit sub-aperture photons are emitted from. The dark color column shows the actual location of the arrays. The lighter columns show the coordinate transformed location of the receive array, based on which aperture is emitting photons. We need to do a coordinate transformation as though all photons are emitted from a single sub-aperture before we can add the fields. The full extent of the linear array is 1.89 times larger, but it is sampled more near the middle of the effective aperture, represented by 9 samples in the middle down to one on either end. This is similar to an apodized aperture.

We could use fewer transmit apertures rather than using them all, and still have the same maximum resolution, but it would reduce the intensity at certain spatial frequencies. The apodizing effect would not be as significant in that case.

An alternate view of the increased resolution is provided in **Fig. 2**.

The main thrust of Tyler (2012) is to argue that we have multiple ways to obtain the same optical distances because of the multiple receive and transmit apertures. This allows us to solve for the difference in path lengths between the sub-apertures, and to adjust out that optical path difference. This can be done closed form, without iteration.

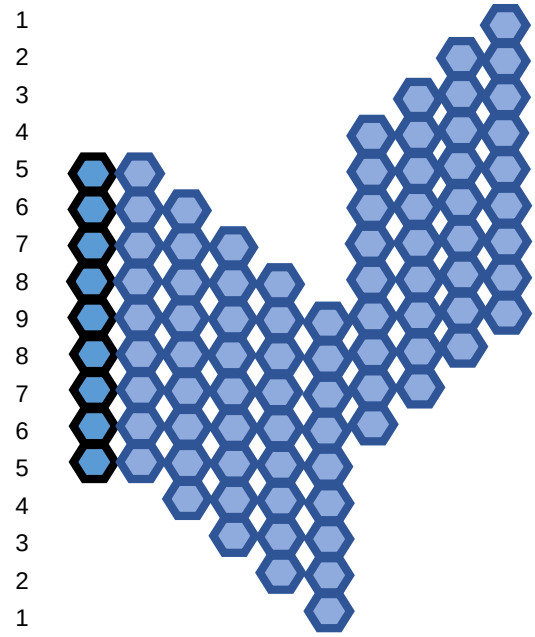


Fig. 1 Effective receiver aperture placement based on transmit sub-aperture utilized. Dark color shows actual location of the arrays. Light color shows perceived location. The number on the left shows how many received sub-apertures are perceived to be at that location.

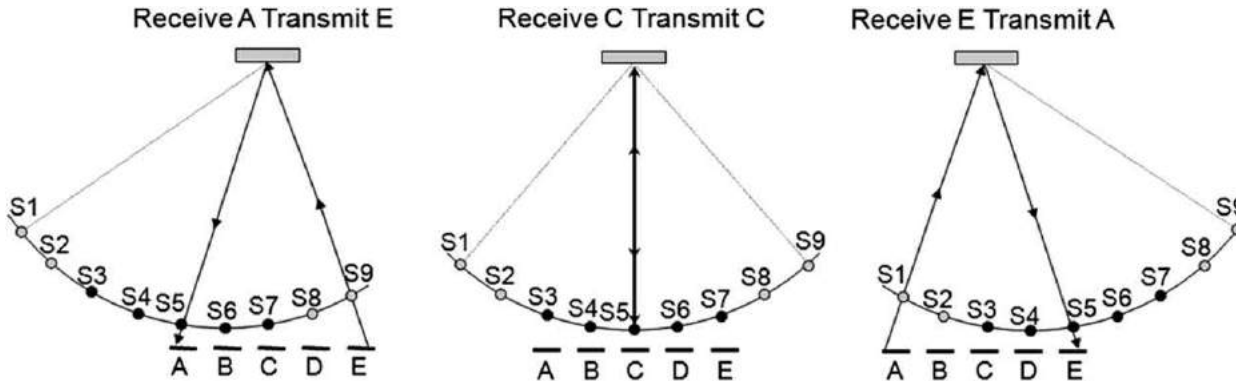


Fig. 2 Speckle measurements create a synthetic aperture that is almost twice as large as the array. The diagrams in this figure illustrate the nature of the speckle return from the object. Taken from Tyler, G.A., 2012. Accommodation of speckle in object-based phasing. Journal of the Optical Society of America A 29 (5), 722–733

Closed Form MIMO

Rabb *et al.* (2011) only use one receive aperture, but multiple transmit apertures. If transmitters are spaced at distances less than the size of a receive sub-aperture diameter, it provides sampling that can allow a closed form solution to differences in atmospheric path across the sub-aperture array (Fienup, 2000). An example of the resulting effective receiver pupil using transmitter diversity is shown in **Fig. 3**. Because of the overlap, one can solve in a closed-form manner for the phase between transmit sub-apertures, allowing its use to compensate for atmospheric phase disturbances in the pupil plane.

Rabb showed that he could calculate phase differences between transmit apertures by using multiple closely spaced transmit apertures with a single receive aperture, by using a linear systems of equations in conjunction with three transmit apertures. In this method, they used bivariate expansion to model intra-aperture phase aberrations. Similarly, Gunturk showed closed form phasing to remove intra-aperture aberrations for a single receive aperture using three transmit apertures but he used Zernike polynomials instead of a bivariate expansion (Gunturk *et al.*, 2012). The wave front difference was used to calculate the coefficients of up to seventh order Zernike polynomials to correct for the phase errors. Both of these methods deal with intra-aperture aberrations, but do not address the need for image phasing in a multi-receive aperture imaging system.

Kraczek *et al.* (2016) showed intra and inter-aperture phase aberrations can be solved in a closed form manner similar to Rabb and Gunturk. The intra-aperture aberrations are removed from the field first using an individual receive aperture and an array of three transmit apertures. Krazcek also considered a fourth transmitter widely separated from the three transmitter cluster. This will greatly extend the size of the synthetic aperture and will increase the frequency response cutoff of the system by approximately double, because the fourth transmitter aperture is placed about the width of the aperture array away from the set of 3 transmit apertures. Krazcek analytically calculated the MTF of this MIMO array, and then "experimentally" verified it in simulation using the slant edge MTF approach (Reichenbach *et al.*, 1991).

To perform inter-aperture closed form phasing, intra-aperture phase errors must be removed. The method used by Krazcek closely follows Gunturk's work, by using three transmit apertures to remove the inter-aperture phase errors.

The left side of **Fig. 4** shows an example of a multi-transmitter, single receive aperture system. The receive aperture is light green and the three transmitter cluster is magenta. The cluster is set to the vertices of an equilateral triangle with side lengths equal to the radius of the receive apertures. The right side of 4 shows the measured received fields, referenced to a common coordinate system. This is similar to **Fig. 3**, taken from Rabb. The overlap regions of the received fields in the pupil plane due to the different transmitter positions are shown in yellow and red. The outer edge of this image is the shape of the synthetic pupil. The synthetic pupil results from all the different fields being referenced to a common coordinate system. This larger synthetic pupil gives rise to a high resolution image. As described in Rabb *et al.* (2010), a shift in the transmitter location has an equal result on the movement of the placement of the received field in the synthetic pupil (Rabb *et al.*, 2010). This necessitates a coordinate change to compensate for different transmitter locations before the pupil plane fields can be coherently added. The synthetic pupil is built from the individual pupil fields being registered into the correct location in post processing. Pupil fields are assembled by Krazcek using known transmitter position since this work was done in simulation.

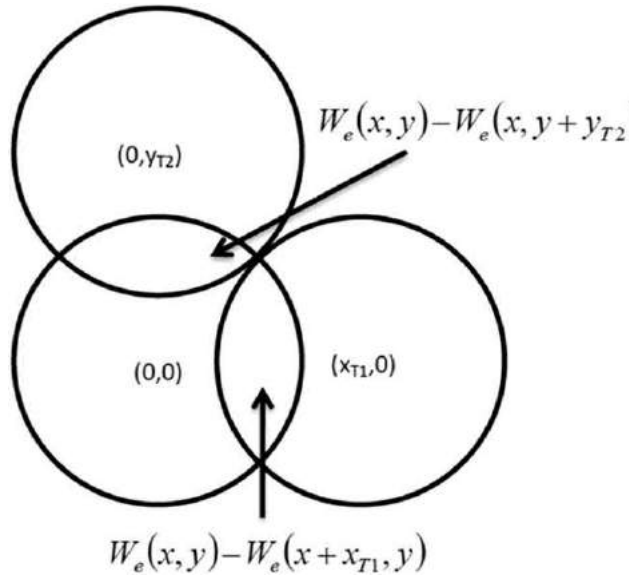


Fig. 3 Multiple sub-aperture overlap using 3 transmitters in a pattern about half a receive sub-aperture diameter apart. When images are adjusted to account for transmitter location, the receiver sub-aperture pupils overlap as shown. Reproduced from Rabb, D.J., Stafford,J.W., Jameson,D.F., 2011. Non-iterative aberration correction of a multiple transmitter system. Optics Express 19 (25), 25048–25056

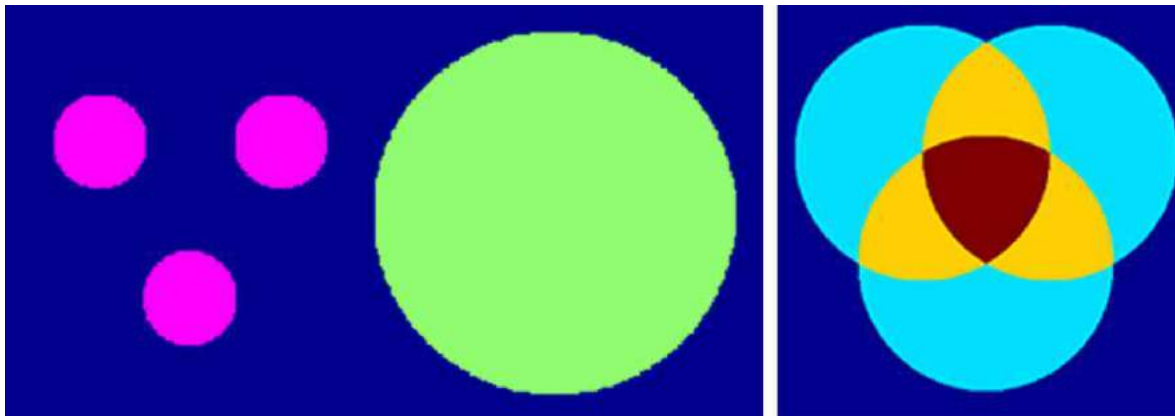


Fig. 4 Left: Single receive aperture, light green, with three transmitter apertures, magenta. Right: Field overlap due to a single receive aperture and three transmitter cluster.

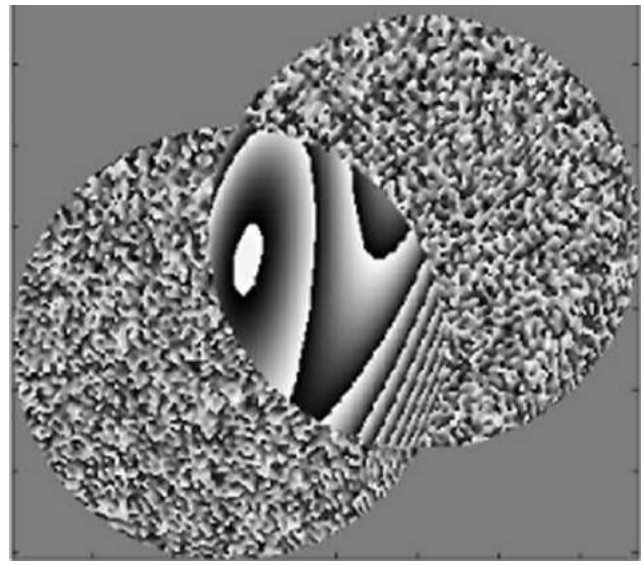


Fig. 5 Difference between the wave fronts of two overlapped pupil fields. The fringed section in the middle is where the two fields overlap.

The fields incident on the receive aperture can be described by

$$U_i(x,y) = P(x,y)\exp(j2\pi W_e(x,y))U_b(x-x_i, y-y_i),\ i=1,2 \tag{3}$$

where $P(x,y)$ is the pupil function of the receive aperture, $W_e(x,y)$ is the wave front error over the aperture, $U_b(x-x_i,y-y_i)$ is the received field with the subscript i referring to the ith transmitter. Defining $W_b(x,y)$ to be the wave front of $U_b(x,y)$, the received wave front is

$$W_i(x,y) = W_e(x,y) + W_b(x-x_i, y-y_i),\ i=1,2 \tag{4}$$

The object wave front is not well defined and it is desirable to calculate the wave front aberrations independent of the intended target. As it is necessary to register the different fields based on the field shifts caused by transmitter positions into a synthetic pupil plane, the numerical dependence on the object wave front can be removed. Taking the difference of the overlapping areas in the received wave fronts gives

$$\begin{aligned}\Delta W(x,y) &= W_1(x-x_1, y-y_1) - W_2(x-x_2, y-y_2) \qquad = (W_e(x+x_1, y+y_1) + W_b(x,y)) \\ &- (W_e(x+x_2, y+y_2) + W_b(x,y)) \qquad = W_e(x+x_1, y+y_1) - W_e(x+x_2, y+y_2)\end{aligned} \tag{5}$$

Due to the coordinate shift, the wave front difference is a difference between the wave front errors, which can be calculated and removed. Fig. 5 shows the difference of the wave front of two overlapped pupil plane fields. The fringed area is where the two field overlap. Notice that this also removes speckle within the region of overlap.

Using the wave front difference, Gunturk is able to remove the intra-aperture phase errors using a set of over-determined linear equations, solving for the Zernike polynomials describing the phase errors across a circular pupil. His method does not work for inter-aperture phase errors because the Zernike coefficients are not necessarily the same between two apertures. The zeroth and first order, piston, tip, and tilt, cannot be solved for in the method describe above because they do not show up in the wave front difference using the same receive aperture. This is because piston is a flat phase and there is no difference in that term across the aperture. Tip and tilt terms have constant slopes across the aperture and therefore also cancel out when taking the difference of the two wave fronts in the same aperture.

A different method must be developed to correct for the inter-aperture tip-tilt-piston aberrations. The top of 6 shows a three transmitter and five aperture array. This is an example of a MIMO system that can be used for inter aperture closed form phasing. The bottom of 6 shows the pupil plane overlap pattern for the system shown in the top 6. Each individual receive aperture provides coverage in the same pattern as shown in 4. There is significantly less overlap between each aperture, but it still exists. The receive apertures should be located as close together as is physically permitted to allow for greater inter-aperture overlap. This greater overlap is beneficial in the registration process. We will define the wave front difference as the wave front error across the overlapped apertures. Assuming that the intra aperture phase errors have been removed, the wave front difference is now defined as

$$\Delta W(x,y) = W_{e,i1,j1}(x,y) - W_{e,i2,j2}(x,y) \quad i1 \neq i2, j1 \neq j2 \tag{6}$$

where i1 and i2 are different transmit apertures and j1 and j2 are different receive apertures (**Fig. 6**).

The left side of 7 shows an example of the wave front difference between two receive apertures. The wave front difference shows only the relative difference between the phase errors on the two apertures. On the left is the wave front difference after intra-aperture corrections have been made and on the right is the same wave front difference after tip-tilt corrections have also been made, resulting in just piston remaining. The gradient can be used on the area between phase wraps to gain an estimate of the tip and tilt between the two apertures. This estimate can then be used to correct tip and tilt. At this point, all that is left is piston. The right side of **Fig. 7** shows an example of the wave front difference after all other phase errors are removed, resulting in just piston phase error being left. The value of the dark section is the piston phase. In a noise free environment both the piston and the tip and tilt corrections can be calculated with very few pixels. In a noisy system more pixels will need to be used. When correcting for these inter-aperture phase errors, one aperture is set as a reference to which all others are corrected. The corrections are daisy chained from the reference field to the field farthest away. As seen in 7, only two transmit apertures are required for inter-aperture phasing. Three transmit apertures are still used for intra-aperture phasing and to facilitate inter-aperture phasing with 2D receive aperture arrays.

Adding a fourth transmitter can increase the size of the synthetic pupil significantly, depending on the location of the fourth transmitter. The top of 8 shows a four transmitter, five receive aperture system. The bottom of 8 shows the shape of the synthetic pupil. The set of five circles on the left hand side comes from the fourth, widely separated, transmitter. Those fields do not need to be overlapped since they come from the same apertures that already had the intra and inter-aperture phase aberrations removed. The exception is that there needs to be overlap between at least one pupil field from the fourth transmitter with one of the pupil

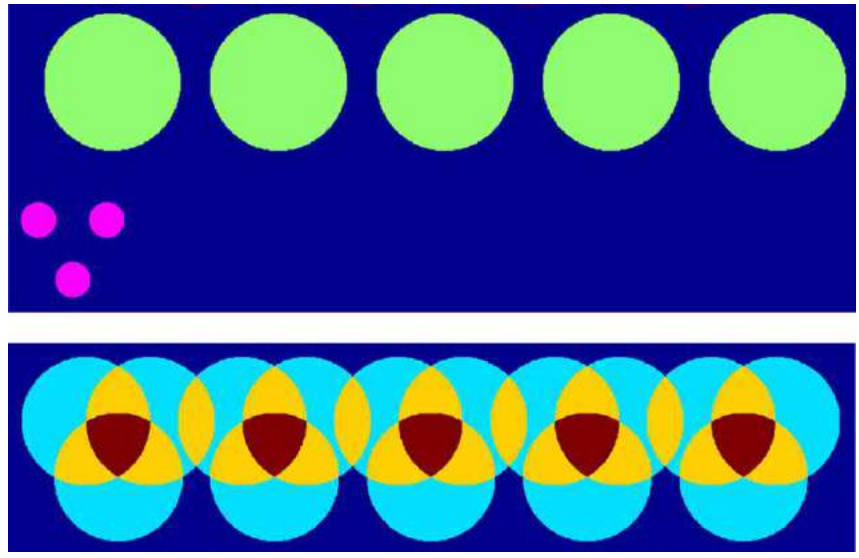


Fig. 6 Top: Example of a three transmitter and five aperture array. The red circles are unfilled receive apertures, the light green circles are filled receive apertures and the magenta circles are transmit apertures. Bottom: Example of the field overlaps from the three transmitter and five aperture array shown on top. The light blue areas are where only one field is present, the yellow are areas of two fields overlapping, and the red is where three fields overlap.

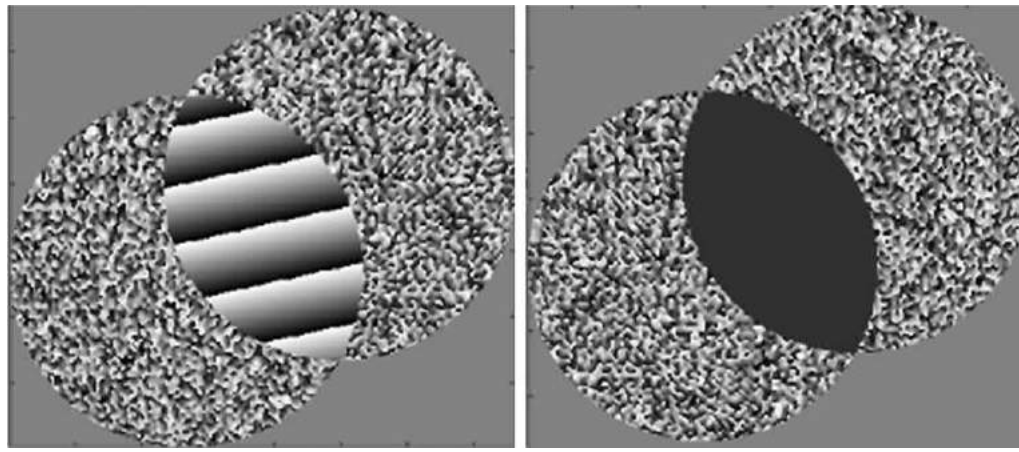


Fig. 7 Wave front difference between two pupils coming from two different apertures. Left: A Piston, tip, and tilt error remains. Right: Piston error remains.

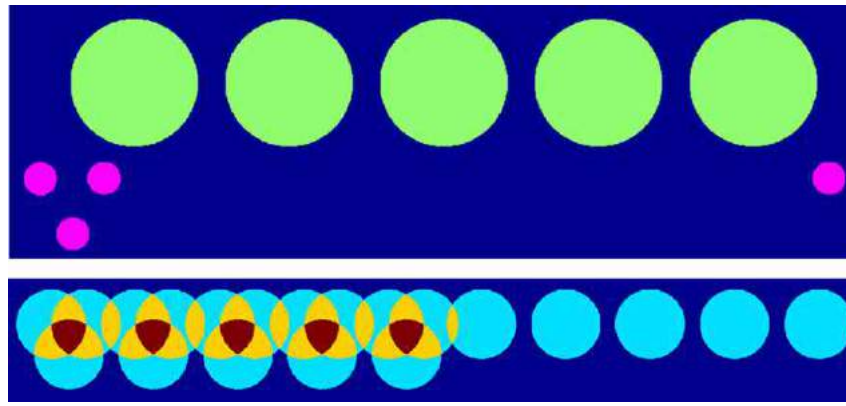


Fig. 8 Top: Example of a four transmitter and five aperture array. The red circles are unfilled receive apertures, the light green circles are filled receive apertures and the magenta circles are transmit apertures. Bottom: Example of the field overlaps from the four transmitter and five aperture array shown on top.

fields from the initial three transmitter cluster. In this example, the farthest left field from the fourth transmitter and the farthest right field from the transmitter cluster are overlapped. This overlap is used to obtain an additional piston correction for all pupil fields from the fourth transmitter and allow for field registration. Sub-pixel registration on the speckle in the pupil plane is used. It is assumed exact positions of all transmit apertures are known but overlap is maintained. An additional piston correction will be necessary to account for the different position of the fourth transmitter fields from where the initial inter-aperture measurements were made; movement of the pupil fields cause an additional piston term in the wave front difference, due to tip-tilt, that needs to be corrected (Fig. 8).

Fig. 9 shows the theoretical modulation transfer function (MTF) in the horizontal dimension for the synthetic aperture in 6 on the left and the synthetic aperture in 8 on the right. Notice that the cutoff frequency for the three transmitter and five aperture array is about half the cutoff of the array with four transmit apertures and five apertures. This is because the synthetic pupil size is about half as large in that dimension.

Trade Space of Current MIMO

MIMO techniques as described here can be implemented either using temporal or spatial heterodyne (digital holography) techniques. It will however be much easier to tag the emitted transmitter signals, allowing simultaneous transmission, if high bandwidth temporal heterodyne is used, since high bandwidth tagging schemes can then be used. RF MIMO techniques that use multiple simultaneous phase centers have been developed (Coutts *et al.*, 2006).

The angular resolution of an array of sub-apertures can be almost twice the resolution of the diffraction limit for a monolithic aperture. In addition, atmospheric turbulence between the object imaged and the sensor can be very quickly and accurately compensated because we do not need to use an iterative approach to finding the required optical path correction to compensate for the turbulence. This compensation is relatively straightforward for turbulence in the pupil plane, but will be more difficult for

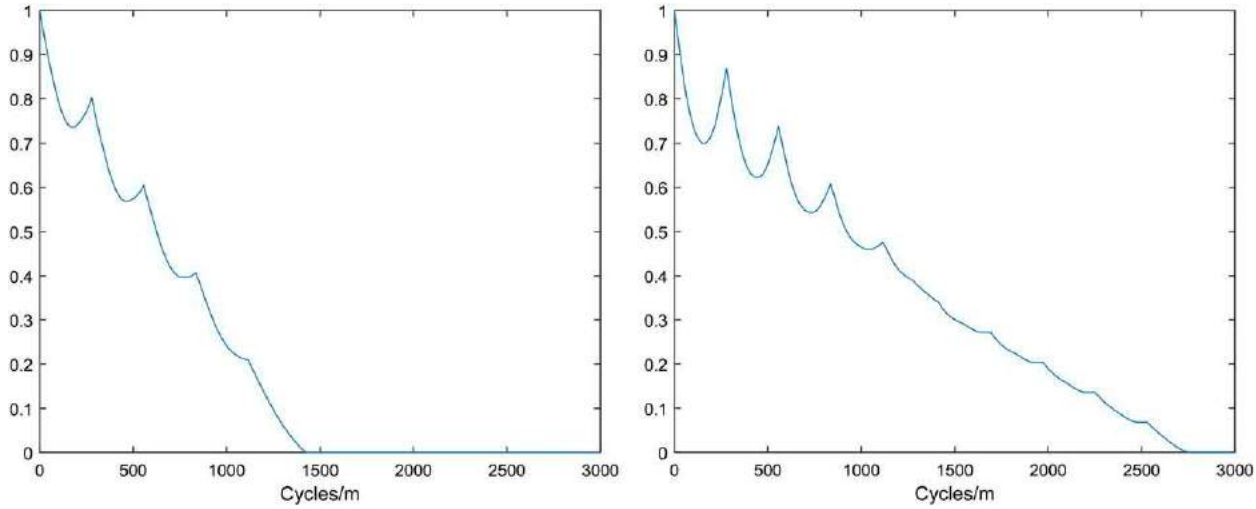


Fig. 9 MTF plots for the arrays shown in 5 and 7. The X-axis is in cycles per mm. Left: Horizontal MTF for array shown in 6. Right: Horizontal MTF for array shown in 7.

volume turbulence. A significant limitation is that the received signal is captured in a smaller receive aperture area. The required laser power will increase by the ratio of the area of the monolithic aperture to the received aperture array area. Also, if the angular resolution is greater than a monolithic aperture of the same size, the cross section of each pixel is lower because of the increased resolution. This results in either more required power, or lower signal to noise. Since each new transmit aperture adds power this can balance out.

Arrays of high temporal bandwidth detectors will be very helpful in implementing MIMO techniques for EO imaging. The papers cited sequenced through the multiple transmitters because they implemented MIMO using a digital holography/spatial heterodyne approach to imaging, using low bandwidth framing detector arrays. If high bandwidth detector arrays and a temporal heterodyne approach to imaging are used then it should be possible to simultaneously emit multiple tagged transmitter beams, and to have each receiver be able to distinguish the transmit aperture any photon came from. High-bandwidth detectors can allow high bandwidth modulations to be imposed on each transmitted beam, and sorted on receive. Temporal heterodyne arrays will need to be AC-coupled, or have high dynamic range, or have high sensitivity such that temporal heterodyne can be implemented using a relatively weak LO.

A second technical hurdle to overcome is volume turbulence. Techniques for calculating, and compensating for, volume turbulence still need to be developed.

MIMO technology will allow imaging with high angular resolution using a much lighter and more compact aperture arrays than a monolithic aperture. An array of small sub-apertures can be much thinner and lighter than a monolithic aperture. Also, a MIMO approach can achieve almost twice the diffraction limited angular resolution of a monolithic aperture.

To be really useful in freezing the atmosphere, MIMO should be implemented using high bandwidth detector arrays, which still need to be further developed. There will be a digital implementation requirement to conduct the required calculations. There will be a need to have many different optical trains, complicating the optical system. Also, narrow line lasers will need to be used to do either spatial or temporal heterodyne. Higher power lasers will be required to image a given area using a MIMO array compared with imaging with a monolithic aperture because the spatial frequency cutoff of the measured signal is being increased more rapidly than the area of the pupil plane receive aperture.

Conclusion

This technology is well suited for long range imaging applications from air or space. MIMO could be used in the cross range dimension along with motion-based synthetic aperture imaging.

In summary, MIMO approaches for active EO sensing can, at a minimum, increase the effective diameter of an aperture array by almost a factor of two, and can allow multiple sub-apertures on receive to be phased using a closed-loop calculation of the phase difference between sub-apertures. This can compensate for atmospheric turbulence at least at some locations between the sensor and the imaged object.

References

Beck, S.M., Buck, J.R., Buell, W.F., *et al.*, 2005. Synthetic-aperture imaging laser radar: Laboratory demonstration and signal processing. Applied Optics 44 (35), 7621–7629.

Coutts, S., Cuomo,K., Mcharg, J., Robey,F., Weikle,D., 2006. Distributed coherent aperture measurements for next generation BMD radar. In: Fourth IEEE Workshop on Sensor Array and Multichannel Processing, pp. 390–393.
Duncan, B.D., Dierking, M.P., 2009. Holographic aperture ladar. Applied Optics 48 (6), 1168–1177.
Fienup, J.R., 2000. Phase error correction for synthetic-aperture phased-array imaging systems. In: Proceedings of SPIE 4123, Image Reconstruction From Incomplete Data 4123, pp. 47–55.
Gunturk, B.K., Rabb, D.J., Jameson, D.F., 2012. Multi-transmitter aperture synthesis with Zernike based aberration correction abstract. Optical Express 20 (24), 5179–5186.
Kraczek, J.R., Mcmanamon, P.F., Watson, E.A., 2016. High resolution non-iterative aperture synthesis. Optical Express 24 (6), 6229–6239.
Krause, B.W., Buck, J., Ryan, C., *et al*., 2011. Synthetic aperture ladar flight demonstration. In: CLEO: Applications and Technology 2011: Laser Applications to Photonic Applications.
McManamon, P.F., Thompson, W., 2003. Phased array of phased arrays (PAPA) laser systems architecture. Fiber and Integrated Optics Journal 22 (2), 79–88.
Rabb, D.J., Jameson, D.F., Stafford, J.W., Stokes, A.J., 2010. Multi-transmitter aperture synthesis. Optical Express 18 (24), 24937–24945.
Rabb, D.J., Stafford, J.W., Jameson, D.F., 2011. Non-iterative aberration correction of a multiple transmitter system. Optical Express 19 (25), 25048–25056.
Reichenbach, S.E., Park, S.K., Narayanswamy, R., 1991. Characterizing digital image acquisition devices. Optical Engineering 30 (2), 170–177.
Richards, M.A., 2005. Fundamentals of Radar Signal Processing: Chapter *. New York: McGraw-Hill.
Skolnik, M.I., 1980. Introduction to Radar Systems: Chapter 14, second ed. New York, NY: McGraw-Hill.
Skolnik, M.I., 1990. Radar Handbook, Chapter 17, by Roger Sullivan, second ed. New York, NY: McGraw-Hill.
Soumekh, M., 1999. Synthetic Aperture Radar Signal Processing With Matlab Algorithms. New York, NY: Wiley.
Tyler, G.A., 2012. Accommodation of speckle in object-based phasing. Journal of the Optical Society of America A 29 (5), 722–733.

InGaAs Linear-Mode Avalanche Photodiodes

Andrew S Huntington, Voxtel, Inc., Beaverton, OR, United States

Nomenclature

a Random variable for the number of primary photoelectrons injected into a multiplier
a_{bg} Random variable for the number of primary background photoelectrons
a_{dark} Random variable for the number of primary dark current electrons
a_{signal} Random variable for the number of primary signal photoelectrons
BW A modulation frequency bandwidth in Hz
F Excess noise factor by which APD output variance exceeds the deterministic gain case
I_{bg} Background photocurrent measured at the APD's terminals
I_{dark} Dark current measured at the APD's terminals
$I_{primary}$ Primary current injected into a multiplier
I_{signal} Signal photocurrent measured at the APD's terminals
k Impact ionization rate ratio for electrons and holes (slower rate over faster rate)
M Mean avalanche gain factor
m Random variable for the per-electron gain factor
n Random variable for the total number of electrons output by a multiplier
N_{CTIA} Input-referred charge noise of an amplifier in units of electrons
n_e Count of electrons within the depletion region of a junction
NEP Photoreceiver noise expressed as an equivalent input optical power level
NEPh Photoreceiver noise expressed as an equivalent number of input photons
n_h Count of holes within the depletion region of a junction
N_Q Photoreceiver noise expressed as an equivalent number of electrons at the APD output
n_{th} Detection threshold expressed as an equivalent number of electrons at the APD output
P_{RX} Photoreceiver output distribution scaled to electron count at APD output
q Elementary charge
QE Quantum efficiency
Q_{signal} Output signal charge in units of electrons
R Responsivity
$S_{I\ TIA}$ Input-referred noise current spectral intensity of an amplifier in units of $A^2\ Hz^{-1}$
$S_{I\ total}$ Total noise current spectral intensity of a photoreceiver
S_I Noise current spectral intensity
t Time
Γ Euler gamma function
v_{se} Electron saturation drift velocity
v_{sh} Hole saturation drift velocity
w Junction depletion width
λ Optical signal wavelength

Introduction

Linear-mode (LM) avalanche photodiodes (APDs) are used to preamplify signal photocurrent in photoreceivers where amplifier circuit noise is the dominant noise component. Since amplifier noise is a fixed quantity that is independent of the APD's gain, the photoreceiver's signal initially increases faster as a function of APD gain than its total noise, improving measures of sensitivity such as the signal-to-noise ratio (SNR). However, LM APDs amplify the shot noise associated with charge carrier generation in the diode junction and also generate excess shot noise associated with the stochastic gain process itself. Since an APD's shot noise increases faster with gain than the amplified signal photocurrent, optimal photoreceiver sensitivity occurs at some finite gain that is determined by factors such as the intensity of the amplifier noise, the magnitude of the APD's dark current, the intensity of the optical signal, and the statistics of the APD's gain process. The interplay of photoreceiver noise components leading to optimal sensitivity at finite gain is illustrated in Fig. 1.

InGaAs LM APD Structure and Manufacturing

Several binary and ternary III–V compound semiconductor alloys are plotted in Fig. 2 according to their crystal lattice parameters and their bandgaps at room temperature. In the diagram, the binary alloys are plotted as points, connected by arcs representing the range of ternary alloy compositions which result from blending pairs of binary alloys. $In_{0.53}Ga_{0.47}As$ is the alloy composition aligned vertically with the triangle representing InP, which means they share the same lattice parameter. $In_{0.53}Ga_{0.47}As$ is favored for use in short-wavelength infrared (SWIR) detectors because its direct and relatively narrow ($\sim$0.74 eV) bandgap results in efficient light absorption out to a cutoff wavelength near 1.7 μm and because comparatively thick ($\sim$μm) single-crystal light absorption layers can be grown with very low defect density, lattice-matched to InP substrates. Although non-lattice-matched InGaAs alloy compositions can be grown on InP substrates to push spectral response further into the SWIR, the use of non-lattice-

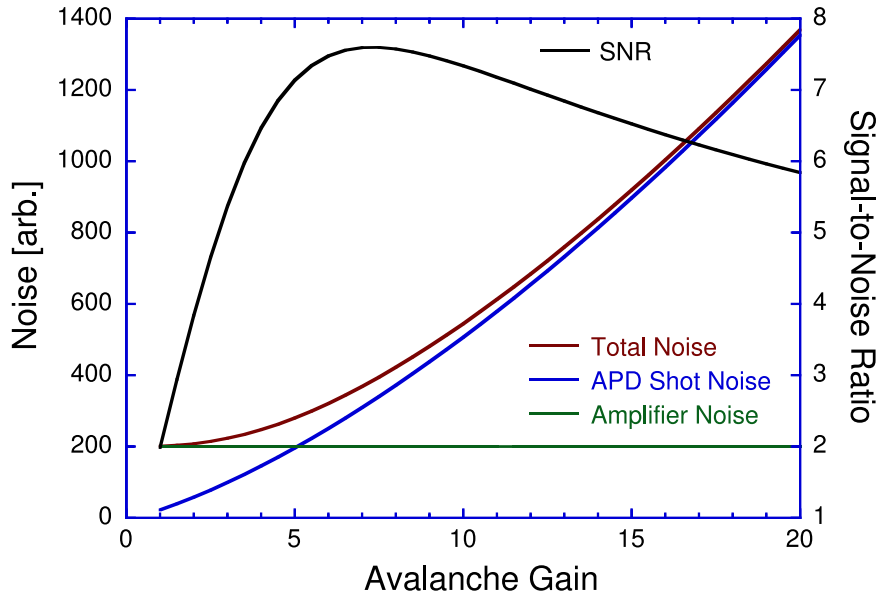


Fig. 1 Total noise, noise components, and signal-to-noise ratio of an avalanche photodiode (APD) photoreceiver vs. mean avalanche gain.

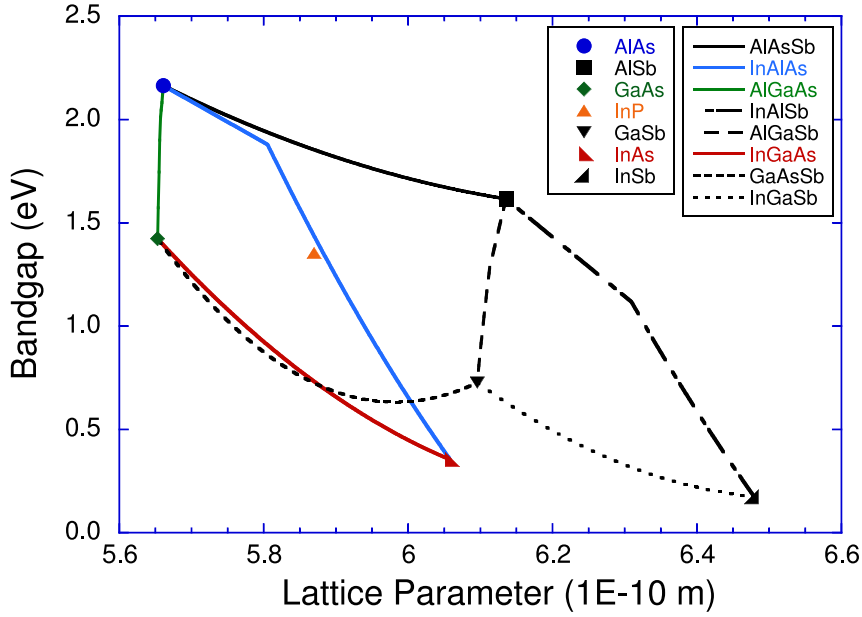


Fig. 2 Room temperature band gap vs. crystal lattice parameter for technologically important III–V compound semiconductor alloys.

matched alloys in APDs is not common, owing to the sensitive dependence of APD dark current on defect density. Dark current considerations also constrain selection of the alloys used for avalanche multiplication layers in InGaAs APDs to either InP itself or the lattice-matched ternary $In_{0.52}Al_{0.48}As$. Wider-bandgap alloys like InP or InAlAs are preferred for use in APD multiplication layers because their use minimizes dark current generation by band-to-band tunneling.

The single-crystal thin films from which InGaAs APDs are fabricated are grown on (100)-oriented or (100)-vicinal InP substrates by molecular beam epitaxy (MBE) or metal-organic chemical vapor deposition (MOCVD). Most InGaAs APDs share a common epitaxial structure in which an InGaAs light absorption layer and a charge carrier multiplication layer are separated by a space charge layer. Also called a field control layer, the charge layer assures that during operation, the electric field in the narrower-bandgap absorber will remain weak enough to limit dark current generation by tunneling when the electric field in the wider-bandgap multiplier is strong enough to generate a useful amount of avalanche gain. The charge layer of an InGaAs APD is the most critical to its function, and calibration of the charge layer's thickness and active doping concentration is the most difficult aspect of InGaAs APD manufacturing. Too many active dopant ions between absorber and multiplier result in an APD that reaches

avalanche breakdown at a reverse bias lower than that required to render it sensitive to optical signals, whereas too few result in excessive dark current generation by tunneling at reverse bias lower than that required to achieve useful avalanche gain. Common methods of doping calibration used in crystal growth labs, such as Hall effect measurements or electrochemical capacitance–voltage (ECV) profiling, typically do not have sufficient precision to calibrate InGaAs APD charge layers. Instead, prior to growth of production wafers, charge layer calibration is commonly accomplished by growth of complete APD layer stacks followed by quick-turn fabrication and current–voltage (I–V) characterization of test structures.

The separate absorption, charge, and multiplication (SACM) layer stack is sandwiched between anode and cathode layers. It is common to grow InGaAs APDs with the cathode side down, and often n-type InP substrates are selected so that electrical contact to the APD's cathode can be made through the wafer substrate. The ordering of the SACM layer stack between anode and cathode is determined by which alloy is selected for the multiplication layer. Because the absorption layer is physically separate from the multiplication layer, and because electrons drift toward the cathode and holes drift toward the anode of a reverse-biased diode, only one carrier polarity can be injected into the multiplier. The impact ionization rate of holes is higher in InP than that of electrons, but lower than that of electrons in InAlAs. Consequently, InP multipliers are grown on the anode side of the absorber, whereas InAlAs multipliers are grown on the cathode side of the absorber. The two common SACM stack orderings are illustrated in Fig. 3.

Detector elements are formed from the epitaxial thin film either by patterned diffusion to define the active area of each element, or by etching away the film surrounding each element. In the most common implementation of the diffused junction design, the anode layer of the APD thin film is not doped during crystal growth. Instead, the p-type dopant species – typically zinc – is introduced during wafer fabrication by diffusion through openings patterned in a dielectric mask layer that is deposited over the thin film. The maximum electric field strength inside the junction of a diode formed by patterned diffusion tends to occur at the edge of the diffusion because charge neutrality requires that the donor ions on the cathode side of the junction be balanced by an equal number of acceptor ions on the anode side. The upward curvature of the depletion region at the edge of the anode diffusion results in ionized acceptors facing ionized donors along a narrower frontage, concentrating the electric field (Fig. 4, left). Shallower anode diffusions result in tighter radii of curvature with higher maximum electric field strength. Multilevel diffusions or diffused guard rings surrounding the central anode can be used to minimize crowding of electric field lines at the perimeter of a

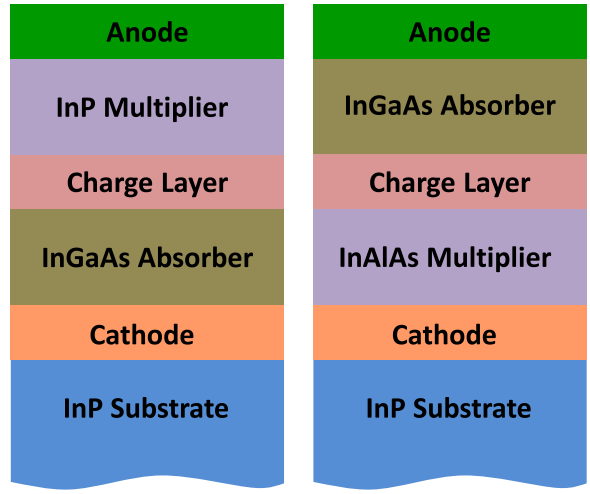


Fig. 3 Typical layer orderings for InP- and InAlAs-multiplier InGaAs avalanche photodiodes (APDs).

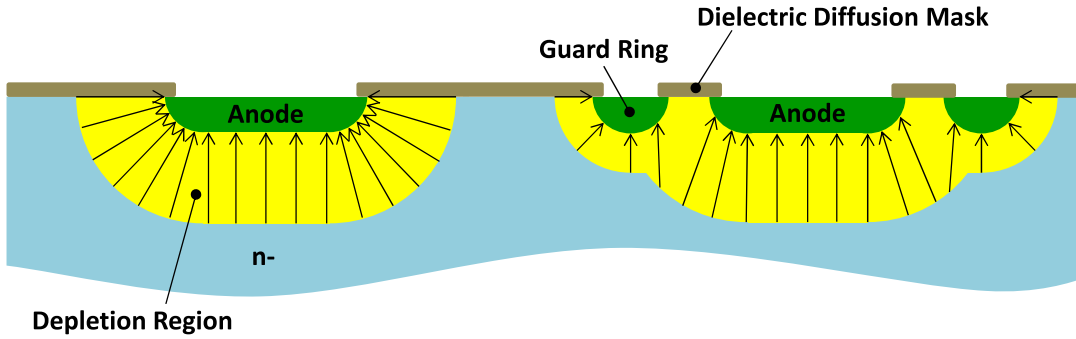


Fig. 4 Sketch of electric field lines in diffused junction photodiodes with (right) and without (left) guard rings.

diffused anode by reshaping the depletion region along the edge of the detector element, preventing edge breakdown (Fig. 4, right).

APDs fabricated by etching are made from thin films in which the top contact layer (typically the anode) is doped during crystal growth. Instead of defining the junction area by patterning the anode doping, the thin film outside the detector element area is etched away down to the substrate, leaving behind a mesa. Whereas the principal concern with diffused junction APDs is managing the electric field intensity at the perimeter of the patterned diffusion, it is sufficient that the sidewalls of an etched mesa APD be smooth, and slope gradually inward from the base of the mesa. Physical irregularities on the mesa sidewall – and especially notches where the slope of the mesa sidewall reverses, creating an overhang – can locally enhance the electric field and create a weak point at the junction perimeter. However, in general, the chemical cleanliness of the mesa sidewall is the principal manufacturing challenge for mesa APDs.

Most of an APD's dark current is generated in its InGaAs absorption layer because, having the narrowest bandgap in the structure, it presents the lowest energetic barrier to generation of charge carrier pairs, and because it is depleted during operation of the APD. In a diffused junction APD the InGaAs layer is safely buried under a wider-bandgap material like InP, and the depletion region intersects the wafer surface along a narrow band of wider-bandgap material at the anode perimeter. In contrast, a swath of depleted narrow-bandgap InGaAs is exposed along the sidewall of an etched mesa APD. Interruption of a crystal lattice by a surface creates a high density of mid-bandgap trap states. When the surface is of a narrow-bandgap semiconductor alloy and lies within the depletion region of a photodiode, a high dark current generation rate results. Chemical formation of materials with higher conductivity than the depleted diode junction is another problem at the mesa sidewall, as this can form a conductive path that shunts the diode. Both issues are addressed during mesa APD manufacturing by preparing mesa sidewalls that are as chemically clean – and devoid of sub-surface damage to the lattice – as possible, and then encapsulating the clean sidewall surface under an impermeable dielectric coating to protect it from environmental degradation. In some cases, chemical passivation treatments – often involving sulfide compounds – are applied to the mesa sidewall prior to encapsulation. Chemical passivation works by forming bonds at the semiconductor surface, the molecular orbitals of which lie outside the semiconductor bandgap, such that the resulting energy states do not behave as traps. The passivation treatment replaces trap states with inactive states and prevents traps from forming by tying up potentially reactive surface sites.

Fabrication of APDs also includes deposition of anode and cathode metal contact pads, and usually deposition of an anti-reflection coating. Different APD shapes and configurations are possible. The most common APD configuration is designed for illumination normal to the wafer surface, with light absorption occurring in a single pass through a comparatively thick ($\sim\mu$m) absorption layer. APDs of this class are commonly fabricated with active area diameters ranging from about 25 to 500 μm. Larger InGaAs APDs are seldom fabricated because detector capacitance, dark current, and manufacturing yield loss all scale linearly with junction area. However, hemispherical immersion lenses are sometimes affixed to APDs to increase their effective light-gathering area. APDs with thick single-pass absorbers may be top- or bottom-illuminated, depending upon whether the optical signal is delivered from the side of the wafer on which the epitaxial thin film was grown, or through the InP substrate. Top-illuminated InGaAs APDs tend to have superior responsivity at wavelengths shorter than the InP absorption band edge at about 950 nm, although the short-wavelength response of bottom-illuminated InGaAs APDs can be enhanced by physically removing the InP substrate. Bottom-illuminated APDs can be bump-bonded directly to application-specific integrated circuits (ASICs) to achieve lower interconnect parasitics than is possible with front-illuminated APDs of equal active area, and the bottom-illuminated configuration is always used for dense two-dimensional imaging arrays of APD pixels. A cross section through an example 25-μm diameter, bottom-illuminated, etched mesa APD pixel is sketched in Fig. 5 to illustrate the major structural parts of such a device. In Fig. 5 a ring-shaped cathode contact around the base of the APD mesa is connected to a pad on top of an adjacent inactive mesa for easy of bump-bonding of both anode and cathode connections. It is common for all the APD pixels of a multielement array to share a small number of common cathode connections.

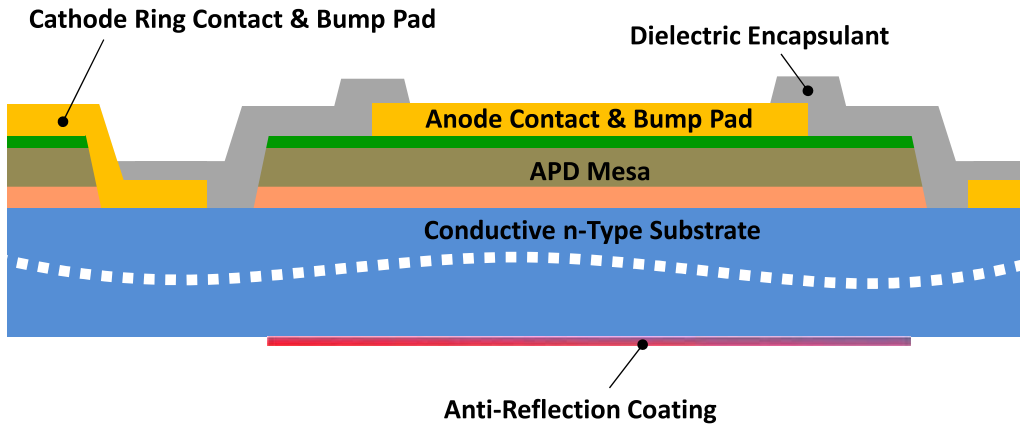


Fig. 5 Cross-sectional sketch of a representative etched mesa InGaAs avalanche photodiode (APD) pixel.

APDs designed for high-bandwidth operation (greater than a few GHz) must minimize their junction thickness in order to cut down the junction transit time. Minimizing the junction thickness minimizes the thickness of the InGaAs absorption layer, so alternative APD configurations like resonant cavity and waveguide designs are necessary to maintain high quantum efficiency (QE). Resonant cavity APDs are illuminated normal to the wafer surface but include mirrors above and below the standard SACM layer stack in order to create a vertical optical cavity that increases the effective number of passes an optical signal takes through the thin absorption layer. In waveguide APDs, the optical signal propagates in the plane of the thin film, so the absorption path length can be increased without increasing the junction transit time. Resonant cavity APDs have the advantage of comparatively easy optical coupling at the resonant wavelength, since the signal may be focused onto the surface of the wafer. Resonant cavity APDs can also be built with very small junction area, which limits capacitance. In comparison, waveguide APDs present the challenge of coupling the optical signal into an edge-facing facet of the in-plane waveguide, and potentially have higher capacitance owing to the larger footprint of the waveguide structure. However, the lattice-matched alloys from which epitaxial distributed Bragg reflectors (DBRs) can be formed for resonant cavity APDs do not provide very much refractive index contrast, necessitating gratings with a large number of periods which are difficult to manufacture. Also, resonant cavity APDs only work well at a single wavelength, since the DBR through which the signal enters is reflective at nonresonant wavelengths. Although these high-bandwidth APD configurations are potentially of interest for light and compact optical free space communications transceivers, APDs are not generally used for very high-bandwidth terrestrial telecommunications because erbium-doped fiber amplifiers (EDFAs) paired with gain-less photodiodes provide superior speed and sensitivity.

Some manufacturing issues have a bigger impact on dense APD pixel arrays than on single-element detectors. Localized morphological defects originate from thin film growth over dust or patches of native oxide remaining on imperfectly cleaned substrate surfaces, or from growth over regions on the substrate surface inadvertently damaged during the thermal desorption process that removes the native oxide. **Fig. 6** is an amplitude mode atomic force microscope (AFM) image of an unusually large dust-related defect. Other morphological defects may be introduced during thin film growth as a metal droplet spat from an effusion cell, or as a patch of non-stoichiometric material created by a local temperature or pressure excursion outside the growth window for the compound semiconductor. When one of these defects occurs within the footprint of an APD's active area it can provide a current path that shunts the diode junction, causing the detector element to fail short. The area density of these defects on an APD wafer is usually on the order of 100 cm^{-2}, and APD active areas are small, so shorted single-element detectors can be screened out and rejected with very modest manufacturing yield loss. However, even when the probability is very low that any given detector element is defective, the likelihood of yielding large format APD arrays with 100% operable pixels drops rapidly with array format. For example, if the likelihood any given 30-μm-diameter pixel is defect-free is about 99.93%, the likelihood that every one of the 1024 pixels in a particular 32×32-format array is defect-free is only about 48.5%; for 64×64-format arrays the estimated yield of completely defect-free array die is only about 5.5%. Insofar as InGaAs APDs operate at voltages that greatly exceed the damage threshold of most CMOS processes and a shorted APD pixel drops most of that high voltage across the readout integrated circuit (ROIC) to which it is connected, it is usually necessary to engineer into the interconnect between APD and ROIC pixels a means of isolating shorted APD pixels.

Limited optical fill factor is another issue particular to APD arrays. Arrays of etched mesa pixels have the advantage of extremely good inter-pixel isolation but cannot make use of the portion of the optical signal which falls between pixel active areas. The dimensional tolerance of the semiconductor process used to pattern and etch pixel mesas determines some minimum separation between adjacent mesas. As the pixel pitch decreases, the dead area consumed by this minimum inter-mesa gap becomes a larger fraction of the total pixel footprint. The ratio of active area to pixel footprint defines an optical fill factor by which the signal is attenuated. Optical fill factor can be recovered by aligning a microlens array to an APD array, such that light incident on the footprint of each microlens element is reimaged onto the active area of the corresponding APD pixel. However, practical implementation can be challenging for finer pixel pitches when the array is to be compatible with small *f*-number external camera

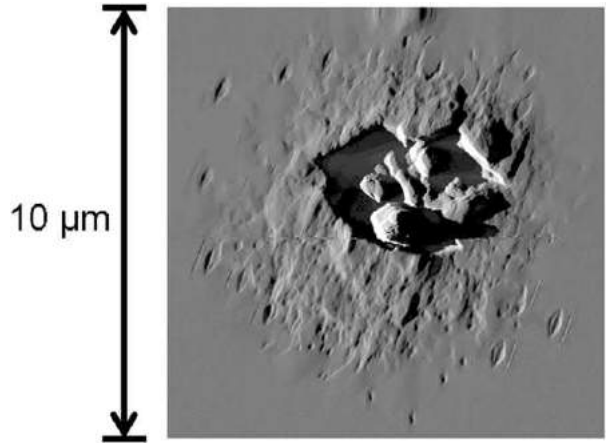


Fig. 6 Amplitude mode atomic force microscope (AFM) image of a large dust-related defect on an avalanche photodiode (APD) wafer surface.

optics of the sort commonly used in compact sensor systems. APD and microlens wafers thinner than about 300 μm – and, depending on size, die which are thinner than about half that – cannot be safely handled as independent pieces without high risk of breakage. Depending on pixel pitch, pixel size, and the *f*-number of external camera optics, a microlens focal length less than 100 μm may be required, necessitating use of special techniques such as substrate removal and attachment of temporary mechanical handling wafers during fabrication.

InGaAs LM APD Device Characteristics

Example current–voltage (*I–V*) and capacitance–voltage (*C–V*) characteristics for a 500-μm-diameter InGaAs LM APD are overlaid in Fig. 7, with reverse current plotted logarithmically against the left-hand axis and capacitance plotted linearly against the right-hand axis. The InGaAs APD in question is one of the larger to be produced commercially. APDs of smaller junction area have proportionally lower dark current and capacitance, although these parameters also depend somewhat on details of structure and manufacture. However, the characteristics in Fig. 7 exhibit features that are common to all InGaAs LM APDs.

Both the current and capacitance characteristics undergo an abrupt step at approximately 22 V reverse bias. This step occurs because the junction's growing depletion region has finished penetrating the doped space charge layer which separates the APD's multiplication and absorption layers. Prior to "punching-through" the charge layer, the junction's depletion region does not extend into the narrow-bandgap InGaAs absorption layer, and a potential energy barrier prevents charge carriers generated in the absorber from reaching the junction and contributing to the diode's reverse current. Reverse current jumps at punch-through because with the potential barrier pulled down, the junction abruptly becomes able to collect the charge carriers generated in the absorption layer. Capacitance drops abruptly at punch-through because the junction's depletion region is able to widen into the substantially undoped absorption layer.

The vertical asymptote in the reverse current characteristics of Fig. 7 occurs at the APD's avalanche breakdown voltage, the critical reverse bias at which the impact ionization process inside the APD junction becomes self-sustaining. Breakdown occurs because secondary charge carriers generated by impact ionization can themselves initiate further impact ionization in a branching chain reaction which need never terminate if the propensity to impact-ionize is high enough. A notional chain of impact ionization events is diagramed in Fig. 8, where time progresses downward on the page and – because of their opposite electric charge – electrons are depicted drifting to the right and holes drift to the left. The starburst symbols from which both the initiating carrier and a new electron–hole pair emerge represent impact ionization events. The fact that impact ionization is a stochastic process is represented in Fig. 8 by the starburst outlines which do not result in additional secondary carriers – these are impact ionization events which might have happened but didn't in the eventuality sketched. Because the multiplication layer is of finite width, carriers can drift out of the multiplier without triggering impact ionization. Below the APD's breakdown voltage, every impact ionization chain eventually terminates in this way, so the avalanche gain is finite. Increasing the reverse bias applied to the APD's terminals exponentially increases the impact ionization rate until, at the breakdown voltage, chains which never terminate become possible. When an APD is biased above its breakdown voltage it behaves like a low value resistor, and the current through the diode is limited by what its bias circuit can supply rather than by the APD junction characteristics. It is advisable to use

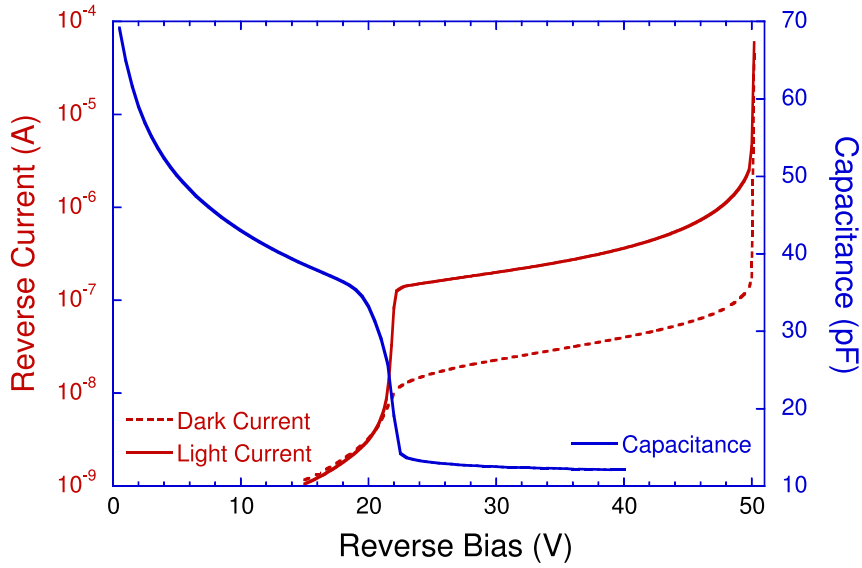


Fig. 7 Current– voltage (*I–V*) and capacitance– voltage (*C–V*) characteristics of a representative 500-μm InGaAs Linear-mode (LM) avalanche photodiode (APD).

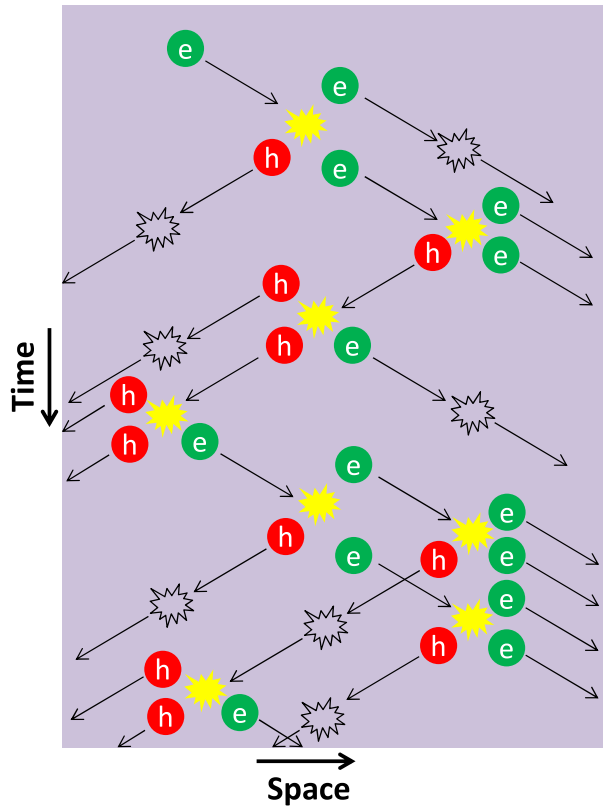


Fig. 8 Diagrammatic sketch of notional impact ionization chains inside an avalanche photodiode's (APD's) multiplier.

current-limited supplies when operating or testing APDs near breakdown, because most InGaAs LM APDs burn out under direct currents in the 1–10 mA range.

QE, avalanche gain (M), and spectral responsivity (R) are usually calculated from direct current (DC) I–V characteristics of the sort plotted in Fig. 7. Responsivity, the APD's ratio of output photocurrent to input optical power, expresses the linear relationship alluded to by the terminology LM APD, and contains factors of gain and QE:

$$R = 8.065544 \times 10^5 \times M \times QE \times \lambda \ (\mathrm{AW}^{-1}) \tag{1}$$

where λ is the optical wavelength in m.

Of the three parameters QE, gain, and responsivity, only responsivity can be measured directly and unambiguously for every InGaAs LM APD. DC responsivity measurements are made by using an optical power meter with a reference photodiode to calibrate the optical power delivered from a monochromatic, continuous-wave (CW) source such as a stabilized diode laser or a monochromator. The DC photocurrent characteristic is found by subtracting the dark I–V characteristic from the light I–V characteristic, so that at any given reverse bias the responsivity is the photocurrent divided by the optical power delivered to the APD's active area.

When operated at a fixed reverse bias, gain saturation causes an InGaAs LM APD's responsivity to drop if the optical signal intensity is increased beyond a certain level. Gain saturation occurs when the photocurrent density is high enough that the mobile charges carrying the current partially compensate the ionized dopant atoms of the APD junction. The impact ionization rate in the APD multiplier is an exponential function of the local electric field strength. Partial compensation of the ionized dopants weakens the electric field, reducing the ionization rate, and therefore the avalanche gain. Different APD designs exhibit gain saturation to different degrees, but in general, the onset of gain saturation occurs at lower optical signal levels when a given APD is operated at higher avalanche gain. This is because the space charge compensation mechanism depends on the charge density carried by the multiplied photocurrent, whereas the optical signal power only determines the primary photocurrent prior to multiplication. Fig. 9 illustrates this point with responsivity data from a 75-μm-diameter InGaAs LM APD. The APD was operated at different reverse biases under 1550 nm illumination at different optical power levels. The lowest operating point depicted, 67.5 V reverse bias, corresponds to an avalanche gain of about $M = 10$. Very little gain saturation is observed for the 67.5 V responsivity curve from 1 nW through 10 μW of optical signal power, whereas the responsivity curve at 70.35 V reverse bias drops from nearly 80 A W^{-1} ($M \approx 80$) for a 1 nW signal to close to 10 A W^{-1} ($M \approx 10$) for a 100 μW signal.

InGaAs LM APDs are usually used to sense weak optical signals, but gain saturation has practical bearing on optical overload behavior and special situations such as optical heterodyne systems in which a weak signal is coherently mixed with a strong optical local oscillator. Gain saturation can also affect the accuracy of APD experiments meant to isolate some property of the

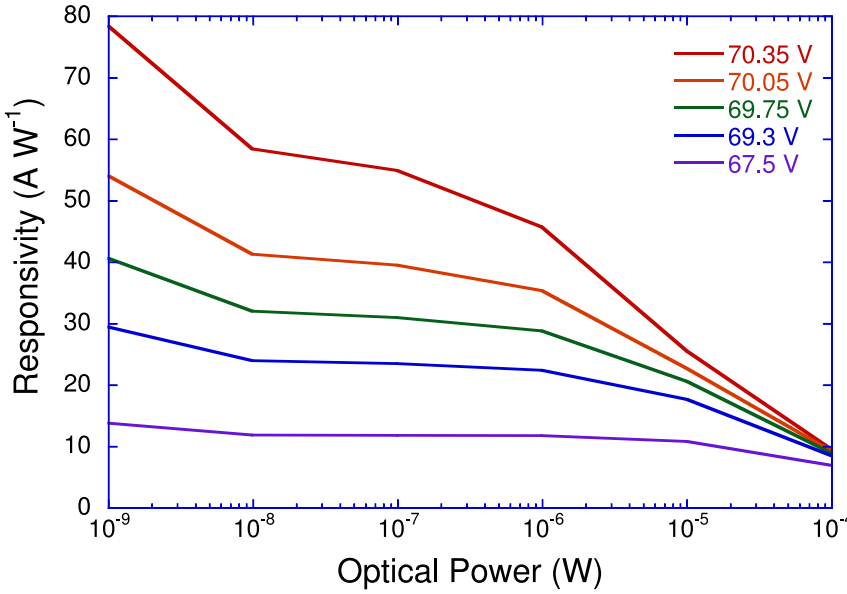


Fig. 9 Avalanche gain saturation as a function of optical power for a representative 75-μm InGaAs linear mode (LM) avalanche photodiode (APD).

photocurrent, in which a measurement taken in the dark is subtracted from the same measurement under strong optical illumination. Spurious measurements can result if the test signal is bright enough to saturate the APD's gain because then the APD's gain operating point is higher for the dark measurement than for the illuminated measurement, and not all of the light-versus-dark difference results from the photocurrent.

Although the DC responsivity of an APD is a commonly reported parameter, APDs are almost always used to detect rapidly modulated or pulsed optical signals. APD responsivity to modulated optical signals decreases with increasing modulation frequency because of the APD's fundamental impulse response, and because of low-pass filtering related to the APD's capacitance in combination with the characteristics of the amplifier which operates the APD. According to the Shockley–Ramo theorem, the instantaneous terminal current of an APD depends upon the instantaneous drift velocities of the electrons and holes in its junction, but can be approximated based on the instantaneous count of electrons and holes present in the junction (n_e and n_h), average electron and hole saturation drift velocities (v_{se} and v_{sh}), and the junction width (w):

$$i(t) \approx \frac{q}{w}\left[v_{se}\, n_e(t) + v_{sh}\, n_h(t)\right]\,(A) \tag{2}$$

where q is the elementary charge.

The fundamental impulse response of an APD has a finite rise time because it takes a finite amount of time for the instantaneous population of secondary carriers in the junction to reach its maximum value. It takes time for primary photocarriers generated in the absorber to reach the multiplier, and once there, carriers which lack sufficient kinetic energy to impact ionize must move through a minimum displacement in the local electric field – the carrier's dead space – in order to pick up the ionization threshold energy. Once energetically active, carriers do not instantly impact ionize. Instead, active carriers travel a variable distance through the multiplier before ionizing – if they ever ionize – based on alloy composition and field-dependent impact ionization rates. Following impact ionization, both the original primary carrier and the newly generated secondary carriers must again pick up the ionization threshold energy before they can trigger new ionizations.

The fall time of an APD's fundamental impulse response is determined by how long impact ionization continues in its multiplier after the peak instantaneous carrier population is reached, and how far secondary carriers must travel from their point of generation in the multiplier to exit the junction. Electrons exit the junction at the cathode and holes exit at the anode, so depending upon the APD's epitaxial stack order (**Fig. 3**), one carrier type has a short distance to travel to exit the junction and the other carrier type must cross the absorption layer. The carrier type with further to travel carries most of the current. When at APD is operated at low gain, generation of secondary carriers may end before any of the secondary carriers of the type with further to travel have exited the junction, in which case the peak of the impulse response coincides with the end of impact ionization and the fall time is determined by the absorber transit time. However, when an APD is operated at higher gain, impact ionization may continue after the peak of the impulse response, extending the fall time of the APD.

Monte Carlo simulations of the impulse response of an APD operating at $M = 12.3$ and $M = 44.0$ are compared in **Fig. 10**. The APD that was simulated has a 1.5-μm-thick InGaAs absorber and a 0.3-μm-thick InAlAs multiplier; including various spacer layers and the charge layer, the total junction width of the simulated APD was 2.25 μm. Saturation drift velocities typical of electrons and holes in these alloys – respectively 10^5 m s^{-1} for electrons and 5×10^4 m s^{-1} for holes – were assumed. The spatial distribution of

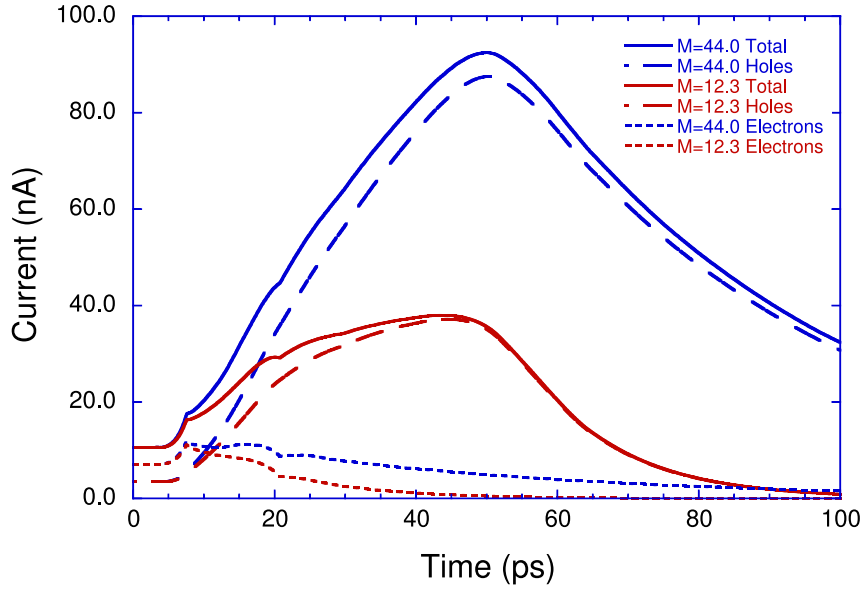


Fig. 10 Monte Carlo simulation of an InGaAs avalanche photodiode's (APD's) photocurrent impulse response.

primary photocarrier generation within the absorber, the dead space effect, and the stochastic nature of impact ionization are treated in this model, but with the approximation of a hard dead space and local field-dependent rather than carrier energy-dependent impact ionization rates, meaning that a carrier's propensity to impact-ionize is treated as abruptly assuming a value that depends on the local field strength once it picks up the ionization threshold energy, as opposed to gradually increasing as it accumulates kinetic energy in excess of the threshold. The approximation that all carriers travel at the average saturation drift velocity at all times is also made, rather than tracking the velocity of individual carriers as they are accelerated by the electric field and lose energy through scattering. These approximations – especially the values chosen for the saturation drift velocities – affect the scaling of the simulated impulse response, but **Fig. 10** illustrates the qualitative features that the carrier type which takes longest to leave the junction dominates the total current, and the fall time is extended at higher avalanche gain. The increase of impulse response duration with gain places a fundamental limit on APD gain-bandwidth product.

The impulse response of an InGaAs LM APD can be measured by exciting it with a laser pulse that is very short compared to the APD's junction transit time, and observing the resulting photocurrent pulse with an oscilloscope, but the input impedance of the scope in combination with the APD's junction capacitance low-pass filters the voltage waveform measured by the oscilloscope. InGaAs LM APD capacitance varies with junction area from less than 0.1 pF for small APD pixels to greater than 10 pF for the largest detector elements. Whereas smaller InGaAs LM APDs typically have sub-nanosecond impulse response, **Fig. 11** illustrates how the 12.3 pF capacitance of a 500-μm-diameter APD leads to much longer impulse response durations.

Similar low-pass filtering effects occur when an APD is connected to an amplifier circuit, and any given amplifier design provides gain over a limited frequency band, so the frequency response of an APD photoreceiver depends as much on the characteristics of the amplifier as on those of the APD. The highest gain-bandwidth products reported for experimental waveguide (Kinsey *et al.*, 2001) and resonant cavity (Lenox *et al.*, 1999) InGaAs LM APDs as individual components is about 300 GHz, but most commercially available InGaAs LM APD photoreceivers have 3 dB bandwidths in the range from tens of megahertz to a few gigahertz. Example plots of photoreceiver responsivity in unit of $V\,W^{-1}$ (i.e., including a factor of the amplifier's transimpedance gain) versus modulation frequency are shown in **Fig. 12** for a photoreceiver assembled from a 75-μm-diameter InGaAs LM APD and a transimpedance amplifier (TIA) chip designed for 2.125 Gbps optical communications.

The photocurrent gain of an APD is the ratio between the photocurrent measured at its terminals to the primary photocurrent generated in its absorber. Avalanche gain is a critical input to APD sensitivity calculations because it affects both signal level and the intensity of excess multiplication noise. APD dark current is also commonly benchmarked to a particular reference gain, and empirical measurement of the APD's excess noise factor requires knowing the gain at which noise power measurements are collected. It is therefore unfortunate that the gain of an InGaAs LM APD often cannot be measured unambiguously.

Although terminal photocurrent is directly measured by subtracting dark from light *I–V* characteristics, measurement of the primary photocurrent can be ambiguous because photocarriers generated in the APD's absorber cannot contribute to the terminal current until the APD is biased above its punch-through voltage. Since the electric field in the APD's multiplier strengthens with increasing reverse bias regardless of whether or not punch-through has yet occurred, it is often the case that at the APD's punch-through voltage, when it first becomes possible to observe photocurrent, impact ionization is already generating avalanche gain in the multiplier.

Bias-dependent collection efficiency of primary photocarriers by the junction is a second factor which can make gain measurements ambiguous. Although the APD abruptly becomes sensitive to light at its punch-through voltage, the collection efficiency of primary photocarriers increases with increasing reverse bias until the InGaAs absorption layer is fully depleted. Since avalanche

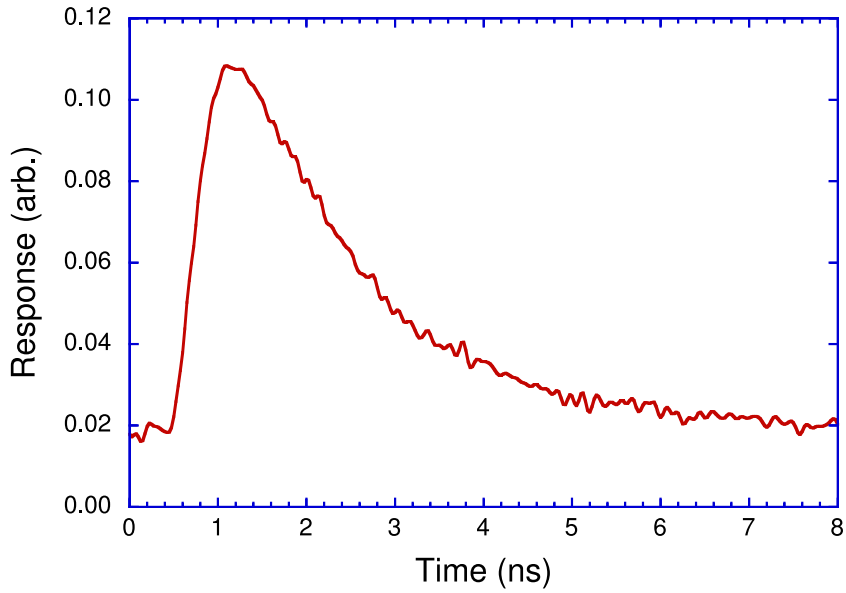


Fig. 11 Measured impulse response of a 500-μm InGaAs linear mode (LM) avalanche photodiode (APD).

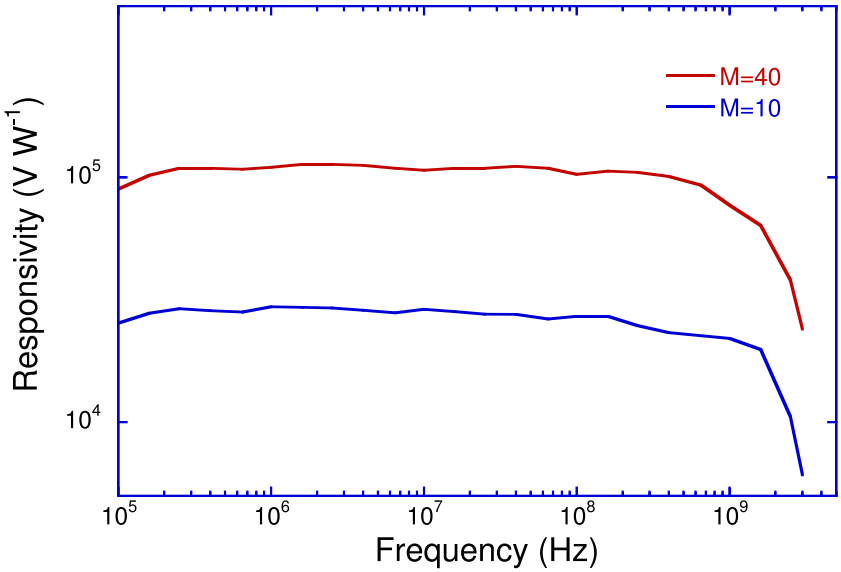


Fig. 12 Measured frequency response of an InGaAs linear mode (LM) avalanche photodiode (APD) photoreceiver assembled from a 75-μm APD and a transimpedance amplifier (TIA) designed for 2.125 Gbps telecommunications.

gain also increases with increasing reverse bias, both the supply of primary photocarriers reaching the multiplier and the gain factor by which they are multiplied increase as the reverse bias is raised past the APD's punch-through voltage. In order to avoid confusing increasing collection efficiency for avalanche gain, gain measurements based on the terminal photocurrent must be made relative to a reference point for which collection efficiency is already at its maximum value. Maximum collection efficiency is usually reached within a few volts of punch-through and is recognizable as a local minimum of the slope of the photocurrent characteristic above the punch-through voltage. The point of minimum photocurrent slope above punch-through can be attributed to the collection efficiency nearing its maximum because any contribution to the slope from avalanche gain increases monotonically with reverse bias. The photocurrent characteristic of an APD for which the primary photocurrent is directly observable has a slope above punch-through that approaches zero where collection efficiency is maximized and increases gradually thereafter as avalanche gain turns on. The point of minimum photocurrent slope above punch-through is also used as a reference point for relative gain measurements in APDs where avalanche gain is already active in the multiplier at punch-through, but in this case, the minimum slope is nonzero and an absolute measurement of the primary photocurrent cannot be made. Fig. 13 replots on linear

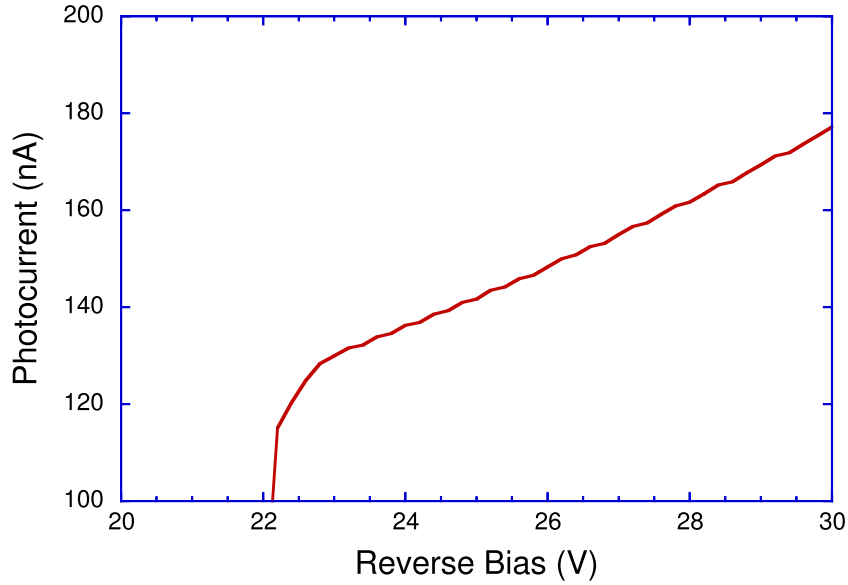


Fig. 13 Photocurrent characteristic of a representative 500-μm InGaAs linear mode (LM) avalanche photodiode (APD).

axes the photocurrent characteristic derived from **Fig. 7**. The minimum slope reference point occurs near 23 V reverse bias but is nonzero, from which one can infer that the multiplier is already generating some avalanche gain at punch-through.

If an InGaAs LM APD's primary photocurrent cannot be directly measured, that prevents exact measurement of its avalanche gain. In that circumstance, measurement of QE is similarly ambiguous, because the gain must be known in order to find the QE from a responsivity measurement. In this common circumstance, physical arguments or measurements on physically analogous photodiodes which lack a charge layer can be used to bound the plausible QE and gain, based on responsivity measurements. For instance, the bottom-illuminated 500-μm-diameter APD whose *I*–*V* characteristics were plotted in **Figs. 7** and **13** has a 1.5-μm-thick InGaAs absorption layer. Ellipsometric measurements on pure InGaAs films give an absorption coefficient of 6700 cm^{-1} at 1500 nm. The minimum plausible QE at 1500 nm, when this APD is operated at 23 V reverse bias, can be estimated by assuming an antireflection coating that transmits 99% of the incident optical power, transmission losses of about 15% through the doped substrate, and absorption of about 63.4% of the optical signal in a single pass through the InGaAs layer, calculated from the known absorption coefficient under the assumption of near perfect collection efficiency. This results in a lower bound estimate of QE = 53.3%. On the other hand, in this design the optical signal is intended to back-reflect from the underside of the APD's anode contact, passing through the absorber a second time. In the best case scenario the substrate absorption is only 5% and the back-reflection is lossless, giving a high estimate for the plausible QE of about QE = 81.4%. Eq. (1) gives unity-gain responsivities at 1500 nm of $R = 0.64\ A\ W^{-1}$ in the worst case QE scenario and $R = 0.98\ A\ W^{-1}$ in the best case. In comparison, an empirical spectral responsivity measurement at 1500 nm with the APD biased at 23 V finds a responsivity of $R = 0.99\ A\ W^{-1}$, implying that at 23 V the APD is actually operating with a gain between $M = 1.0$ and $M = 1.5$. The continuous increase of photocurrent with reverse bias observed in **Fig. 13** shows that by 23 V some avalanche gain is already occurring, which is consistent with the range estimated for the gain. Example responsivity and QE spectra for this APD that reflect the gain calibration ambiguity are plotted in **Fig. 14**.

Multiplied shot noise is the dominant noise source in an InGaAs LM APD. Because InGaAs LM APDs are exclusively used to sense fast optical signals, any significant $1/f$ noise generated by an InGaAs LM APD is usually at frequencies below the low-frequency cutoff of its amplifier. Some references include a Johnson–Nyquist noise term in their analysis of LM APD noise but this is unnecessary because diodes only have Johnson–Nyquist noise near zero bias, and because the series contact resistance of an APD is typically too low to contribute measureable noise. Sometimes noise is attributed to a hypothetical load resistor connected in parallel to the APD, but this circuit configuration is practically never employed. Instead, photoreceiver sensitivity analysis should consider the multiplied shot noise of the APD and an input-referred amplifier circuit noise term that is particular to the amplifier design.

An APD's multiplication noise results from the random variation of its avalanche gain. Analysis of APD noise is based on the counting distribution derived by Robert J. McIntyre (McIntyre, 1972). The probability that an APD multiplier operating at some average gain (M) will output a certain number of electrons (n) given an input of a certain number of primary electrons (a) is

$$P_{\mathrm{McIntyre}}(n) = \frac{a\,\Gamma\left[\frac{n}{1-k}+1\right]}{n\,(n-a)!\;\times\Gamma\left[\frac{n\,k}{1-k}+1+a\right]} \times \left[\frac{1+k(M-1)}{M}\right]^{a+\frac{n\,k}{1-k}} \times \left[\frac{(1-k)\,(M-1)}{M}\right]^{n-a} \tag{3}$$

where k is the ratio of impact ionization rates for the two carrier types (the slower rate over the faster rate) and Γ is the Euler gamma function.

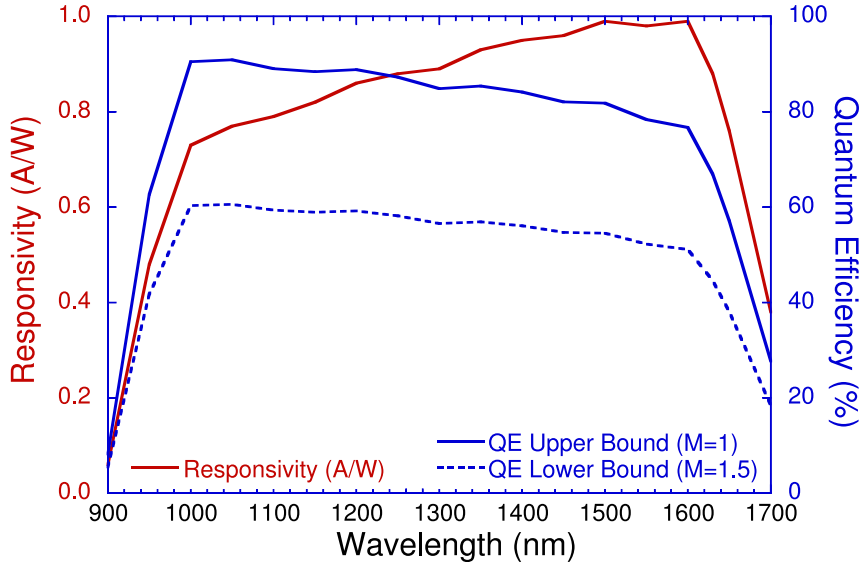


Fig. 14 Measured responsivity and upper and lower bounds on quantum efficiency (QE) for a representative 500-μm InGaAs linear mode (LM) avalanche photodiode (APD).

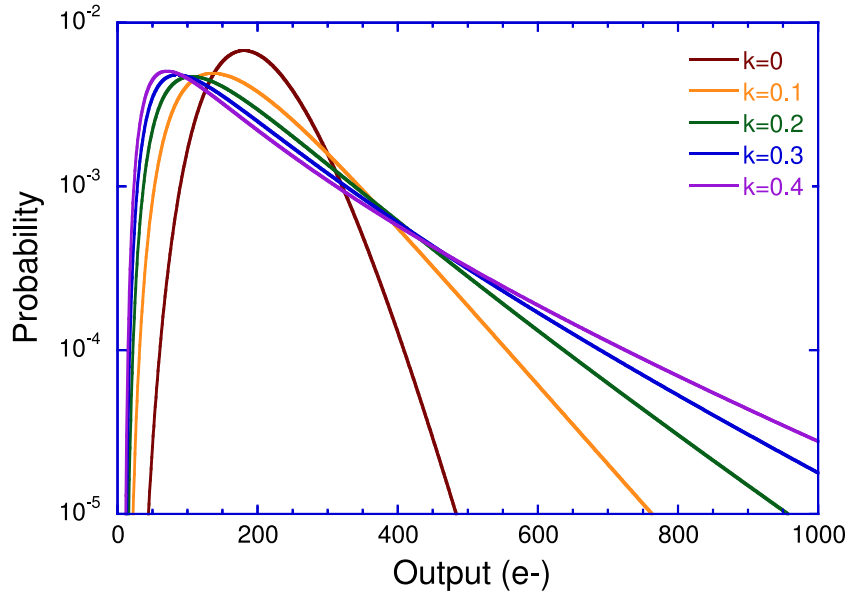


Fig. 15 Output count distributions for linear mode (LM) avalanche photodiodes (APDs) operating at M=20 and responding to 10 primary photoelectrons, but differing in terms of k.

McIntyre distributions for APDs operating at the same average gain ($M=20$) and illuminated by the same signal strength ($a=10$ primary photoelectrons) but differing in k are plotted in **Fig. 15**. These distributions illustrate the practical meaning of different values of k. For the same input signal strength and the same average gain, an APD with lower k will have a higher probability of detecting a signal and a lower probability of generating a false alarm.

These statements assume that the APD is employed in a photoreceiver equipped with a binary decision circuit that rejects inputs below a certain detection threshold, and that the mean signal photocurrent is larger than the mean dark current. In this common scenario, a single detection threshold is simultaneously in the high-output tail of the dark current distribution but comfortably lower than the bulk of the photocurrent distribution's probability density (**Fig. 16**), such that the longer tail of the high-k distribution increases the probability of false alarm but reduces signal detection probability by decreasing the distribution's median output value. The detection threshold is employed to reject false alarms arising from total photoreceiver noise, of which the APD's dark current is one component. At the same time, the detection threshold must not be set so high that it also rejects outputs arising from valid photocurrent signals. An output distribution with a higher median for a given input is desirable because

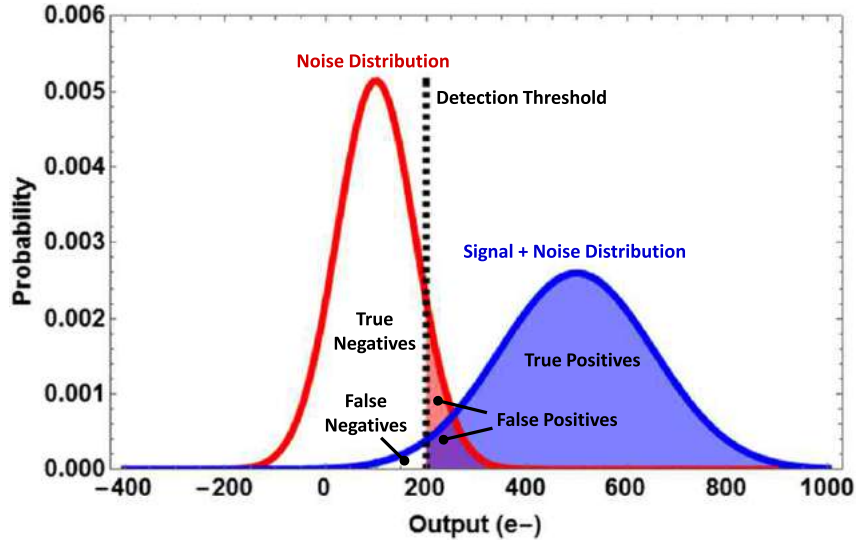


Fig. 16 Illustration of signal and noise distributions for an avalanche photodiode (APD) photoreceiver equipped with a thresholded decision circuit.

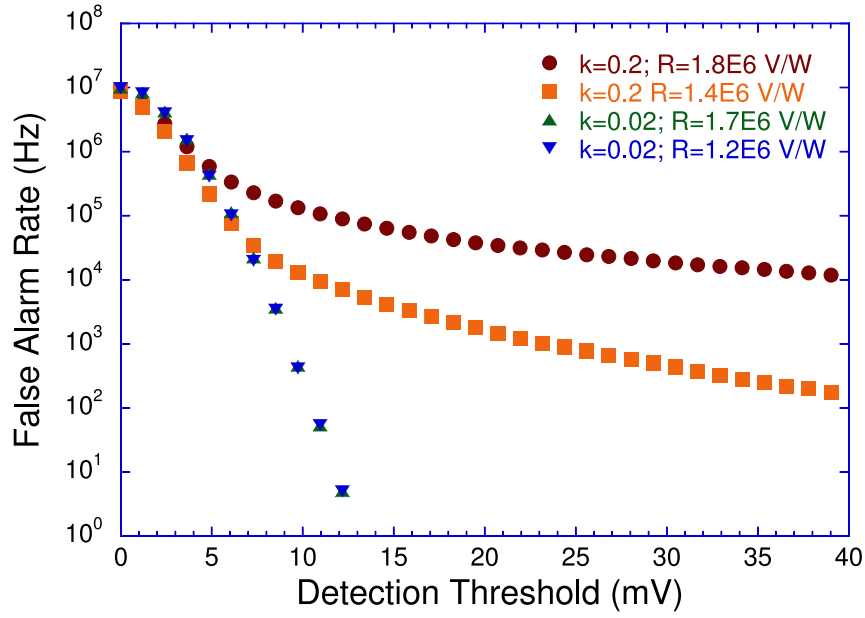


Fig. 17 Comparison of false alarm rate vs. detection threshold profiles for avalanche photodiode (APD) photoreceivers with different values of *k*.

the high median will allow one to set the detection threshold higher without sacrificing signal detection efficiency. On the other hand, a reduced likelihood of very high-output events will help minimize the false alarm rate (FAR) arising from "lucky" dark current electrons that happen to individually experience very high avalanche gain.

Photoreceiver performance actually depends upon the convolution of the APD's output distribution with the amplifier's Gaussian circuit noise and not just the APD's output distribution, but the general observations about how k relates to signal detection and false alarm performance illustrated for the APD output distributions of Fig. 15 still hold in a more rigorous analysis. Fig. 17 is an empirical demonstration of the advantage a low-k APD conveys from the standpoint of FAR performance. FAR measurements are plotted in Fig. 17 for two otherwise identical APD photoreceivers, both operated with similar low-frequency responsivities to 1550 nm light (marked in V W^{-1} because of the TIA's transimpedance), which employ 75-μm-diameter InGaAs LM APDs respectively characterized by $k=0.2$ (red square; orange rectangle) and $k=0.02$ (blue and green triangles). The FAR versus detection threshold behavior of the two receivers is similar at low threshold but diverges sharply at higher threshold, giving the photoreceiver with a $k=0.02$ APD a significant sensitivity advantage in applications that are intolerant of high FARs.

The exact McIntyre distribution often must be used to calculate quantities like FAR or bit error rate (BER) because those photoreceiver characteristics depend on the shape of the tail of the APD's output distribution, and the McIntyre distribution has positive skew. In most cases, the standard FAR and BER formulas which apply to photodiode-based receivers do not provide accurate results when applied to APD-based receivers – even when the standard deviation is adjusted to account for the APD's excess multiplication noise – because those formulas assume the probability density of false alarms falls off as a function of output amplitude with a Gaussian profile. In general, the only circumstances in which the Gaussian approximation provides accurate results are when APDs operate at low gain or with very low k. The FAR of InGaAs LM APD photoreceivers is better modeled by

$$\mathrm{FAR} = \sqrt{\frac{2\pi}{3}}\, N_Q\ \mathrm{BW}\ \ P_{RX}(n_{\mathrm{th}})\ (\mathrm{Hz}) \tag{4}$$

where N_Q is the noise of the photoreceiver (APD and amplifier) referenced to the output of the APD in units of electrons, BW is the bandwidth in Hz of the signal chain into the threshold comparator, and $P_{\mathrm{RX}}(n_{\mathrm{th}})$ is the McIntyre distribution of the APD's multiplied dark current convolved with the amplifier's input noise, evaluated at an electron count corresponding to the detection threshold.

The most common measures of photoreceiver sensitivity such as SNR and NEP define noise based on the standard deviation of an APD's output distribution, and so are not affected by the divergence of the tail of the McIntyre distribution from the Gaussian approximation, which normally becomes and issue a few standard deviations beyond the mean. For this type of analysis, it is sufficient to compute the variance of the APD's output. In units of electrons the APD output variance is found using the Burgess variance theorem:

$$\begin{aligned}\mathrm{var}(n) &= M^2\ \mathrm{var}(a) + \langle a\rangle \mathrm{var}(m)\\ &= M^2\ F\ \langle a\rangle\ \ (\mathrm{e}^{-2})\end{aligned} \tag{5}$$

where the excess noise factor (F) is defined as:

$$F \equiv \frac{\langle m^2\rangle}{M^2} \tag{6}$$

In Eq. (5) the symbol m is a per-electron gain random variable, and the symbols a, M, and n have the same meaning as in Eq. (3). The noise factor of Eq. (6) is deemed "excess" because if APD gain was deterministic, Eq. (5) would simply read $\mathrm{var}(n) = M^2\ \langle a\rangle$. Eq. (5) assumes that the primary photocarrier count represented by a is generated by a Poisson process, so the substitution $\mathrm{var}(a) = \langle a\rangle$ can be made.

For most linear-mode APDs, including InGaAs APDs, the excess noise factor has the gain-dependence derived by McIntyre for thick, uniform junctions (McIntyre, 1966):

$$F = M\left[1 - (1-k)\left(\frac{M-1}{M}\right)^2\right] \tag{7}$$

In Eq. (7), the parameter k is the same ratio of impact ionization rates appearing in Eq. (3). When $k > 0$, k is the slope of the excess noise curve as a function of gain, in the limit of high gain. For single-carrier multiplication, $k = 0$, and $F \to 2$ in the limit of high gain. Another feature of single-carrier $k = 0$ multiplication is that avalanche breakdown cannot occur because the only carriers capable of impact ionizing all drift in the same direction across the junction. The gain curve of a $k = 0$ APD does not exhibit the vertical asymptote shown in **Fig. 7**, enabling stable operation at higher gain than a $k > 0$ APD. Different InGaAs LM APDs vary with respect to the value of k that characterizes their multiplication noise. Holes ionize more readily than electrons in InP, and InGaAs APDs with comparatively thick ($\sim$μm) InP multipliers typically have $k = 0.4$. The impact ionization rate of electrons is higher than that of holes in InAlAs; APDs with thick ($\sim$μm) InAlAs multipliers are characterized by $k = 0.3$ whereas those with InAlAs multipliers thinner than about 0.3 μm have $k < 0.2$ (Lenox *et al.*, 1998). InGaAs LM APDs with multipliers engineered to suppress hole multiplication have been reported in which $k = 0.04$ (Williams *et al.*, 2013).

In many applications it is common to work in units of current rather than electron count. The noise spectral intensity theorem for APDs that is equivalent to Eq. (5) is

$$S_I = 2\ q\ M^2\ F\ I_{\mathrm{primary}}\ (A^2\mathrm{Hz}^{-1}) \tag{8}$$

where I_{primary} is the total unmultiplied current (the sum of signal and background photocurrent and dark current) prior to injection into the APD's multiplier. In principle, there can be dark current leakage paths in an APD which bypass the multiplying junction which would not be multiplied and therefore which would not be part of I_{primary} in Eq. (8). Also, in some exotic InGaAs LM APDs a substantial portion of the total dark current is generated inside the multiplier by tunneling rather than originating in the absorber, and so – because of a different path length through the multiplier – experiences a different amount of avalanche gain and contributes a different amount of multiplied shot noise than the photocurrent. However, in practice, dark current which originates in the absorber and which experiences the APD's full multiplication dominates in most commercially available InGaAs LM APDs, and the noise on that dark current is well modeled by Eq. (8).

As explained in connection to gain and QE measurements, unambiguous measurements of primary photocurrent cannot be made on an InGaAs LM APD if avalanche gain in its multiplier turns on at a lower reverse bias than its punch-through voltage. For APDs where I_{primary} is not observable, the best practice when making APD photoreceiver sensitivity calculations with Eq. (8) is to

use the terminal current in place of the quantity $M \times I_{primary}$, and to use the best estimate of the gain for the second factor of M. Uncertainty about $I_{primary}$ also affects the accuracy of experimental measurements of the excess noise factor, because the most common method of measuring F involves measuring noise power spectral intensity, S_p, with a noise figure meter, converting S_p to S_I using an impedance that is characteristic of the test system but which must be measured, and then applying Eq. (8) to find F from S_I (which requires an estimate of M). For this reason, the most accurate measurements of F are made on special test structures which lack charge layers, and which do not include InGaAs absorbers. Instead, photocarriers are generated by illuminating the test structure with a short-wavelength laser, and both the primary current and the avalanche gain can be unambiguously measured because there is no charge layer to prevent collection of photocarriers at low reverse bias.

Eqs. (5) and (8), respectively, calculate multiplied shot noise variance expressed in units of electrons or in units of current; the square root gives the standard deviation, which is the most common measure of noise. The total noise of an APD photoreceiver assembled from a charge-integrating capacitive-feedback transimpedance amplifier (CTIA) is

$$\begin{aligned} N_Q &= \sqrt{N_{CTIA}^2 + \left(\langle a_{dark}\rangle + \langle a_{bg}\rangle + \langle a_{signal}\rangle\right) M^2 F} \\ &= \sqrt{N_{CTIA}^2 + \left[\frac{\tau}{q}\left(I_{dark} + I_{bg}\right) + Q_{signal}\right] M F} \quad (e^-) \end{aligned} \tag{9}$$

where N_{CTIA} is the input-referred charge noise of the CTIA in units of electrons, τ is the effective DC current integration time of the CTIA, I_{dark}, and I_{bg} are, respectively, the dark current and background photocurrent as measured at the APD's terminals, and Q_{signal} is the product of the APD's QE and gain factor (M) with the optical signal level in units of photons.

The total noise current spectral density of an APD photoreceiver assembled from a resistive-feedback transimpedance amplifier (RTIA) is

$$\sqrt{S_{I\ total}} = \sqrt{S_{I\ TIA} + 2\,q\,M\,F\,(I_{dark} + I_{bg} + I_{signal})} \quad (AHz^{-1/2}) \tag{10}$$

where $S_{I\ TIA}$ is the input-referred noise current spectral intensity of the RTIA and I_{signal} is the product of the APD's spectral responsivity and the optical signal power. If the optical signal intensity is rapidly modulated rather than being quasi-CW, the RMS optical signal power is used.

The total noise calculated in Eq. (9) can be referred to the photoreceiver input, expressed in units of noise-equivalent photons, by dividing N_Q by the product of QE and M:

$$\mathrm{NEPh} = \frac{Q_N}{\mathrm{QE} \times M} \quad (\text{photons}) \tag{11}$$

Similarly, the total noise calculated in Eq. (10) can be referred to the photoreceiver input, expressed in units of noise-equivalent power, by dividing ${S_{I\ total}}^{1/2}$ by the responsivity, R:

$$\mathrm{NEP} = \frac{\sqrt{S_{I\ total}}}{R} \quad (W\ Hz^{-1/2}) \tag{12}$$

The spectral density forms of current noise and NEP can be converted to in-band figures by multiplying either by $BW^{1/2}$.

Optical SNR, such as plotted in **Fig. 1**, is calculated as the ratio of the optical signal in photons to NEPh or in units of power to the in-band NEP.

References

Kinsey, G.S., Campbell, J.C., Dentai, A.G., 2001. Waveguide avalanche photodiode operating at 1.55 μm with a gain-bandwidth product of 320 GHz. IEEE Photonics Technology Letters 13 (8), 842–844.

Lenox, C., Nie, H., Yuan, P., *et al.*, 1999. Resonant-cavity InGaAs/InAlAs avalanche photodiodes with gain-bandwidth product of 290 GHz. IEEE Photonics Technology Letters 11 (9), 1162–1164.

Lenox, C., Yuan, P., Nie, H., *et al.*, 1998. Thin multiplication region InAlAs homojunction avalanche photodiodes. Applied Physics Letters 73 (6), 783–784.

McIntyre, R.J., 1966. Multiplication noise in uniform avalanche photodiodes. IEEE Transactions on Electron Devices ED-13, 164–168.

McIntyre, R.J., 1972. The distribution of gains in uniformly multiplying avalanche photodiodes: Theory. IEEE Transactions on Electron Devices ED-19 (6), 703–713.

Williams, G.M., Compton, M., Ramirez, D.A., Hayat, M.M., Huntington, A.S., 2013. Multi-gain-stage InGaAs avalanche photodiode with enhanced gain and reduced excess noise. IEEE Journal of the Electron Devices Society 1 (2), 54–65.

Very High Range Resolution Lidars

Zeb W Barber, Montana State University – Spectrum Lab, Bozeman, MT, United States

Introduction

Light detection and ranging (lidar) or laser radar (ladar) technologies have already made high impacts in wide range of industrial, construction, scientific, and defense fields, and future applications of lidar will multiply as performance and cost improve. Consumer lidars already exist in the form of laser rangefinders for golfers and hunters, but the push toward autonomous vehicles and robots will increase the proliferation of lidar technology to unprecedented levels (Ackerman, 2016; Khoshelham and Zlatanova, 2016; see Section Relevant Websites). An important performance metric for all lidars is the resolution of the range measurement, which determines the suitability for different applications. Range resolution at the level of a meter to a few centimeters are the most common and are good for general range finding, mapping, and outdoor navigation applications. With very high range resolution lidar technology (as small as a few tens of micrometers) based on coherent frequency modulated continuous wave (FMCW) (De Groot and McGarvey, 1992) or concepts such as dual-comb coherent optical sampling (Coddington *et al.*, 2009); very different classes of applications are enabled. These include: noncontact metrology, high resolution 2D and 3D imaging, and coherent synthetic aperture imaging. This article focuses on very high resolution lidar technologies that enable range measurements with resolution of less than 1 cm. This is a convenient threshold as generally the lidar and ladar technologies required to achieve this level of range resolution are significantly different than more common brief-pulse, direct-detect lidars due to the difficulty of achieving sub-70 ps ($\sim$1 cm range difference) timing resolution.

Range Resolution Versus Range Precision

Before describing very high range resolution lidar technology, it is important to clarify the definition of range resolution. The definition of range resolution used in this article is consistent with that used in the radar community (see Section Relevant Websites). Namely, resolution is defined as the minimum range separation that two point targets of equal power can have and still be resolved (see **Fig. 1**). As all lidars and ladars are based upon measuring the round-trip time-of-flight of light between the ladar system and the target, range resolution is then fundamentally limited by the temporal bandwidth of the optoelectronics as $\Delta R \geq c/2B$, where B is the bandwidth and c is the speed of light. This definition is directly analogous to the Rayleigh resolution criterion used in optical imaging, where the image resolution is fundamentally limited by the spatial bandwidth of the imaging system. For example, in brief-pulse, direct-detection lidar, the fundamental timing (range) resolution is just set by the temporal duration, $\Delta\tau$, of the optical pulse (assuming the detector and electronics are sufficiently fast). For a Fourier limited Gaussian pulse the temporal width is just the inverse of the bandwidth as, $\Delta\tau = 1/B$. Given this temporal pulse width, if two targets are separated by less than $\Delta\tau$ in delay then echo pulses from each target merge into a single slightly wider return pulse and the targets are unresolved.

Unfortunately, this definition of range resolution is not used consistently throughout the literature or in manufacturer specifications. Often metrics better described as range precision (σ_R in **Fig. 1**) or the range grid/bin size (dR in **Fig. 1**) are specified as resolution. Range precision measures the statistical spread of multiple ideally identical range measurements of a single target. Using signal processing and estimation on a lidar signal with sufficient signal-to-noise ratio (SNR) the range precision achieved can be significantly better than the range resolution supported by the bandwidth. Part of the confusion of range resolution and range precision is that many lidars for 3D imaging are designed to only return a single range for each measurement, which blurs the

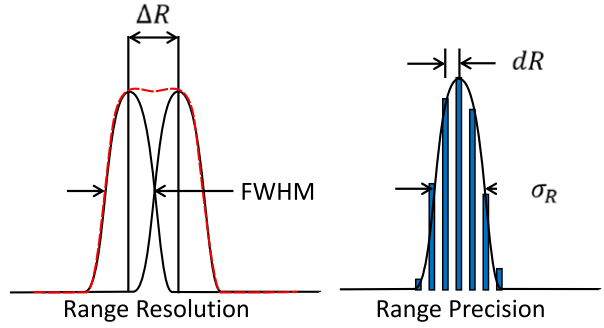


Fig. 1 Graphical definitions of range resolution – minimum separation to resolve two targets, and range precision – deviation of range estimates from a constant target.

 doi:10.1016/B978-0-12-803581-8.09424-8

distinction as the lidar can measure range displacements or differences in range to targets with precision much smaller than the resolution with sufficient SNR. However, the distinction is important as range resolution is an intrinsic property of the lidar system independent of the return power, while range precision is an extrinsic property dependent on the return power and resulting SNR. Therefore, when comparing different lidar technologies, range resolution provides a more fundamental comparison.

An advantage of resolving power in lidar is the ability to reduce the effect of range pulling from weaker nearby scatterers. An extreme example of range pulling can be observed with a common type of time-of-flight based ranging based on phase measurements of periodic intensity modulated optical signals (Fujima *et al.*, 1998). In these systems, the range to a target can be inferred by the period of the modulation and the amount of phase lag on the received light induced by the time-of-flight to the target and back. This type of time-of-flight distance measurement method has been used in surveying and metrology equipment for a long time and is now being used for 3D imaging (Iddan and Yahav, 2001) using special photonic mixing device, complementary metal-oxide-semiconductor (CMOS) cameras also known as "time-of-flight" cameras that are now provided by many vendors (see Section Relevant Websites). With a modulation frequency of 50 MHz (in the range common for these type of systems) and 12 bit resolution of the phase, the range could potentially be measured with a very good precision of $\sigma_R = c/(2f_{\mathrm{mod}} \bullet 2^N) = 0.73$ mm. However, the drawback with this method is that there is essentially no range resolution whatsoever. The return phase of the received light is just the radio frequency (RF) coherent sum of all the scatterers in the scene. This means that if there are multiple returns in the range dimension (e.g., imaging through a window, screen, or foliage; or in a scene with a strong specular reflections which generates multiple path returns) the range will be a weighted average and very sensitive to stray scattering. Additionally, the range is only unambiguous out to one half of a wavelength of the modulation frequency ($\sim$3 m for 50 MHz), which limits the performance of the systems. In surveying and metrology applications these issues are resolved by using retro-reflecting targets to ensure the target of interest dominates the return signal reducing range pulling, and changing the modulation frequency slightly to implement a Vernier effect to resolve ambiguities.

Finally, range resolution is defined analogous to the Rayleigh criterion because many applications of interest require imaging, i.e., resolving multiple returns in the range dimension. For these applications, the resolving power of the lidar system is most important. A good example is the use of lidar in foliage and camouflage penetrating 3D imaging. In order to range gate out the cover, the lidar system must be able to resolve the return from the cover from the object beneath, which means the minimum distance between the cover and the object is proportional to the range resolution. Another example is the use of very high resolution lidar for optical metrology, where reflections from multiple surfaces need to be resolved to measure thicknesses and positions. However, the best example is the use of coherent lidar in synthetic aperture imaging lidar (SAIL or SAL) (Beck *et al.*, 2005). Here the Rayleigh analogy is complete as the image resolution in one dimension is directly set by the range resolution.

Very High Range Resolution Lidar Techniques

Brief-Pulse, Direct-Detect Lidar

For a lidar system to meet a threshold of 1 cm range resolution (timing resolution less than 66.7 ps), the corresponding bandwidth must be in excess of 15 GHz. To achieve this level of resolution with the straightforward brief-pulse, direct-detect lidar approach is very challenging, in particular if one is to simultaneously achieve high sensitivity (i.e., equivalent to longer ranges). Several type of pulsed laser sources with sub-100 ps pulse width exist to support the range resolution including microchip passively Q-switched lasers (Spühler *et al.*, 1999) and modulated master oscillator/fiber power amplifier (MOPA) systems. However, high energy lasers needed for long range lidar applications are less available due to the high peak powers this entails. On the detection side, bandwidths exceeding 50 GHz is achievable with modern optoelectronics developed for the telecommunication industry, however, these components meet severe trade-offs between sensitivity and bandwidth. Typically, these very fast linear PIN detectors are fiber coupled and use a 50 Ω load, which limits the NEP to approximately 10 pW/rtHz when dark current and a 3 dB noise figure (NF) low noise amplifier are included. For a 15 GHz bandwidth pulse this translates to a unity SNR sensitivity of about 1 μW, which is the equivalent of over 500 return photons per pulse. Linear avalanche photodiode (APD) based receivers can achieve a factor of 10 better, but rarely achieve bandwidth greater a few GHz. Operating APDs in Geiger mode can achieve single photon sensitivity and timing resolution of sub-50 ps (Butera *et al.*, 2016) and are the preferred approach for achieving the highest sensitivity and resolution for brief-pulse, direct-detect lidars. Yet, in general, approximately 70 ps/1 cm range resolution is the limit for brief-pulse, direct-detect lidar and prospects for further improvement are not promising due to electronic bandwidth limitations.

Streak Cameras

The use of streak cameras with brief-pulse laser illumination has been used with success for high resolution lidar (Gelbart *et al.*, 2002; Knight *et al.*, 1989; Yang *et al.*, 2012). Streak cameras are optical detectors that use time-to-space mapping mechanism to allow high time resolution. The most common type of streak cameras use a photocathode to generate free-electrons which are accelerated and deflected in space using time varying electric fields. Streak cameras can achieve very high timing resolution (as low as 200 fs; Hamamatsu, 2008) and sensitivity as low as a single photoelectron. However, there are trade-offs with high temporal resolution. This includes timing errors dominated by trigger jitter against the electric field sweep waveform. The timing jitter can be

significantly worse (>50 ps) than the timing resolution of the streak camera reducing absolute ranging capability and making it difficult to average low-level signals to increase sensitivity. Another problem that arises with high timing resolution and sensitivity is reduced dynamic range caused by the need to run at high photomultiplier gain. Finally, as the range window/range resolution ratio is set by the spatial resolution of the camera, the range window to range resolution ratio (i.e., time bandwidth product) can be much smaller than with other approaches. However, a significant benefit of many streak cameras is the ability to use the second spatial dimension of the camera for line imaging or other modalities such as spectroscopy.

Coherent Mode-Locked Lasers

Coherent mode-locked lasers have emerged as a powerful tool for a large range of applications including various forms of lidars. For very high resolution lidar applications several different techniques (Coddington *et al.*, 2009; van den Berg *et al.*, 2015; Delfyett *et al.*, 2012) have been developed that utilize the coherence of the mode-locked laser rather than just relying on the mode-locked laser's ability to make short pulses of light. The key difference is that these methods utilize interferometric or coherent detection methods to resolve time differences much smaller that the timing resolution of the optical detector.

Mode-locked laser based multi-wavelength interferometry

This class of mode-locked laser based very high resolution lidar combines the spectral comb structure of the mode-locked laser with interferometry and spectrally resolving elements (van den Berg *et al.*, 2012, 2015; Wu *et al.*, 2016). The basic operation utilizes a Michelson interferometer to coherently interfere the mode-locked laser with itself. The output port of the interferometer is then spectrally dispersed in space to resolve each comb line (see **Fig. 2**). Each resolved comb line in this interferogram then becomes an individual CW interferometer obeying Michelson interference rule of $I(\lambda) = I_0\left\{\frac{1}{2} + \frac{1}{2}\cos(4\pi R/\lambda)\right\}$. If the repetition rate of the comb is not large (i.e., <1 GHz) individually resolving each comb line with an optical grating spectrometer can be difficult, so some researchers have utilized an etalon based dispersive element known as a virtually imaged phased array (VIPA) (Shirasaki, 1996) that provides a high, but periodic, dispersion in another dimension to fully resolve the comb. Fully resolving the comb is important to eliminate the pulse structure of the mode-locked laser on the optical detector, without which the pulses would also need to be temporally overlapped to see the spectral interference.

The reason why this technique can be classified under the category of lidar ranging is that with a large number of comb lines a range profile can be constructed by a fast Fourier transform (FFT) of the interferogram. The range resolution is set by the bandwidth $\Delta R = c/(2N\, f_{rep})$, where N is the number of resolved comb lines. However, due to the equally spaced comb structure, the interferogram is discretely sampled at the repetition rate which forms a frequency domain Nyquist limit. The range profile is then ambiguous with a period of $\tau_{max} = 1/f_{rep} \Rightarrow R_{max} = c\tau_{max}/2 = c/2f_{rep}$. For example, a 1 GHz repetition rate leads to a 15 cm ambiguity period. These ambiguities can be resolved by changing the repetition rate of the laser to sample with a different repetition rate.

It is important to note that this technique relies on the coherence of the mode-locked laser to maintain contrast of the interferogram. This means the coherence length of each individual comb lines must larger than the range to the target. However, this technique does not require coherence between the comb lines. A laser made-up of many independent lasers with tight tolerance of the line frequencies would work as well. This also means that the technique is robust to the temporal shape of the transmitted mode-locked laser pulse, however relative dispersion and/or line-to-line relative phase errors between the Tx/Rx path and the reference path are important. In addition to the laser coherence, this technique utilizes coherent detection which means the target must maintain holographic stability with the Tx/Rx system during the integration time of the camera.

Dual-comb interferometry based distance measurement

A second method of mode-locked laser based very high resolution lidar is based on an innovative detection method that uses a second mode-locked comb as a local oscillator rather than a copy of the signal laser (Coddington *et al.*, 2009; Wu *et al.*, 2016; Durán *et al.*, 2016; Zhang *et al.*, 2014). The second comb has a slightly different repetition rate so that the relative pulse delay on

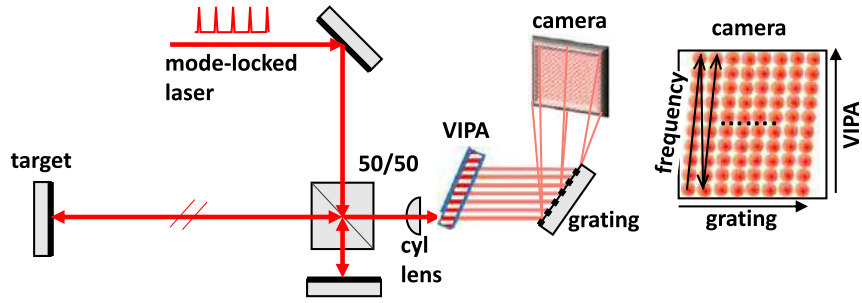


Fig. 2 A schematic of one possible realization of a mode-locked laser based multi-wavelength interferometric lidar system. Adapted with permission from van den Berg, S.A., Persijn, S.T., Kok, G.J.P., Zeitouny, M.G., Bhattacharya, N., 2012. Many-wavelength interferometry with thousands of lasers for absolute distance measurement. Physics Review Letters 108, 183901. doi:10.1103/PhysRevLett.108.183901.

the detector scans with a period equal to the inverse of the difference in the repetition rates ($T_{meas}=1/\Delta f_{rep}$). If balanced homodyne detection is used the output is zero unless the pulses overlap in time. When the pulses do overlap in time the coherent interference determines the strength of the homodyne signal (see **Fig. 3(a)**). If the detector output is digitized with a sample rate synched to the local oscillator (LO) comb, then the output is a sampled version of the optical field of the signal laser with very high time resolution. Lidar ranging is simply accomplished by using one of the combs as a transmitter (see **Fig. 4**). The range profile then is a series of echo pulses in the dual-comb heterodyne output (or nonlinear sampling (Zhang *et al.*, 2014)), which not only provides the envelope of the pulse but also the pulse phase. The sampling resolution of the range profile is related to the repetition rates as $\Delta t=\Delta f_{rep}/(f_{rep,sig}\cdot f_{rep,LO})$. For example, with $\sim$100 MHz repetition rates with an $\Delta f_{rep}=1$ kHz the sampling resolution is 100 fs or effectively a 10 THz sampling rate. This time sampling condition sets the range resolution, but also must satisfy Nyquist criteria on the bandwidth of the comb to prevent aliasing. The Nyquist condition is $\Delta t\leq 1/2B$, where B is the bandwidth of the combs. This then sets the fundamental trade-offs between range resolution, unambiguous range window, and measurement rate (Wu *et al.*, 2014). These trade-offs often necessitate the use of long measurement periods and/or use of optical bandpass filters to limit the optical bandwidth.

In addition to the time domain description of this dual-comb approach, as shown in **Fig. 3(b)**, the frequency domain description reveals it also as a massively multi-heterodyne detection technique (Coddington *et al.*, 2010). The difference in the repetition rate of the two combs causes the heterodyne beat frequency of adjacent comb modes to increase linearly with mode

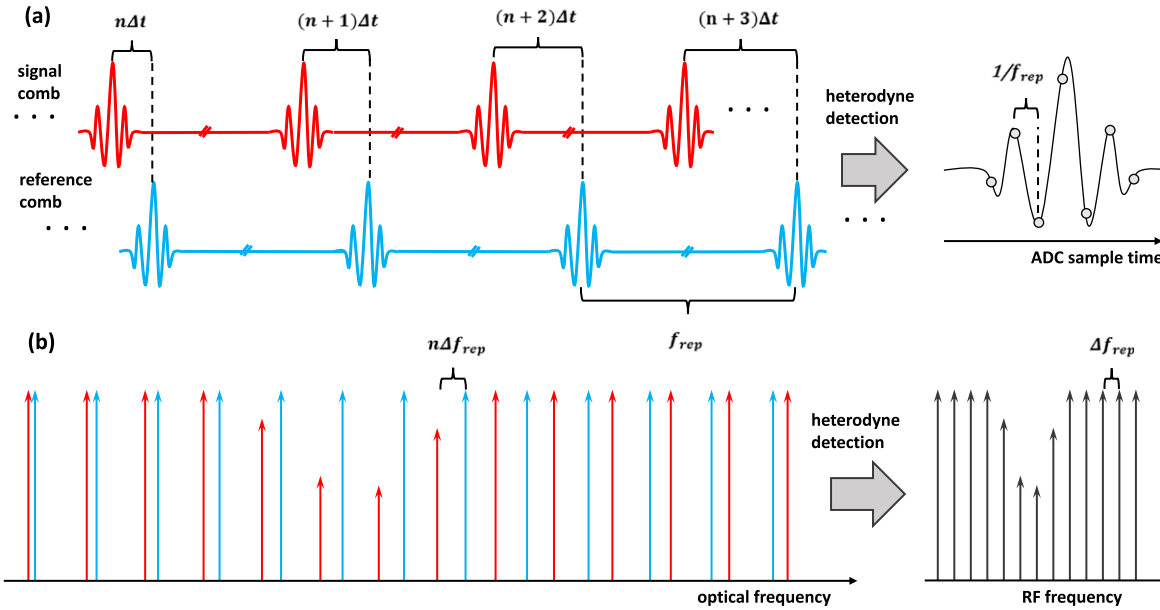


Fig. 3 Graphical description of the dual-comb approach in (a) the time domain as a cross-correlation and (b) the frequency domain as a multi-heterodyne signal showing how the optical comb translates to a radio frequency comb including individual line amplitudes. Adapted with permission from Coddington, I., Swann, W.C., Newbury, N.R., 2010. Coherent dual-comb spectroscopy at high signal-to-noise ratio. Physical Review A 82, 043817. doi:10.1103/PhysRevA.82.043817.

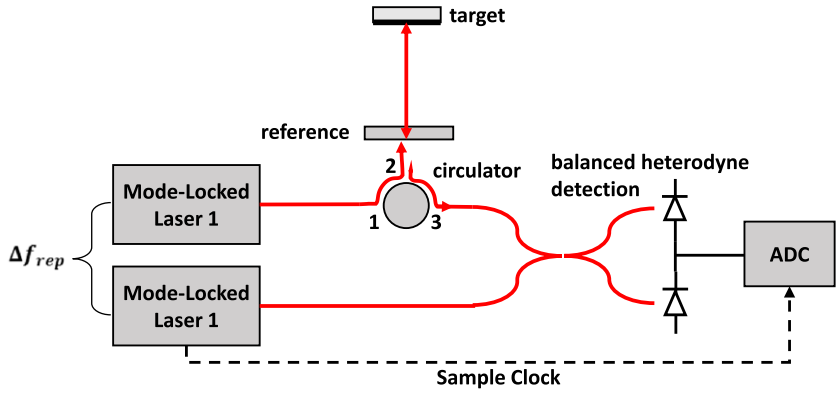


Fig. 4 Simple schematic setup used to perform dual-comb very high resolution ranging approach.

number. This effectively translates the phase and amplitude of the optical frequency comb onto a fine resolution RF comb with spacing equal to Δf_{rep}. Therefore if the amplitude of an optical comb line is reduced, for example, by a molecular absorption in the path, this will be observed on the corresponding RF comb line. While this RF comb can be obtained by use of an RF spectrum analyzer, an FFT of the time domain dual-comb signal provides a phase coherent spectrum. From this perspective the dual-comb approach is similar to Fourier transform infrared red (FTIR) spectroscopy, where the scanned delay between the signal and reference paths is provided by Δf_{rep} rather than the physical motion of an interferometer mirror.

Dual-comb interferometry has relatively strict requirements on the mutual coherence of the mode-locked lasers. The technique was originally demonstrated using highly stabilized mode-locked lasers (also known as optical frequency combs), where both the repetition rate and carrier offset degrees of freedom are locked to stable RF sources. This ensures that the relative phases of all the mode-locked lines are well defined during the measurement time and allows one to simply time average the interferogram on an oscilloscope to achieve improved SNR. An additional benefit of using highly stabilized comb lasers is that the optical phase of the return pulses in the interferogram allows interferometric distance measurement along-side the time-of-flight range measurement provided by the envelope. As demonstrated in Liu *et al.* (2011) high resolution lidar ranging can be achieved using free-running mode-locked lasers as long as the linewidth of the individual comb lines are less than the difference in the repetition rate, Δf_{rep}. However with this technique, the absolute phase of the pulse drifts with time necessitating more advanced signal processing techniques to perform averaging of the signal envelope which provides the time-of-flight lidar measurement. Other difficulties of this technique include: (1) The large optical pulse amplitudes, which can lead to nonlinearities in the photodetector and limited dynamic range of the measurement. This has been mitigated by using large fixed fiber dispersion to spread the pulse in time then using digital pulse compression in the signal processing stage. (2) The range ambiguity caused by the periodic pulse structure. This can be mitigated by making Vernier type measurements by changing the repetition rates of the lasers (Nakajima and Minoshima, 2015).

FMCW Lidar

The final very high range resolution lidar method that will be described is coherent FMCW lidar. FMCW ranging techniques have long been utilized in the RF domain for radar applications (Stove, 1992). The benefits of FMCW techniques include: continuous wave (as opposed to pulsed) output that provides higher energy per measurement at low peak powers which is also a better match to continuous waveform (CW) amplifiers, shot-noise-limited single-photon sensitivity, and high dynamic range. While many types of frequency modulated waveforms have been investigated, the dominant waveform type in use for both the radar and lidar domains is the linear frequency chirp (LFC). The LFC waveform is a linear ramp of the frequency versus time that provides a unique linear correspondence of time and frequency (see **Fig. 5**). Combined with coherent detection against a local copy of the LFC waveform (i.e., a LO), this structure produces a coherent beat signal, the frequency of which indicates the delay between the two LFC waveforms as $f(\tau)=\kappa\tau$, where κ is the chirp rate. Extending this to multiple range returns as in a radar or lidar system, the full range profile can be obtained from the spectrum of the coherently detected signal (generally in the form of an FFT of the digitized time domain signal). This procedure, known as stretched processing, makes a simple and bandwidth efficient signal processing system as compared to other coded waveform ranging techniques.

The main benefit of this stretched processing approach is that when the maximum relative delay of the signal versus the LO is small compared to the chirp time, T_c, the bandwidth of the coherent beat note is much less than the overall chirp bandwidth, B. This allows the use of relatively low bandwidth photodetectors and analog-to-digital converters (ADCs) while maintaining the full range resolution provided by the chirp bandwidth. Low bandwidth detectors have higher dynamic range and make it easier to achieve shot-noise-limited detection, while high speed ADCs are expensive and are improving only slowly (Walden, 1999) and the best speed versus dynamic range performance of ADCs are currently in the couple hundred megasample per second (MS/s) regime (see Section Relevant Websites). The trade-off with this bandwidth compression is to limit the maximum delay and/or range

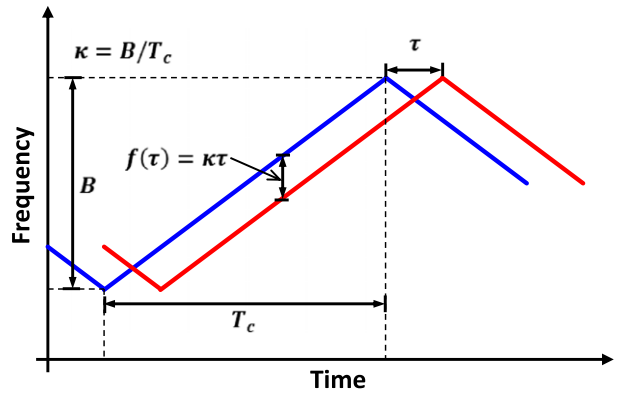


Fig. 5 Diagram of the time-to-frequency mapping performed by the coherent mixing of linear frequency chirps as in frequency modulated continuous wave (FMCW) lidar.

window. For very long range lidar applications this can be a significant issue. As an example, assume an LFC with $B=100$ GHz of bandwidth and a chirp duration of $T_c=1$ ms, which translates to a chirp rate of $k=100$ MHz/μs. At this chirp rate, a 2 km range ($\tau=13.3$ μs round-trip delay) produces a frequency of ~ 1.33 GHz. This is much smaller than the 100 GHz chirp bandwidth, but still relatively large compared to the <100 MHz regime in which one would like to operate. Additionally, to sample this high frequency signal at more than Nyquist requires significant memory and signal processing requirements. Assuming 3 GS/s sampling rate for the full T_c would produce 3 million samples that need to be processed with an FFT. For these reasons long range FMCW lidar systems may utilize hardware based range shifting mechanisms such as frequency shifters or multiple chirp sources to limit the bandwidth requirements of the photodetector, ADC and digital-signal-processing system.

A second issue with FMCW lidar that has not been brought up with the other coherent techniques is the relative motion of the target and Tx/Rx system which create Doppler shifts ($f_D=v_l/\lambda$) where v_l is the line-of-sight velocity. In direct-detect lidar systems only the optical power is detected and the optical frequency shifts caused by the Doppler shift are too small to greatly affect the measurement. For coherently detected systems, however, the Doppler shift translates in to RF frequency shifts (~ 645 KHz/(m/s) for 1550 nm light) of the detected signal. This has effects in all coherent lidar systems, but they are most consequential in FMCW lidar with LFC's because the frequency shift of the coherently detected signal directly translates to a shift in apparent range. This range-Doppler ambiguity can be resolved by making measurements at different chirp rates. One common method is to make measurements on both the up frequency chirp and the down frequency chirp of a triangle type chirp waveform as in Fig. 5. By changing the sign of the chirp rate, the relative sign of the Doppler frequency shift and range delay are reversed and the range-Doppler ambiguity can be resolved from the two measurements.

Very High Resolution FMCW Lidar

The rest of this article focuses on FMCW very high resolution lidar with discussion of the technical aspects of the technique, laser sources, and its application to imaging and metrology.

Chirp Nonlinearity

To take advantage of the bandwidth compressing ability of FMCW lidar in the very high resolution regime requires chirped laser sources with large frequency tuning range. While microwave photonic methods of generating optical chirps using microwave chirps and external modulators has been used successfully for resolutions of about 1 cm, the most successful demonstrations of very high resolution lidar to date have relied on the use of chirps created by directly tuning the output frequency of a CW laser. The main difficulty with this approach is that widely tunable lasers are generally not very stable and often do not tune linearly. An important aspect LFC based FMCW lidar is that the linear chirp enables the use of a simple FFT to compress the pulse and obtain a high resolution range profile. If the chirp is nonlinear then the linear mapping of frequency-to-time is not perfect and, as illustrated in Fig. 6, the time domain FMCW signal is no longer a single frequency sinusoid that can be compressed with a simple FFT. If the nonlinear form of the chirp is well-behaved and known it can be possible to compensate for the nonlinearities in the FMCW signal with additional signal processing. However, often the nonlinearities are unknown or vary from chirp-to-chirp, which means that some method is required to measure and correct the nonlinearities on each chirp.

The most common method for measuring the nonlinearities of a chirp from a tunable laser is to simply measure the FMCW signal from a single fixed delay, often from a fiber based reference interferometer. If the delay is fixed and not wavelength dependent (Barber *et al.*, 2010) then the FMCM signal should be a single sinusoid and the phase of the time domain signal should

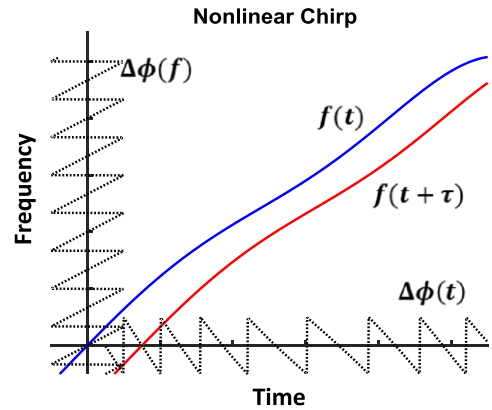


Fig. 6 Resulting FMCW signal from a nonlinear chirp. The blue and red lines represent the frequency of the imperfect chirp waveform vs. time. The dotted lines represent the relative phase of the interference signal in the time (along horizontal axis) and frequency domain (along the vertical axis).

advance perfectly linearly in time (ignoring phase wraps). Deviations from a single frequency sinusoid indicate local variations in the chirp rate with advancing phase indicating locally higher chirp rate and retarding phase indicating lower chirp rate. By tracking the phase of the FMCW signal the nonlinearities of the chirp can be determined. As illustrated in **Fig. 6**, with a fixed delay the relative phase of the two chirps referenced in the frequency domain is consistent because the relative phase (up to a constant overall offset phase) is just $\Delta\phi(f') = 2\pi f' t$, where f' is the instantaneous frequency. However, due to the chirp nonlinearities the time domain signal is not consistent with the frequency domain signal. An important aspect of this reference delay approach that should not be overlooked is that the interferometer signal is ambiguous to the sign of the chirp rate. This means that if the chirp is not monotonic there can be errors in tracking the phase when the chirp rate goes through zero. One way to break this ambiguity, is to add an acousto-optic modulator (AOM) to one arm of the interferometer. This shifts the beat note away from DC so that positive chirp rates show up as a positive frequency shift from the AOM frequency and negative chirp rates show up as negative shifts.

Active Chirp Laser Stabilization for FMCW Lidar

While this delay based chirp referencing method is common, how it is used varies. In very short range applications such as swept source optical coherence tomography (SS-OCT) (Huber *et al.*, 2005) the reference interferometer output is often just used to trigger samples of the ADC in the signal path at every zero phase crossing. This forces the time domain samples to be regular in frequency. Other works use the reference output to determine the chirp nonlinearities and use digital signal processing to correct the FMCW signal in post-processing (Beck *et al.*, 2005; Ahn *et al.*, 2005; Cabral and Rebordão 2007). The last approach – and what will be described in further detail here – is to use the reference interferometer output as an error signal in a phase-locked loop (PLL) to actively stabilize the frequency of the chirp (Roos *et al.*, 2009; Satyan *et al.*, 2009; Greiner *et al.*, 1998). In general, the sweep rate of most tunable lasers can be actively stabilized to achieve a linear frequency chirp suitable for lidar range measurements. However, the frequency tuning, free-running linewidth, and feedback characteristics of the particular laser will determine the details of the stabilization approach and the resulting chirp properties.

Broadband (5 THz) chirp stabilization

A first example of active chirp stabilization is that described in Roos *et al.* (2009). The laser used was developed for telecommunications test and measurement and is able to tune mod-hop free over an extremely large range (up to >80 nm from 1525 to 1605 nm). The relatively long external cavity and tunes mechanically by changing the angle of the feedback grating and so has a slow frequency sweep period, >500 ms, with a maximum chirp rate of ~ 10 THz/sec. Around this relatively slow sweep there are faster frequency tuning mechanisms including: a piezoelectric actuator that provides a few kHz bandwidth and 1 GHz of frequency tuning range; and direct feedback to the current of the laser diode which provides about 1 MHz feedback bandwidth. To stabilize this laser to a 70 m reference fiber interferometer required several layers of passive feedforward linearization and active stabilization using both the piezoelectric actuator for larger, slower deviations from linearity and direct current feedback for fast, small fluctuations from the finite linewidth of the laser. With both active and passive linearization a 5 THz chirp from this laser was linearized to less than 100 kHz rms (Barber *et al.*, 2011). This bandwidth provides less than than 50 μm resolution at large ranges (several hundred meters limited by the rms deviations). In addition, one of the important results of this work was the realization that over this large of bandwidth (~ 40 nm) the wavelength dispersion of the fiber interferometer had a great effect on the linearity of the chirp. Therefore, dispersion compensating fiber was used in the reference interferometer to minimize the wavelength dispersion. Without dispersion management, the ability to perform pulse compression with a simple FFT for longer range (greater than a couple meters) FMCW lidar was compromised (Barber *et al.*, 2010).

For use in metrology applications where high accuracy distance measurements are needed, in addition to linearizing the chirp, the chirp rate also must be accurately calibrated. To provide calibration independent of a distant standard, the chirp rate of the laser could be calibrated by measuring the time versus optical frequency against a fixed set of widely separated frequency references. Two examples that were performed on this laser were: (1) The use of the NIST calibrated hydrogen cyanide (HCN) absorption lines as absolute frequency references as in **Fig. 7**, and (2) using the comb lines of a stabilized mode-locked laser as a frequency ruler (Barber *et al.*, 2011; Giorgetta *et al.*, 2010). The former method can achieve sub part per million calibration, and the latter could achieve part per billion calibration. The optical frequency comb based method was adapted in Baumann *et al.* (2014a,b, 2013) to provide dynamic high accuracy chirp nonlinearity measurements that could be used for post-processing based correction in high range resolution FMCW lidar based 3D metrology.

80 GHz distributed feedback based chirp stabilization

A second example of an actively linearized chirp laser is one based on a distributed feedback (DFB) laser (Satyan *et al.*, 2009). DFB lasers are common in the telecommunications industry due to their robust single-mode, narrow linewidth, and stable wavelength operation. While stable relative to other diode laser types, DFB lasers can be frequency tuned relatively rapidly and reliably using the laser current with a tuning rate on the order of 1 GHz/mA. A schematic for the active linearization and stabilization of a DFB chirp laser is shown in **Fig. (8a)**. This setup is similar to that used in Ref. Roos *et al.* (2009) for the 5 THz tunable chirp laser except that the fiber delay line is shorter to better match the relatively larger linewidth of the DFB laser (~ 500 kHz) versus that of the external cavity diode laser (<50 kHz). However, this setup is more advanced in that the RF reference of the phase and frequency

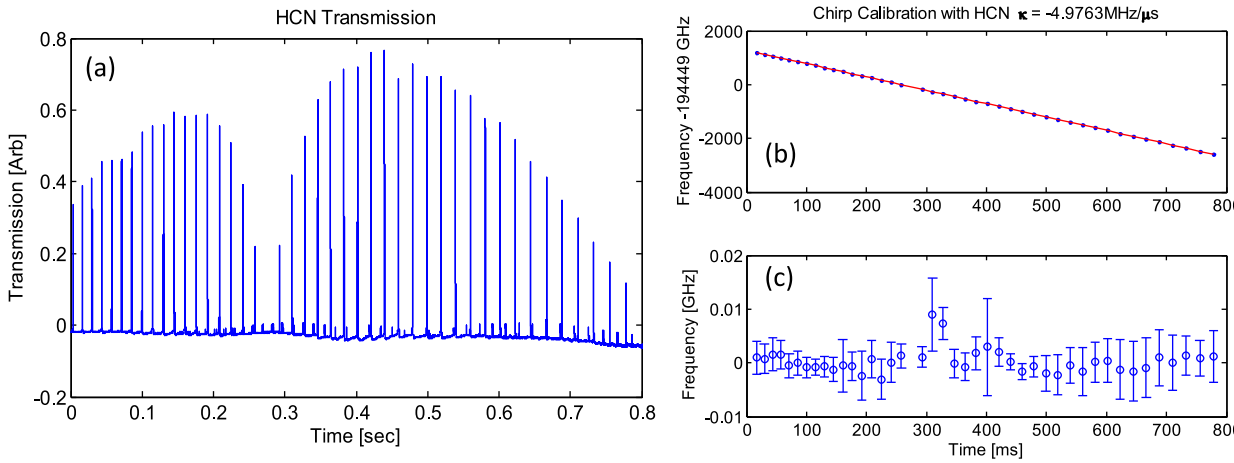


Fig. 7 (a) The transmission of a wideband chirp laser from 1530 to 1570 nm region through 10 Torr hydrogen cyanide (HCN) fiber coupled cell; (b) shows the measured HCN absorption peaks vs. chirp time use to calibrate the chirp rate and residual quadratic chirp of the 5 THz chirp laser; and (c) the residuals of the fit in (b). The error bars are the absolute frequency uncertainty provided by NIST (Gilbert *et al.*, 2005).

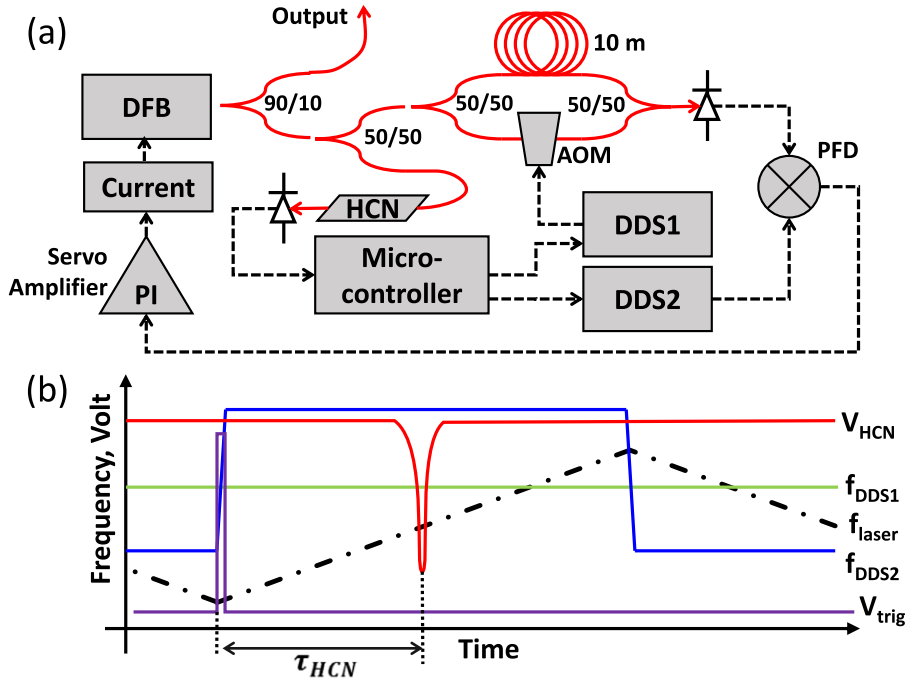


Fig. 8 (a) A schematic of a distributed feedback (DFB) chirp linearization system and center frequency stabilization based on an optical absorption line in hydrogen cyanide (HCN); and (b) a timing diagram representing how the different frequencies and voltages are used to stabilize the up-chirp, down-chirp, and the center frequency of the laser. A full description is provided in Section 80 GHz distributed feedback based chirp stabilization. DDS, direct digital synthesizer; PFD, phase and frequency detector; PI, proportional integral amplifier.

detector (PFD) used to generate the error signal for the PLL is derived from a highly controllable direct digital synthesizer (DDS) signal source. As in Roos *et al.* (2009) both the up-chirp and the return down-chirp of the laser are actively stabilized by switching the frequency of the reference to be shifted above or below the AOM frequency, respectively. However, instead of rapidly switching from one frequency to the other, here the reference frequency makes a controlled sweep between the two settings (see the f_{DDS2} line in **Fig. 8(b)**). This controlled sweep allows the PLL loop to remain locked at the turn-around point. This has the effect of making the center frequency of the chirp also stabilized to the fiber interferometer. This is in great contrast to the 5 THz laser chirp where the laser came completely unlocked at the turn-around points, leading to >100 MHz level jumps of the center frequency from chirp-to-chirp. Further investigations of the DFB laser, showed that while this locking around the corner provided good short term stability, drifts in the fiber length and possible timing instabilities in the triggering of the up and down sweeps could lead to slow drift of the center frequency of the laser. This long term drift could be controlled by slight (1 Hz level) adjustments of the

AOM frequency, but to provide truly long term absolute stability this setup implemented a slow digital feedback loop using an absorption line in a HCN spectroscopic cell as a frequency reference. The feedback loop was implemented using a microcontroller that measured the time difference, τ_{HCN}, between the start of the chirp and incidence of the absorption peak using a simple threshold comparator circuit. The AOM frequency is then adjusted through f_{DDS1} to stabilize the time difference. This stabilized center frequency is beneficial for SAL imaging or other interferometric applications that require the frequency of the chirps to be stable.

Applications of Very High Range Resolution Lidar

Aspherical optical metrology

The ability to perform lidar with resolutions at smaller than 50 μm over large distances (greater than 10 m) lends itself very naturally to think about length metrology as an application. What is more, because FMCW lidar has the ability to resolve multiple range returns, new metrology concepts can be explored. In particular, if one applies the FMCW lidar system to an optical system the distance to every surface (even if buried inside) can be measured. This is similar to optical time domain reflectometry (OTDR) systems that are used in the fiber telecommunications industry to diagnose and find faults, but on a much finer scale of microns instead of meters. Additionally, while the range resolution of the FMCW lidar system is 50 μm, the range precision and accuracy on the measurement of distance can be much smaller due to the ability to find the center of a range peak much better than its width. **Fig. 9(a)** shows the range profile of a microscope slide placed in a slightly focused beam at about 2.8 m of total range delay. The two peaks are well resolved by the ≈50 μm range resolution at about 1.5 mm apart, which is the optical thickness of the microscope slide. Both peaks have high SNR of greater than 35 dB. This high signal to noise allows a fit of the peaks to determine their centers to high precision. This can be confirmed by taking a series of measurements of the range profile. While the individual standard deviation of the fit to the individual range peaks is approximately 700 nm (not shown), the thickness as measured by subtracting the two individual range measurements show less than 50 nm standard deviation, which is approximately 1000 times less than the range resolution of an individual peak.

Extending this to three-dimensional metrology shows a potential application of very high resolution lidar to optical metrology. **Fig. 10** shows work performed to characterize highly aspherical optics. This work used an *X*–*Y* stage to make a grid of range profile

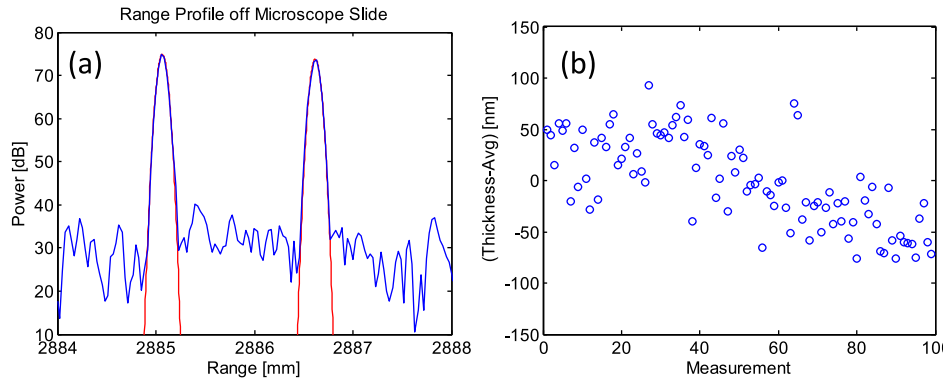


Fig. 9 (a) Range profile of the front and back surface of a microscope slide placed at about 2.8 m range delay using FMCW lidar with an ∼5 THz chirp and (b) series of optical thickness measurements taken by subtracting the range of the back and front surface. This series of measurements shows less than 50 nm standard deviation.

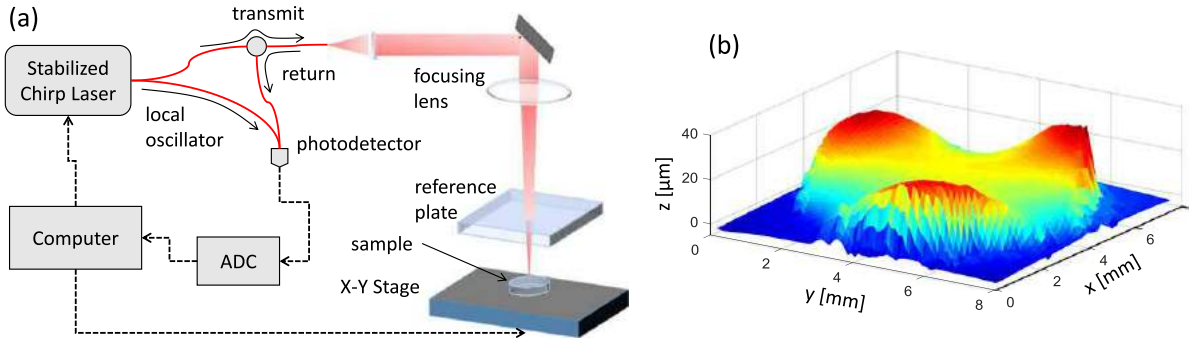


Fig. 10 (a) *X*–*Y* scanning setup to perform multi-surface profilometry and (b) 3D surface profile (vertical dimension exaggerated to show range precision) of complex aspheric optic provided by Wavesource Inc.

measurements on an optical flat whose surface was machined with a strong Zernike aberration function profile. This complex surface was machined by Wavesource Inc. as part of a research project to test the potential of FMCW lidar for optical metrology and as confirmation and testing of Wavesource Inc.'s diamond turning capability. The profile had steep slopes (nearly 50% grade at points), which make standard interferometric methods of characterization difficult. In addition to the reflection on the top surface, the high range resolution allows both the top and bottom surface to be measured simultaneously. To improve the precision and accuracy of the measurement a reference reflection near to the sample is used to remove any path length fluctuations that occur in the measurement chain. As it can be seen from **Fig. 10(b)** the ~50 μm surface elevations are highly resolved. Comparison with the predicted surface profile from the diamond turning process showed good quantitative agreement.

Noncontact 3D metrology

In addition to shorter range optical metrology, very high resolution lidar is also suited to 3D metrology applications over larger distances and volumes. In the 3D metrology space the highest accuracy volume metrology systems utilize interferometric based optical distance measurements that require some type of retro-reflective target to achieve high precision and accuracy (~25 μm) desired by the most demanding applications. The use of a retro-reflective targets forces the operator to manually contact the target to the part under measurement reducing productivity. The use of very high resolution FMCW lidar that can achieve high SNR from diffuse surfaces could allow noncontact and unsupervised metrological scanning. Indeed, 3D architectural and terrestrial lidar scanners (Petrie and Toth, 2008) have become one of the largest markets for lidars. Utilizing high resolution lidars in these systems to make more precise and accurate range measurements is an obvious application space.

In the area of 3D metrology with FMCW lidar, investigations have been made into the fundamental limits of precision when measuring the range to a diffuse surface (Baumann *et al.*, 2014a,b). One difficulty that was discovered was the influence of surface roughness and laser speckle on the FMCW ranging process that can lead to unpredictable range pulling with size on the order of the range resolution. A similar effect also can lead to pseudo-Doppler range pulling when the measurement beam moves rapidly across a surface. Despite these limits it was shown that with a 1 THz bandwidth laser these effects could be held to less than 10 μm (Baumann *et al.*, 2014a,b).

Improving the range measurement in a 3D lidar scanner also necessitates that the transverse precision and accuracy of the laser scanner is commensurate with the range measurement. While angular scanning mechanisms can be made quite accurate, difficulties still arise when making precise noncontact measurements due to the size of the incident laser beam on the target. With the high precision and accuracy available with very high resolution FMCW lidar, researchers have investigated 3D metrology methods that utilize only distance measurements to position points in 3D space (Warden *et al.*, 2015; Warden, 2014; Mateo and Barber, 2015). These approaches use similar techniques to GPS where the distance between the point of interest and a few (or several) known points in 3D space is used to determine the position of the unknown point. One difficulty with these techniques is being able to provide measurement coverage in a large volume while maintaining sufficient SNR to make precise and accurate range measurements.

Advanced coherent imaging with FMCW lidar

One of the most exciting applications for coherent FMCW lidar is in the area of coherent imaging including Synthetic Aperture (Imaging) Lidar or Ladar (SAL or SAIL; Beck *et al.*, 2005; Bashkansky, 2002; Krause *et al.*, 2011; Crouch and Barber, 2012), holographic aperture lidar (HAL; Duncan and Dierking, 2009), sparse aperture imaging, digital holography, or combinations of these (Crouch *et al.*, 2015; Krause *et al.*, 2012). These coherent imaging methods seek to exploit the phase information obtained with coherent detection methods to improve the ability to image a scene and/or reduce resource requirements in an imaging system. As a direct optical analog to common radar methods, FMCW lidar is ideal for translating synthetic aperture radar (SAR) imaging methods to the optical domain. In fact, SAL was one of the motivations for the development of the 5 THz chirp laser described in Section Broadband (5 THz) chirp stabilization. An earlier investigation into SAL (Beck *et al.*, 2005), described the lack of suitable chirp laser sources as an impediment to the development of the technique. The benefit of very high resolution lidar FMCW lidar is that relatively high resolution images, i.e., large number of pixels, can be made with table-top sized experimental setups, making the investigation of SAL technique easier (Crouch and Barber, 2012).

SAL is a 2D range/cross-range imaging technique that provides imaging resolution within the diffracted limited beam of the transmitter. In one dimension the image resolution is provided by the range resolution of the lidar system, which does not depend on diffraction and means that high range resolution provides better image resolution. It also means that the target/scene must present range diversity to the lidar receiver, usually accomplished by illuminating the target at an angle. In the other dimension, the image resolution is provided by exploiting relative motion between the target/scene and the lidar transmitter/receiver. Relative motion perpendicular to the line-of-sight provides different points of view of the target and traces out a larger effective synthetic aperture for imaging. To successfully use this synthetic aperture, the system must coherently combine the information from the many range profile measurements. This necessitates the use of coherent lidar.

A basic design and simplified signal processing architecture for strip-map SAL is shown in **Fig. 11**. In strip-map SAL the motion of the Tx/Rx aperture, which is perpendicular to the ranging direction, also scans the Tx/Rx beam across the target. This limits the maximum size of the synthetic aperture in strip-map to the size of the beam on target. The data collection and signal processing system is relatively simple. At each aperture position the coherent FMCW lidar signal, before processing with an FFT to give a range profile, is stored in a 2D matrix called the phase history data. Because no lenses or range dimension FFT is used, this phase history data is a Fourier domain projection of the scene in the range dimension and the cross-range dimension (along the direction of motion or track). In the far-field limit, the SAL image can then be formed by a 2D FFT of the phase history data. In more physical

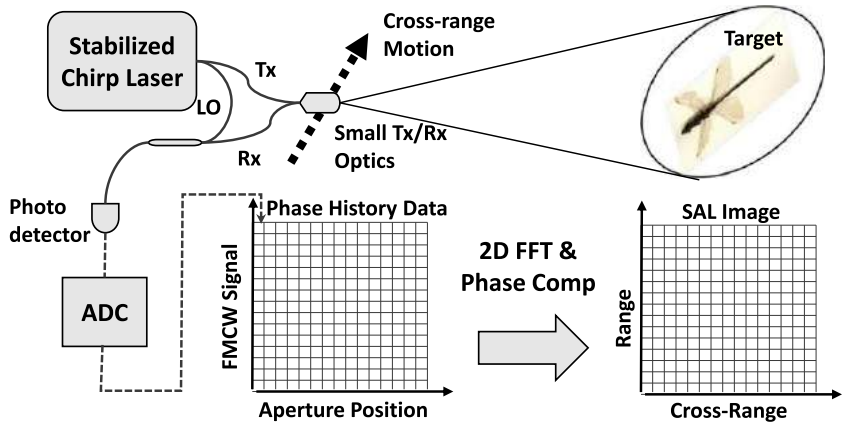


Fig. 11 Simple schematic of SAL setup and signal processing architecture.

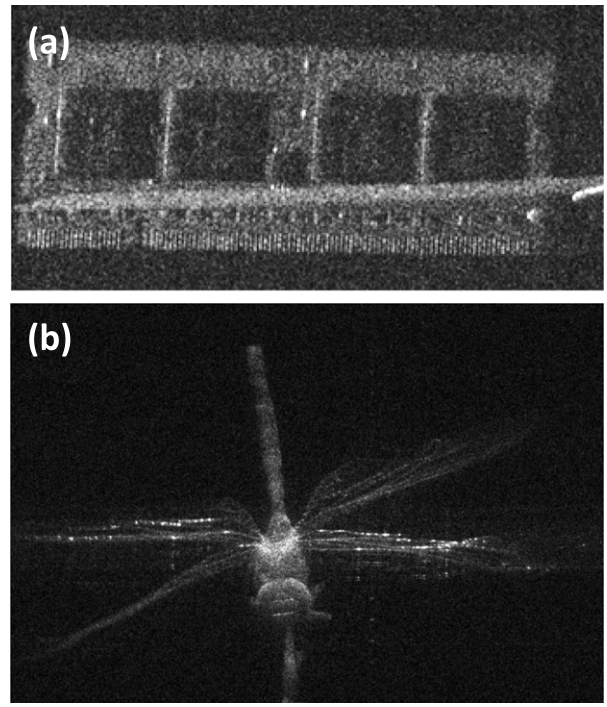


Fig. 12 Two SAL images made at ~2m range using a 5 THz chirp laser (a) is a RAM memory chip and (b) is a dried dragonfly specimen. Republished with permission from Crouch, S.C., 2012. Synthetic aperture LADAR techniques. Master's Thesis, Montana State University – Bozeman, College of Letters & Science. Available at: http://scholarworks.montana.edu/xmlui/handle/1/1125 (accessed 30.11.16).

situations, some phase compensation of the phase history data is needed first to account for not being infinitely far away. **Fig. 12** shows two high resolution images made on a table-top setup with the laser described in Section Broadband (5 THz) chirp stabilization.

An important aspect of SAL imaging is that it is a coherent process, which means that the phase of the phase history data is crucial. Yet, the phase of the FMCW signal is sensitive to the relative motion of the Tx/Rx aperture and the scene on the wavelength scale, i.e., holographically sensitive. While this would seem to make the SAL data collection task near impossible over any kind of distance or time, there are methods to compensate for the non-ideal motions. First, if there are known bright points in a scene these can be used as a phase reference. Second, algorithms developed for SAR such as the phase gradient autofocus (PGA) algorithm show remarkable ability to estimate and correct piston motion phase errors using the correlation of phase errors in the range dimension (Wahl *et al.*, 1994). The PGA algorithm has also shown to work quite well in the low SNR regime where the image signal less than an order of magnitude larger than the shot-noise background (Barber and Dahl, 2014). The proof of the feasibility for real-world SAL imaging is that air-to-ground SAL imaging has actually been demonstrated from an aircraft (Krause *et al.*, 2011), however work is still required for practical applicability.

Conclusions

The objective of this article was to introduce the reader to technologies, techniques, and applications in the area of very high range resolution (<1 cm) lidar. In general, these techniques are not brief-pulse, direct-detection lidars. They do, however, often utilize coherent lidar techniques with lasers that have unique spectral domain coherence. Clever exploitation of this spectral coherence reduces the requirements for high speed optoelectronics allowing range resolution of tens of microns – the equivalent of THz bandwidths which is far beyond current electronics. The article then focused on coherent FMCW lidar to introduce interesting applications for very high resolution lidar in metrology and coherent imaging. While precision optical distance measurement techniques for metrology have been around for a long time, the difference between these older methods and the newer lidar techniques is the ability to fully resolve a range profile, not just single ranges. With the ability to resolve range profiles, multiple returns in the path of the lidar can be measured simultaneously, which provides unique capabilities in the area of optical metrology. In addition, the high sensitivity of these coherent lidar techniques provide the ability to measure distances to diffuse surfaces precisely and accurately. Lastly, very high resolution coherent FMCW lidar has been used to demonstrate coherent imaging modalities such as synthetic aperture lidar (SAL).

Acknowledgments

This article represents my understanding of the state of the field of very high resolution lidars. I have been working in this field for only 8 years and a lot of great work had been performed before I entered the field and continues to this day. About once a month I receive peer-review requests for articles in the area of coherent lidar and the field only seems to be growing. As one person, it is difficult to keep up to date on all the advances in even this relatively narrow field, and I am sure that I overlooked much excellent work in high resolution lidar. While I decided to heap all the errors, omissions, and opinions contained in this article onto myself as sole author, none of my knowledge of the field could have been gained without the help and assistance of many people including current and former students, colleagues and coworkers, industry and government collaborators, and acquaintances.

In particular I need to acknowledge: My former students Stephen Crouch and Ana Baselga Mateo who performed the work for their thesis research that led to my understanding of the issues and applications of very high resolution FMCW lidar. My colleagues at Bridger Photonics and Blackmore Sensors and Analytics, Dr. Peter Roos and Dr. Randy Reibel who introduced me to FMCW ladar and helped me earn my first research grants in the area and much sub-award funding over the years. Wavesource Inc. for collaborating on initial optical metrology measurements by providing samples and sub-award funding. My collaborators at NIST in Boulder, CO Ian Coddington, Esther Baumann, Fabrizio Giorgetta, and Nathan Newbury for showing initial interest in this area and then by greatly pushing the field of precision metrology with very high resolution FMCW lidar. Finally, I would like to thank collaborators in the AFRL/RYMM – LADAR Technology Branch.

Additional acknowledgment must go to Nate Newbury for providing permission to adapt Figure 3 from their excellent paper on dual-comb measurement methods. Thanks is required for Dr. Steven van den Berg of the Dutch Metrology Institute for giving permission to adapt Figure 2 from an article on his excellent work in mode-locked laser based multi-heterodyne interferometry.

Parts of this work were supported by grants through the Montana Board of Research and Commercialization Technology (MBRCT), also by the National Science Foundation through GOALI grant #1031211, and by the United States Air Force through a Young Investigator Program Award.

See also: Micro-Lidars for Short Range Detection and Measurement

References

Ackerman, E., 2016. Cheap lidar: The key to making self-driving cars affordable. IEEE Spectrum: Technology, Engineering, and Science News. Available at: http://spectrum.ieee.org/transportation/advanced-cars/cheap-lidar-the-key-to-making-selfdriving-cars-affordable (accessed 30.11.16).

Ahn, T.-J., Lee, J.Y., Kim, D.Y., 2005. Suppression of nonlinear frequency sweep in an optical frequency-domain reflectometer by use of Hilbert transformation. Applied Optics 44, 7630–7634. Available at: https://www.osapublishing.org/abstract.cfm?uri=ao-44-35-7630 (accessed 30.11.16).

Barber, Z.W., Babbitt, W.R., Kaylor, B., Reibel, R.R., Roos, P.A., 2010. Accuracy of active chirp linearization for broadband frequency modulated continuous wave ladar. Applied Optics 49, 213–219. Available at: http://www.opticsinfobase.org/abstract.cfm?uri=ao-49-2-213 (accessed 26.01.15).

Barber, Z.W., Dahl, J.R., 2014. Synthetic aperture ladar imaging demonstrations and information at very low return levels. Applied Optics 53, 5531–5537. doi:10.1364/AO.53.005531.

Barber, Z.W., Giorgetta, F.R., Roos, P.A., *et al.*, 2011. Characterization of an actively linearized ultrabroadband chirped laser with a fiber-laser optical frequency comb. Optics Letters 36, 1152–1154. Available at: http://www.opticsinfobase.org/abstract.cfm?uri=ol-36-7-1152 (accessed 26.01.15).

Bashkansky, M., 2002. Synthetic aperture imaging at 1.5μ: Laboratory demonstration and potential application to planet surface studies. SPIE 4849, 48–56. doi:10.1117/12.460767.

Baumann, E., Deschênes, J.-D., Giorgetta, F.R., *et al.*, 2014a. Speckle phase noise in coherent laser ranging: Fundamental precision limitations. Optics Letters 39, 4776–4779. doi:10.1364/OL.39.004776.

Baumann, E., Giorgetta, F.R., Coddington, I., *et al*., 2013. Comb-calibrated frequency-modulated continuous-wave ladar for absolute distance measurements. Optics Letters 38, 2026–2028. doi:10.1364/OL.38.002026.

Baumann, E., Giorgetta, F.R., Deschênes, J.-D., *et al*., 2014b. Comb-calibrated laser ranging for three-dimensional surface profiling with micrometer-level precision at a distance. Optics Express 22, 24914–24928. doi:10.1364/OE.22.024914.

Beck, S.M., Buck, J.R., Buell, W.F., *et al*., 2005. Synthetic-aperture imaging laser radar: Laboratory demonstration and signal processing. Applied Optics 44, 7621–7629. doi:10.1364/AO.44.007621.

Butera, S., Vines, P., Tan, C.H., Sandall, I., Buller, G.S., 2016. Picosecond laser ranging at wavelengths up to 2.4 μm using an InAs avalanche photodiode. Electronics Letters 52, 385–386. doi:10.1049/el.2015.3995.

Cabral, A., Rebordão, J., 2007. Accuracy of frequency-sweeping interferometry for absolute distance metrology. Optical Engineering 46.doi:10.1117/1.2754308.

Coddington, I., Swann, W.C., Nenadovic, L., Newbury, N.R., 2009. Rapid and precise absolute distance measurements at long range. Nature Photonics 3, 351–356. doi:10.1038/nphoton.2009.94.

Coddington, I., Swann, W.C., Newbury, N.R., 2010. Coherent dual-comb spectroscopy at high signal-to-noise ratio. Physical Review A 82, 043817. doi:10.1103/PhysRevA.82.043817.

Crouch, S., Barber, Z.W., 2012. Laboratory demonstrations of interferometric and spotlight synthetic aperture ladar techniques. Optics Express 20, 24237. doi:10.1364/OE.20.024237.

Crouch, S., Kaylor, B.M., Barber, Z.W., Reibel, R.R., 2015. Three dimensional digital holographic aperture synthesis. Optics Express 23, 23811. doi:10.1364/OE.23.023811.

De Groot, P., McGarvey, J., 1992. Chirped synthetic-wavelength interferometry. Optics Letters 17, 1626–1628. Available at: https://www.osapublishing.org/abstract.cfm?uri=ol-17-22-1626 (accessed 30.11.16).

Delfyett, P.J., Mandridis, D., Piracha, M.U., *et al*., 2012. Chirped pulse laser sources and applications. Progress in Quantum Electronics 36, 475–540. doi:10.1016/j.pquantelec.2012.10.001.

Duncan, B.D., Dierking, M.P., 2009. Holographic aperture ladar. Applied Optics 48, 1168–1177. Available at: http://www.opticsinfobase.org/ao/fulltext.cfm?uri=ao-48-6-1168&id=176694 (accessed 23.04.15).

Durán, V., Andrekson, P.A., Torres-Company, V., 2016. Electro-optic dual-comb interferometry over 40 nm bandwidth. Optics Letters 41, 4190. doi:10.1364/OL.41.004190.

Fujima, I., Iwasaki, S., Seta, K., 1998. High-resolution distance meter using optical intensity modulation at 28 GHz. Measurement Science and Technology 9, 1049. Available at: http://iopscience.iop.org/article/10.1088/0957-0233/9/7/007/meta (accessed 30.11.16).

Gelbart, A., Redman, B.C., Light, R.S., Schwartzlow, C.A., Griffis, A.J., 2002. Flash lidar based on multiple-slit streak tube imaging lidar. SPIE 4723, 9–18. doi:10.1117/12.476407.

Gilbert, S.L., Swann, W.C., Wang, C.-M., 2005. Hydrogen cyanide H13C14N absorption reference for 1530 nm to 1565 nm wavelength calibration–SRM 2519a. NIST Special Publication 260, 137. Available at: http://132.163.4.18/srm/upload/SP260-137.pdf (accessed 9.02.15).

Giorgetta, F.R., Coddington, I., Baumann, E., Swann, W.C., Newbury, N.R., 2010. Fast high-resolution spectroscopy of dynamic continuous-wave laser sources. Nature Photonics 4, 853–857. doi:10.1038/nphoton.2010.228.

Greiner, C., Boggs, B., Wang, T., Mossberg, T.W., 1998. Laser frequency stabilization by means of optical self-heterodyne beat-frequency control. Optics Letters 23, 1280–1282. doi:10.1364/OL.23.001280.

Hamamatsu 2008. Guide to Streak Cameras, Hamamatsu Photonics K.K. Available at: http://www.hamamatsu.com/resources/pdf/sys/SHSS0006E_STREAK.pdf (accessed 31.10.16).

Huber, R., Wojtkowski, M., Fujimoto, J.G., Jiang, J.Y., Cable, A.E., 2005. Three-dimensional and C-mode OCT imaging with a compact, frequency swept laser source at 1300 nm. Optics Express 13, 10523–10538. doi:10.1364/OPEX.13.010523.

Iddan, G.J., Yahav, G., 2001. Three-dimensional imaging in the studio and elsewhere. SPIE 4298, 48–55. doi:10.1117/12.424913.

Khoshelham, K., Zlatanova, S., 2016. Sensors for indoor mapping and navigation. Sensors 16.doi:10.3390/s16050655.

Knight, F.K., Klick, D., Ryan-Howard, D.P., *et al*., 1989. Laser radar reflective tomography utilizing a streak camera for precise range resolution. Applied Optics 28, 2196. doi:10.1364/AO.28.002196.

Krause, B.W., Buck, J., Ryan, C., *et al*., 2011. Synthetic aperture ladar flight demonstration. In: Conference on Lasers and Electro-Optics, CLEO, IEEE, pp. 1–2.

Krause, B.W., Tiemann, B.G., Gatt, P., 2012. Motion compensated frequency modulated continuous wave 3D coherent imaging ladar with scannerless architecture. Applied Optics 51, 8745–8761. Available at: http://www.osapublishing.org/abstract.cfm?uri=ao-51-36-8745 (accessed 2.08.15).

Liu, T.-A., Newbury, N.R., Coddington, I., 2011. Sub-micron absolute distance measurements in sub-millisecond times with dual free-running femtosecond Er fiber-lasers. Optics Express 19, 18501. doi:10.1364/OE.19.018501.

Mateo, A.B., Barber, Z.W., 2015. Multi-dimensional, non-contact metrology using trilateration and high resolution FMCW ladar. Applied Optics 54, 5911. doi:10.1364/AO.54.005911.

Nakajima, Y., Minoshima, K., 2015. Highly stabilized optical frequency comb interferometer with a long fiber-based reference path towards arbitrary distance measurement. Optics Express 23, 25979. doi:10.1364/OE.23.025979.

Petrie, G., Toth, C., 2008. Terrestrial laser scanners. In: Shan, J., Toth, C.K. (Eds.), Topographic Laser Ranging and Scanning. Boca Raton, FL: CRC Press, pp. 87–128. Available at: http://www.crcnetbase.com/doi/abs/10.1201/9781420051438.ch3 (accessed 30.11.16).

Roos, P.A., Reibel, R.R., Berg, T., *et al*., 2009. Ultrabroadband optical chirp linearization for precision metrology applications. Optics Letters 34, 3692–3694. Available at: http://www.opticsinfobase.org/abstract.cfm?uri=OL-34-23-3692 (accessed 26.01.15).

Satyan, N., Vasilyev, A., Rakuljic, G., Leyva, V., Yariv, A., 2009. Precise control of broadband frequency chirps using optoelectronic feedback. Optics Express 17, 15991–15999.

Shirasaki, M., 1996. Large angular dispersion by a virtually imaged phased array and its application to a wavelength demultiplexer. Optics Letters 21, 366–368. Available at: https://www.osapublishing.org/abstract.cfm?uri=ol-21-5-366 (accessed 1.12.16).

Spühler, G.J., Paschotta, R., Fluck, R., *et al*., 1999. Experimentally confirmed design guidelines for passively Q-switched microchip lasers using semiconductor saturable absorbers. Journal of the Optical Society of America B 16, 376. doi:10.1364/JOSAB.16.000376.

Stove, A.G., 1992. Linear FMCW radar techniques. IEE Proceedings F – Radar Signal Process 139, 343–350. doi:10.1049/ip-f-2.1992.0048.

van den Berg, S.A., Persijn, S.T., Kok, G.J.P., Zeitouny, M.G., Bhattacharya, N., 2012. Many-wavelength interferometry with thousands of lasers for absolute distance measurement. Physical Review Letters 108, 183901. doi:10.1103/PhysRevLett.108.183901.

van den Berg, S.A., van Eldik, S., Bhattacharya, N., 2015. Mode-resolved frequency comb interferometry for high-accuracy long distance measurement. Science Reports 5, 14661. doi:10.1038/srep14661.

Wahl, D.E., Eichel, P.H., Ghiglia, D.C., Jakowatz, C.V., 1994. Phase gradient autofocus – A robust tool for high resolution SAR phase correction. IEEE Transactions on Aerospace and Electronic Systems 30, 827–835. doi:10.1109/7.303752.

Walden, R.H., 1999. Analog-to-digital converter survey and analysis. IEEE Journal on Selected Areas in Communications 17, 539–550. doi:10.1109/49.761034.

Warden, M.S., 2014. Precision of frequency scanning interferometry distance measurements in the presence of noise. Applied Optics 53, 5800. doi:10.1364/AO.53.005800.

Warden, M.S., Campbell, M., Hughes, B., Lewis, A., 2015. GPS Style Position Measurement With Optical Wavelengths. Washington, DC: OSA.

Wu, H., Zhang, F., Liu, T., Balling, P., Qu, X., 2016. Absolute distance measurement by multi-heterodyne interferometry using a frequency comb and a cavity-stabilized tunable laser. Applied Optics 55, 4210. doi:10.1364/AO.55.004210.

Wu, G., Zhou, Q., Shen, L., *et al*., 2014. Experimental optimization of the repetition rate difference in dual-comb ranging system. Applied Physics Express 7, 106602. doi:10.7567/APEX.7.106602.

Yang, H., Wu, L., Wang, X., *et al*., 2012. Signal-to-noise performance analysis of streak tube imaging lidar systems I Cascaded model. Applied Optics 51, 8825. doi:10.1364/AO.51.008825.
Zhang, H., Wei, H., Wu, X., Yang, H., Li, Y., 2014. Absolute distance measurement by dual-comb nonlinear asynchronous optical sampling. Optics Express 22, 6597. doi:10.1364/OE.22.006597.

Relevant Websites

https://www.wired.com/2015/09/laser-breakthrough-speed-rise-self-driving-cars/
A.C.M.C.M. Business, Laser Breakthrough Could Speed the Rise of Self-Driving Cars, WIRED.
http://web.stanford.edu/∼murmann/adcsurvey.html
Boris Murmann: ADC Survey.
http://www.radartutorial.eu/01.basics/Range%20Resolution.en.html
Radar Basics: Range Resolution.
https://en.wikipedia.org/w/index.php?title=Time-of-flight_camera&oldid=751907785
Time-of-Flight Camera.

Multi-Dimensional Laser Radars

Vasyl V Molebny, Academy of Technological Sciences of Ukraine, Kiev, Ukraine

Introduction

Despite the word "ranging" in the term "laser detection and ranging" implies acquiring the information on range, the other parameters measured by laser radar can have its own dimensionality, for example, object orientation in space and its velocity, light scattering in the media, refraction of the media, temperature, humidity, gases concentration, etc. Some of them are single-dimensional, like temperature. Some can have several dimensions, like temperature distribution in space or temperature dependence on time. Other types of examples are combinations of several parameters distributed in space and time, like 3D velocity of an object having its 3D position and 3D orientation in space, both changing in time, thus climbing up to a hypothetical ten-dimensional (10D) laser radar system. Extending the popular term "3D (three-dimensional)", Molebny and Steinvall (2014) introduced the term "multi-dimensional" meaning that the laser radar outputs the information defined by three or more independent descriptors, which can be regarded as dimensions. This publication will serve as a canvas for our article. Examples of information delivered by three-dimensional laser radars are illustrated in **Table 1**.

Examples of multi-dimensional laser radar systems are illustrated in **Table 2**.

Measurement of any of the dimensions can be provided with a single transmitter and a single receiver or with several of each of them. Moreover, there could be a combination of different laser radar systems or even a combination of laser radars with passive infrared systems or with microwave radars. Last decade brought hundreds of studies and corresponding publications on a variety of designs, applications, theories, experiments and field tests. We report here some representatives of them.

Stepping up the Dimensionality

The number of dimensions acquired by laser radar can be increased when adding new functions. For example, scanning the optical axis of a simple rangefinder and storing the measured range at different positions of the optical axis one can reconstruct the profile of the target thus getting its three-dimensional image.

The range information for reconstruction of the 3D image can be got not only by changing the direction of the optical axis sequentially in time (scanning), but by illuminating (flashing) the space in a wide field of view and receiving the information from all directions simultaneously.

In the first case, the range dimension is supplemented with two new dimensions (x, y) or (α, ϵ). In the second case, the two-dimensional image is supplemented with the range information. Here are the examples of these two approaches.

Table 1 Three-dimensional description of the parameters acquired by a laser radar

D_1	D_2	D_3	*Comments*
x	x	R	Range (R) in orthogonal (x, y, R) system of co-ordinates
x, α	y, ε	R, d	Range (R) or depth (d) in orthogonal (x, y, R), (x, y, d), or spherical (α, ε, R), (α, ε, d), system of co-ordinates
α	ε	V_r	Radial velocity in spherical (α, ε, V_r) system of co-ordinates
x, α	y, ε	A, f	Distribution of amplitude (A) or frequency (f) of vibrations on the target surface
f	t	A	Frequency (f) and amplitude (A) of vibrations of a point on the target surface in time (t)
x, α	y, ε	β, c, h, n, T, p	Distribution in space of scatter (β), gas concentration (c), humidity (h), refraction (n), temperature (T), polarization (p)

Table 2 Examples of multi-dimensional description of laser radar information (4D and more)

D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}	*Comments*
α	ε	V_r	t							Radial velocity varying in time
x	y	T	V_x	V_y						Temperature and velocity distribution in a plane
x	y	z	V_x	V_y	V_z					Velocity distribution in a volume
x	y	z	V_x	V_y	V_z	t				Velocity distribution in the volume varying in time
α	ε	R	V_χ	V_η	V_ζ	α_χ	ϵ_η	$(\varphi)_\zeta$	t	Range, velocity and object orientation varying in time

 doi:10.1016/B978-0-12-803581-8.09425-X

Ranging Supplemented by Two-Dimensional Scanning

One of the very first 3D laser radars developed under NASA contract used two-axis piezoelectric driven mirror scanners providing 350×350 resolvable scan elements (Flom, 1972). These elements, for both the transmitter and the receiver, were aligned and their positions were electronically controlled, thereby creating a synchronously scanned transmitter-receiver in a total search field of view 30×30 degrees. The wavelength of the GaAs laser was 0.9 μm. Pulse repetition rate was controlled depending on the range to the target.

In 1960s, the antimissile defense (AMD) was a strategic problem at both sides of Atlantic. To improve the precision of the AMD radars, an experimental laser radar was built and tested (Molebny *et al.*, 2010) with impressive dimensions and parameters. It consisted of 196 lasers and the same number of range-gated photomultipliers (4 groups of 7×7 arrays, totaling a 14×14 array). Initial target designation from microwave radar was got with an error ± 3.5 arc. min. Therefore, the total field of view (7×7) arc. min. was split into (100×100) arc. sec. sub-zones, to be illuminated in sequence. Each ruby laser pulse had energy of 1 J pulse repetition frequency 10 Hz, pulse duration 30 ns.

In 1980s, an imaging CW heterodyne ladar was tested (Wang *et al.*, 1984). The CO_2 laser produces a 5 W beam at 10.6 μm. The beam is split into a signal beam and a local oscillator beam. The signal beam is frequency shifted, intensity modulated at 15 MHz, and scanned across the field of view by a galvanometer scanner controlled by a microcomputer. The scanner covers 512 lines in a (12×12)-deg. field of view. From the received signal, 15 MHz is filtered, amplified, and detected to produce amplitude and phase data. These data are then recorded on a magnetic tape. The recorded data are read by a computer in the laboratory and presented on a video monitor. Reflectance has a speckled appearance, horizontal stripes in the range images represent 10 m ambiguity steps produced by the amplitude modulated ranging system. Range resolution was about 0.3 m at a range of 50 m.

Two-Dimensional Imaging Supplemented by Scanning in *z* Direction

The technology of scanning in z direction can be implemented by range gating – a technique widely used in microwave radars. In the case of laser radar, pulses of picosecond duration allow to section the space in very thin "slices". High-speed cameras with a highly sensitive CCD and short gate times provide gated viewing with automatic gain control versus range, whereby foreground backscatter can be suppressed. Compensation can be automated for decreasing reflection with range as z (Flom, 1972), or an exponential damping of the light in absorbing media. A 3D laser radar prototype implementing this technology was reported (Busck and Heiselberg, 2004). The system uses a picosecond Q-switched Nd:YAG laser at 532 nm with a 32 kHz pulse repetition frequency, which triggers a (752×582)-pixel CCD camera.

Developing the idea of cutting the range in slices, a set of image slices at different distances can be reconstructed. Some of these slices can be removed (e.g., obscuring objects, camouflage, clutter, etc.). In one of the approaches of such processing, the first step consists in selecting the distances of interest in the histogram, for example, distances around the ranges of 2.5, 7.5, and 16.0 m (**Fig. 1**), where maximal numbers of pixels are concentrated. Additional processing can be made involving a 3D mesh to analyze the cloud points both in (x, y) and z directions. Processing for laser gated viewing was developed by TNO Defence, Security and Safety (Netherlands) (Bovenkamp and Schutte, 2010) based on Intevac laser gated viewer. 1.5-μm eye safe laser was used and a CMOS camera for gated viewing (minimum gate shift was 1.0 m, minimum gate width -20 m). The results of image processing are demonstrated in **Fig. 2**.

Intevac's Electron Bombarded Active Pixel Sensor (EBAPS) technology is based on a III-V semiconductor photocathode in proximity-focus with a high resolution, backside-thinned, CMOS chip anode. The electrons emitted by the photocathode are

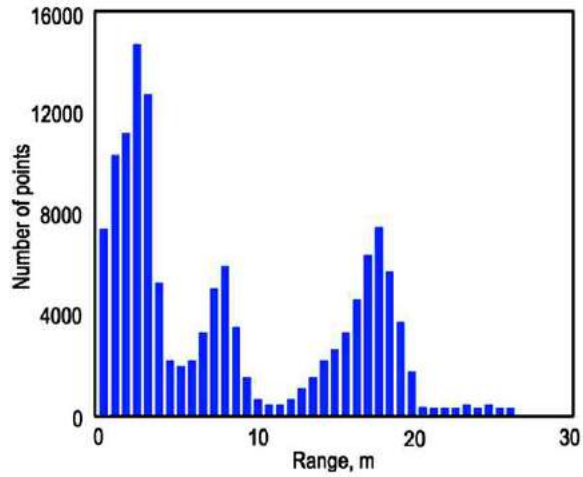


Fig. 1 Histogram of range image. Author's courtesy Bovenkamp, E.G.P., Schutte, K., 2010. Laser gated viewing: An enabler for automatic target recognition. Proc. SPIE7684, 768435.

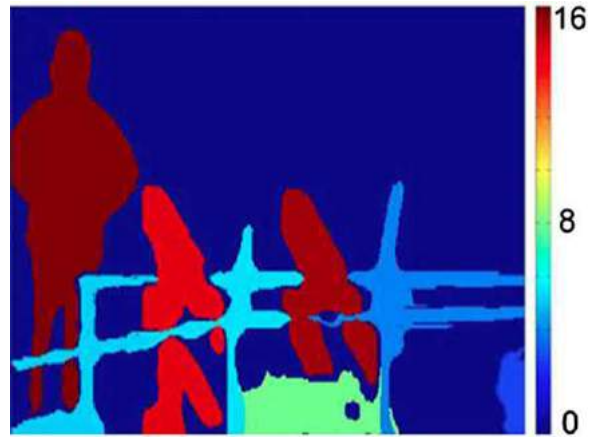


Fig. 2 2D projections of objects at different distances. Reproduced from Bovenkamp, E.G.P., Schutte, K., 2010. Laser gated viewing: An enabler for automatic target recognition. Proc. SPIE7684, 768435.

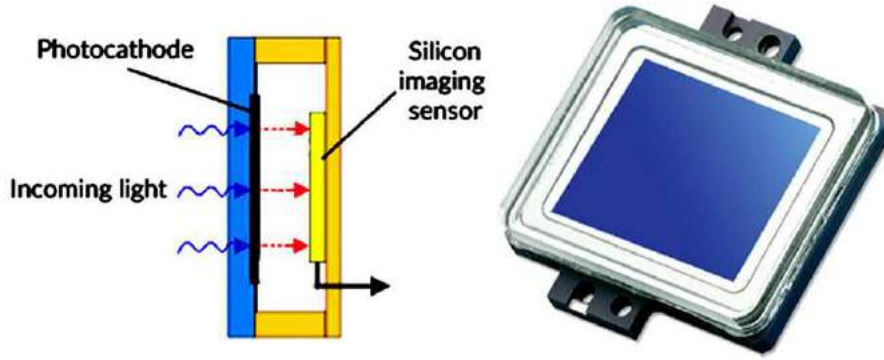


Fig. 3 Intevac's electron bombarded active pixel sensor (EBAPS). Reproduced from Electron Bombarded Active Pixel Sensor (EBAPS). Available at: http://www.intevac.com.

directly injected in the electron bombarded mode into the CMOS anode, where the electrons are collected, amplified and read-out to produce digital video directly out of the sensor (Electron Bombarded Active Pixel Sensor (EBAPS)) (**Fig. 3**).

It is the further development of the hybrid photomultiplier tube sensitive from 900 to 1300 nm referred to as an intensified photodiode (Bradbury *et al.*, 1997). It uses an active electron photocathode and a GaAs Schottky avalanche photodiode as an anode. Its first stage amplification consists of a very low noise electron-hole multiplication. This is then followed by traditional APD multiplication. These two processes are sufficient to overcome preamplifier noise, and the photodetector is therefore capable of photon counting.

Another option of 3D imaging with laser radar is the illumination of the whole field of view with one or small number of short pulses. This technology called "flash ladar" is based on range-gated imaging. **Fig. 4** illustrates the time "slices" 1, 2, 3, …, 20 operated by a (128 × 128) FPA (focal plane array) with a ROIC (read out and integration circuit). This flash ladar was developed by Northrop Grumman Aerospace Systems (Wong *et al.*, 2010). It can store an image every 0.5 ns. The ROIC captures 20 time lapsed images for each pulse which are then used to calculate the range estimates of the target area. Laser flash lamp pumps Nd: YAG OPO, 1.57 μm, 50 mJ, 6.7 ns, 30 Hz. Time per slice is 2.2 ns.

Advanced Scientific Concepts (ASC) developed the TigerEye flash 3D ladar video camera and DragonEye space camera (Stettner, 2010) that can assist almost any manned or unmanned vehicle with collision avoidance, navigation, etc. Each of 128 × 128 pixels is "triggered" independently, allowing capture of 16,384 range data points to generate a 3D point cloud image up to 30 frames per second.

Selex developed advanced infrared detectors for multimode active and passive imaging applications (Baker *et al.*, 2008) that can be electronically switched from a thermal imaging function to a laser gated imaging function so that passive thermal imaging, solar flux imaging and laser-gated imaging can be designed into one electro-optic system. The wavelength sensitivity of these detectors (APD arrays in HgCdTe) ranges from 1 μm to 4.5 μm which is promising for active multispectral imaging in the SWIR and MWIR regions.

Another promising technology of 3D imaging developed at Lincoln Lab (Albota *et al.*, 2002a,b), is photon counting. The technique is based on a 2D detector array. Rather than measuring intensity, these detectors measure the photon time-of-arrival.

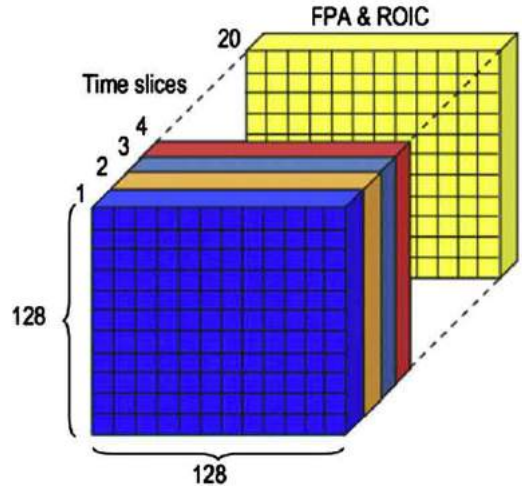


Fig. 4 Image building in a flash ladar. Reproduced fromWong, C.M., Logan, J.E., Bracikowski, C., Baldauf, B.K., 2010. Automated in-track and cross-track airborne flash LADAR image registration for wide-area mapping. Proc. SPIE7684, 768427.

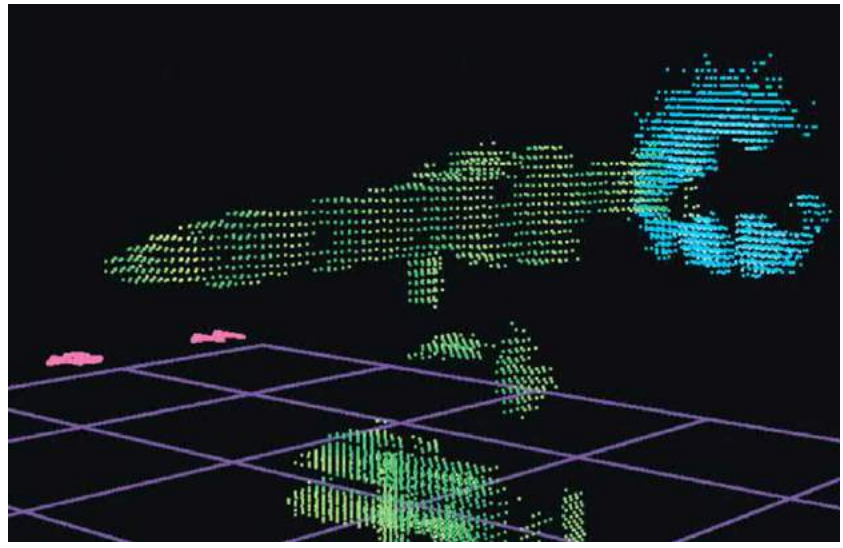


Fig. 5 Ladar reconstructed image of a missile model with a 32×32 APD array, 16 kHz repetition rate. Reproduced from Cho, P., Anderson, H., Hatch, R., Ramaswami, P., 2006. Real-time 3D ladar imaging. Lincoln Lab. J.16 (1), 147–164.

With each pixel coded in range, the ladar produces a 3D image (angle-angle-range). The illuminator is a passively Q-switched laser. The electro-optical receiver is a 4×4 array of avalanche photodiodes (APDs) that is passively quenched and gated to operate in photon-counting (Geiger) mode. Each of the 16 APD pixels in the detector array is linked to timing circuitry. The field of view of the 4×4 array is scanned to generate larger images with up to 128×128 pixels.

Doubled Nd:YAG laser microchip generates 250 ps pulses at 532 nm. It is pumped with 1.2 W at 808 nm. Repetition rate is 1 kHz, output energy per pulse is 3 μJ. This energy is enough to image the targets nearer than 1 km. 3D images at ranges beyond 1 km with a 30 μJ per pulse source were also reported. In response to a single photoelectron, Geiger-mode APDs yield a fast, high-amplitude electrical pulse that triggers the timing circuitry. The next step at Lincoln Lab was designing a 32×32 APD array with 16 kHz repetition rate (Cho *et al.*, 2006). Single photon counting and linear mode photon counting made great progress in the last decade. An example of 3D imaging is demonstrated in **Fig. 5**.

For operation in the short-wavelength infrared (1.06 pm), arrays of InP-based avalanche photodiodes were developed in the 128×32 format (Verghese *et al.*, 2009).

One of the perspective applications of laser radar is identification capability based on measuring the range profiles reconstructed in the form of histograms from a combination with a detection sensor. The detection sensor gives the location of the threat, while the laser range profiler (LRP) can identify the target based on high range-resolution data. Heuvel van den *et al.* (2008) tested the LRP with the Infrared Search and Track (IRST) sensor on board of a naval vessel.

It was shown that range profiles of ships can be obtained up to a range of 10 km. The range profiles have a good correspondence with the geometry of the ship and do not vary rapidly with aspect angle like microwave radar range profiles. **Fig. 6** shows the LRP prototype used for the MCG/8 NATO trial in Stavanger, Norway. **Fig. 7** shows the 3D model of a frigate illuminated from the front side. The model was used to calculate the range profile. An example of the reconstructed range profile at a distance 3.6 km is shown in **Fig. 8**.

The profiler gives the range and can discriminate between false alarms and potential threats. False alarms that can limit IRST usefulness like birds, cloud edges, and sun glints can be easily discriminated from real targets by a laser range profiler.

Fig. 6 Laser Range Profiler for the MCG/8 trial. Reproduced from Heuvel van den, J.C., Pace, P., Bekman, H.H., *et al.*, 2008. Experimental validation of ship identification with a laser range profiler. Proc. SPIE 6950, 69500V.

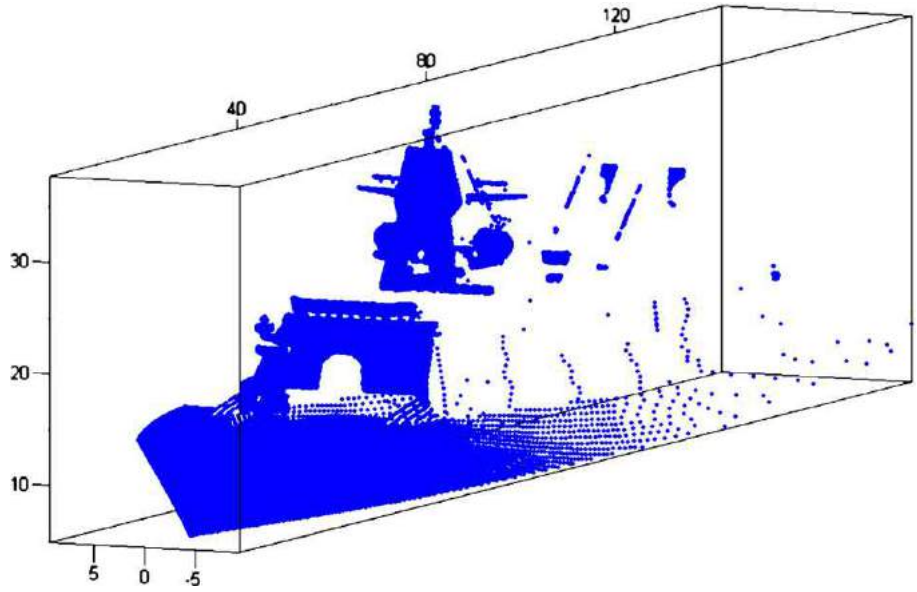


Fig. 7 Plot of data points used to generate a range profile. Reproduced from Heuvel van den, J.C., Pace, P., Bekman, H.H., *et al.*, 2008. Experimental validation of ship identification with a laser range profiler. Proc. SPIE 6950, 69500V.

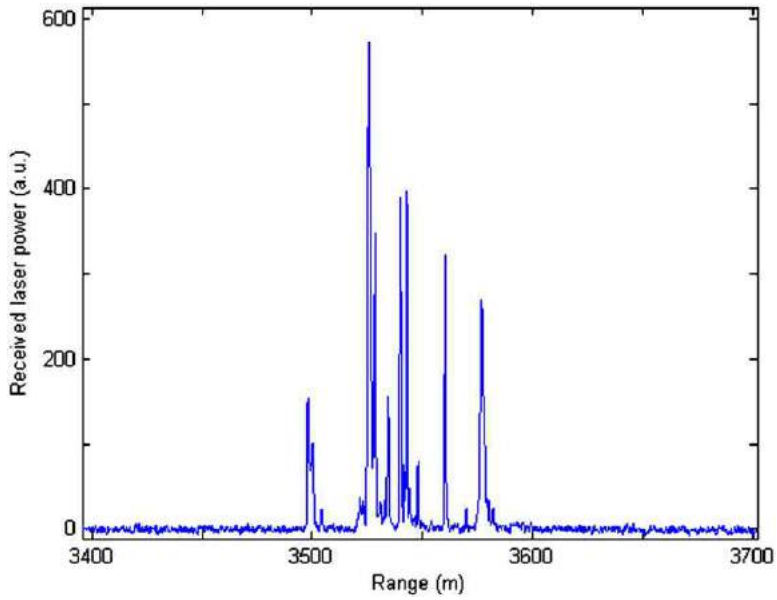


Fig. 8 Laser range profile of a ship. Reproduced from Heuvel van den, J.C., Pace, P., Bekman, H.H., *et al.*, 2008. Experimental validation of ship identification with a laser range profiler. Proc. SPIE 6950, 69500V.

Fig. 9 Marine target sensing testbed with 1550-nm laser and two receivers. Below is optics for two cameras and Celestron telescope. Reproduced from Steinvall, O., Tulldahl, M., 2017. Laser range profiling for small target recognition. Opt. Eng. 56 (3), 031206.

The LRP has the favorable properties of a good range resolution in combination with low sea-surface clutter. These properties have been validated by experiments: a range resolution of 0.6 m has been achieved. Sea-surface reflection was negligible.

Steinvall and Tulldahl (2017) described similar technique for identification of small targets. They used an LRP based on a laser with a pulse width of 6 ns. They showed both experimental and simulated results for laser range profiling of small boats out to a 6–7-km range and a unmanned aerial vehicle (UAV) mockup at close range (1.3 km). The naval experiments took place in the Baltic Sea using many other active and passive electro-optical sensors in addition to the profiling system (**Fig. 9**). The UAV experiments showed the need for a high-range resolution, thus a photon counting system in addition to the more conventional profiler was used in the naval experiments. The typical resolution of 0.7 m obtainable with a conventional range finder type of sensor can be used for target classification with a depth structure over 5–10 m or more, but for smaller targets such as an UAV a high resolution (in the described case 7.5 mm) is needed to reveal depth structures and surface shapes. **Fig. 10** shows an example of range profile from the boat at 0-deg course. The large peak is attributed to the fore and the next to the front of the cabin at a 2.5-m distance from the fore. The extent of the pulse shows a boat length exceeding 7 m.

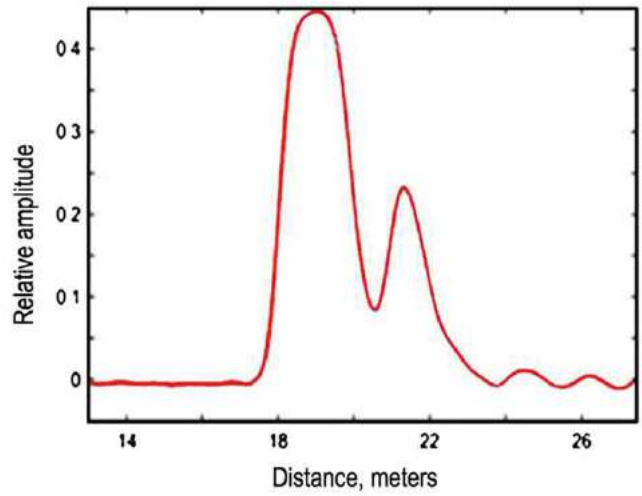


Fig. 10 The waveform referred as the pixel number when sweeping along the line of sight. Reproduced from Steinvall, O., Tulldahl, M., 2017. Laser range profiling for small target recognition. Opt. Eng. 56 (3), 031206.

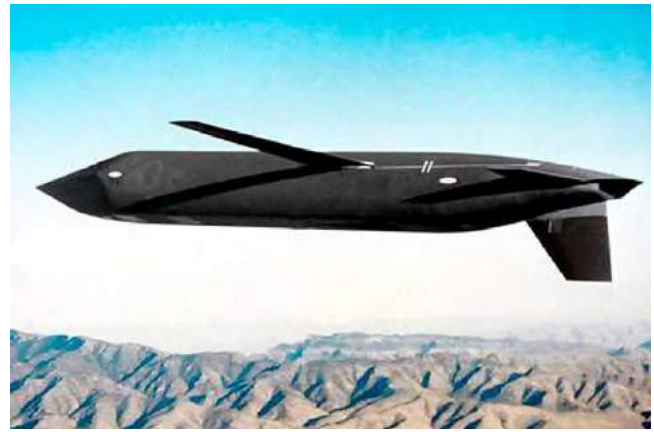


Fig. 11 Subsonic advance cruise missile AGM 129A. Reproduced from Directory of U.S. Military Rockets and Missiles. Raytheon (General Dynamics) AGM-129 ACM. Available at: http://www.designation-systems.net/dusrm/m-129.html.

Adding the Velocity as a Dimension

Precision navigation to the designated target or to any other landing site on Earth or extraterrestrial objects, rendezvous and docking with orbiting spacecraft require more information than range and/or altitude. The solutions for navigation can be flying on a preprogrammed route, via image recognition of a series of radar photographs of the terrain or surface altitude profile. Early variants of the Tomahawk (Brigety, 2007) are equipped with an inertial navigation system (INS), Terrain Contour Mapping (TERCOM), and the Digital Scene Matching Area Correlator (DSMAC). TERCOM uses a highly sensitive radar altimeter embedded in the missile to identify surface features by their height. In the terminal phase, the weapon requires even more accurate guidance than TERCOM can provide. For this reason, Tomahawk relies on DSMAC which uses an optical sensor to scan the ground over which the missile is flying. The on-board computer compares the acquired image to the stored one and adjusts the missile's course so it is following the preplanned route. The missile uses also inertial navigation to update the navigation points.

A modified approach was implemented in the subsonic advance cruise missile (ACM) AGM 129A (**Fig. 11**). Its external shape was optimized for low observable characteristics. To provide a low detectability, the microwave radar was excluded from the set of sensors. For guidance, the AGM 129A uses an inertial navigation system together with a TERCOM enhanced by highly accurate speed updates from laser Doppler velocimeter. The accuracy is quoted between 30 and 90 m.

For the extraterrestrial applications, a Doppler lidar was developed by NASA under the Autonomous Landing and Hazard Avoidance Technology (ALHAT) project (Pierrottet *et al.*, 2009; Amzajerdian *et al.*, 2012). The lidar precision vector velocity data enable the navigation system to continuously update the vehicle trajectory toward the landing site. The ALHAT project included

tests on the Morpheus lander (Fig. 12) developed not to fly in space but to demonstrate technologies that could support future human or robotic exploration (Morpheus lander).

The prototype of Doppler lidar in ALHAT project measures three components of the velocity of which the velocity vector is calculated. A fiber laser is used to generate a very narrow line width, which is frequency modulated and amplified. The output is split into three components in order to distribute the power to three optical channels corresponding to the velocity vector components (Fig. 13).

Using the laser waveform modulated linearly with time the lidar obtains high-resolution range and velocity information. Fig. 14 shows the waveform's frequency content versus time, and the resulting intermediate frequency that contains the range and velocity information. The lidar design uses an optical homodyne receiver configuration, in which a portion of the transmitted beam serves as the reference local oscillator for the optical receiver. The prototype was preliminary tested on the Eurocopter AS350D helicopter. Its optical head is mounted inside a gimbal spherical shroud, pointing in the nadir position while collecting the data.

Interference of oppositely chirped pulses was exploited for simultaneous, range and Doppler velocity measurements (Henderson *et al.*, 1993). In this instrument, echo signals are Doppler shifted in frequency and are also delayed in time relative to

Fig. 12 Morpheus lander in the last moments of landing with ALHAT laser radar instruments. Reproduced from Morpheus launder. Available at: https://morpheuslander.jsc.nasa.gov.

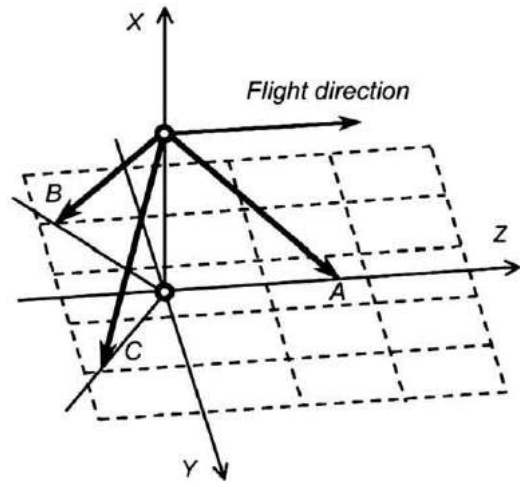


Fig. 13 Unit vectors describing the sensor geometry. Reproduced from Pierrottet, D., Amzajerdian, F., Petway, L., *et al.*, 2009. Flight test performance of a high precision navigation Doppler lidar. Proc. SPIE 7323, 732311.

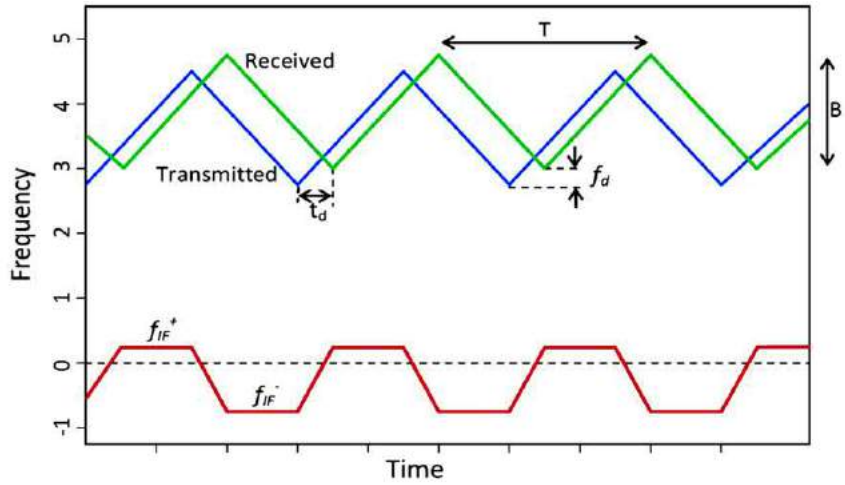


Fig. 14 Linearly modulated laser frequency (top). The difference of transmitted and received signals (lower trace) contains both range and velocity information. Reproduced from Pierrottet, D., Amzajerdian, F., Petway, L., *et al.*, 2009. Flight test performance of a high precision navigation Doppler lidar. Proc. SPIE 7323, 732311.

Fig. 15 SAVIS laser Doppler vibrometer. Reproduced from Lutzmann, P., Göhler, B., Putten, F., Hill, C.A., 2011. Laser vibration sensing: Overview and applications. Proc. SPIE 8186, 818602.

the reference pulse trains. This results in the generation of up and down beat tones. Sub millimeter resolution ranging was performed and simultaneous, range and Doppler velocity measurements were experimentally demonstrated using a target moving faster than 330 km/h.

Vibrations

The detection of vibrations at long range is of interest in a number of applications: different kinds of operating mechanisms (civil and military), civil infrastructure (bridges, dams, etc.), security (detection of acoustic signals induced in objects nearby), medicine, and so on.

Laser vibrometers typically use the Doppler effect and the coherent detection at any optical wavelength (1.5 μm, 2 μm, 10.6 μm). To get spatially resolved vibrational information, scanning and even multi-beam lasers are used. The review papers of Lutzmann *et al.* (2011), (2017) are focused on works at Fraunhofer IOSB target classification and identification, including camouflaged or partly concealed targets, detection of buried land mines, remote sensing of human breathing, some other applications. Authors used an in-house-developed coherent laser radar (**Fig. 15**). The sensor is all-fiber-based and operates at a wavelength of 1.5 μm. With an output power of almost 1 W, a spectral line width smaller than 1 kHz and a field of view of 75 μrad, this system is optimized for performing long-range measurements at distances of up to several kilometers.

One of the applications of the developed vibrometer was monitoring the wind turbines. Turbines tend to vibrate strongly, and laser Doppler vibrometry can usefully supplement the existing on-board sensors. The measurements were done on a medium-size wind turbine with a hub height of about 70 m with a rotor diameter of 25 m. The laser Doppler vibrometer was placed about 270 m away from the base point of the wind turbine. The aim was to measure the vibration characteristics of the mast and the nacelle. **Fig. 16** shows the wind turbine with indication of points on the mast where the measurements were done and an example of a spectrogram of the signal recorded from a point in the upper point of the mast. Following the time axis one can see the first natural frequency which is about 0.33 Hz. To study the turbine in multiple points simultaneously, a multi-beam vibrometry system would shorten the measuring time.

It is of great interest to the security and military forces, to use remote vibration sensors to detect and assess human state of health or activity. Typical targets are battlefield or earth-quake casualties, and also humans perceived as potential threats. The LDV can give important hints about human activity and health.

The data from 1.5-μm vibrometer were collected from stationary subjects the laser beam being mostly directed on their bare necks at a distance of about 90 m. **Fig. 17(a)** displays strong pulse beat spikes recorded immediately after a short run with a pulse rate of 110 beats per minute, where the breathing is weakly overlaid by a rate of ∼43 breaths per minute. The lower part (b) displays the signal sequence a little later, where a deep breathing is dominant (breathing rate: 25 breaths per minute) and the body has partly recovered.

An example (Lutzmann *et al.*, 2000) of the visualized distribution of vibrations of a functioning engine is shown in **Fig. 18**. **Fig. 19** illustrates (Polytec 2006) the vibrations of a car's door (left) and of a turbine's blades (right).

Understanding the characteristics of the vibration of musical instruments is important in studying the acoustic characteristic and improving their manufacture technique. **Fig. 20** demonstrates the acoustic vibrations of the violin acquired with the (Polytec 2006) vibrometer using the Doppler principle.

The drumhead has been studied using various methods. Zhang and Su (2005) used the method of Fourier Transform Profilometry (FTP) introduced by Takeda and Motoh (1983). The method is particularly requiring only one image of the deformed fringe pattern to reconstruct the 3D shape of the measured dynamic object. A sinusoidal structured pattern, which is projected onto the surface of a measured drum, is dynamically deformed with the vibration of the membrane. The sequence of the deformed fringe images is grabbed by a high-speed CCD camera, and then a series of wrapped phase maps are produced after the processing with Fourier transform, filtering and inverse Fourier transform. **Fig. 21** shows two different phases of drum vibration.

Compared with the vibration pattern measurement by laser Doppler vibrometry, this method has lower accuracy, but has an obvious advantage that all the data of the whole membrane at one sampling instant can be obtained simultaneously, furthermore its information recording time can last along with the whole vibration process to demonstrate the whole vibration from uncreated to fade away.

To get a spatial distribution of vibrations on the surface of the investigated object, a scanning vibrometry is applied. It allows analysis of the structure with a very fine spatial resolution, not modifying its dynamic behavior, decreasing the testing time if a

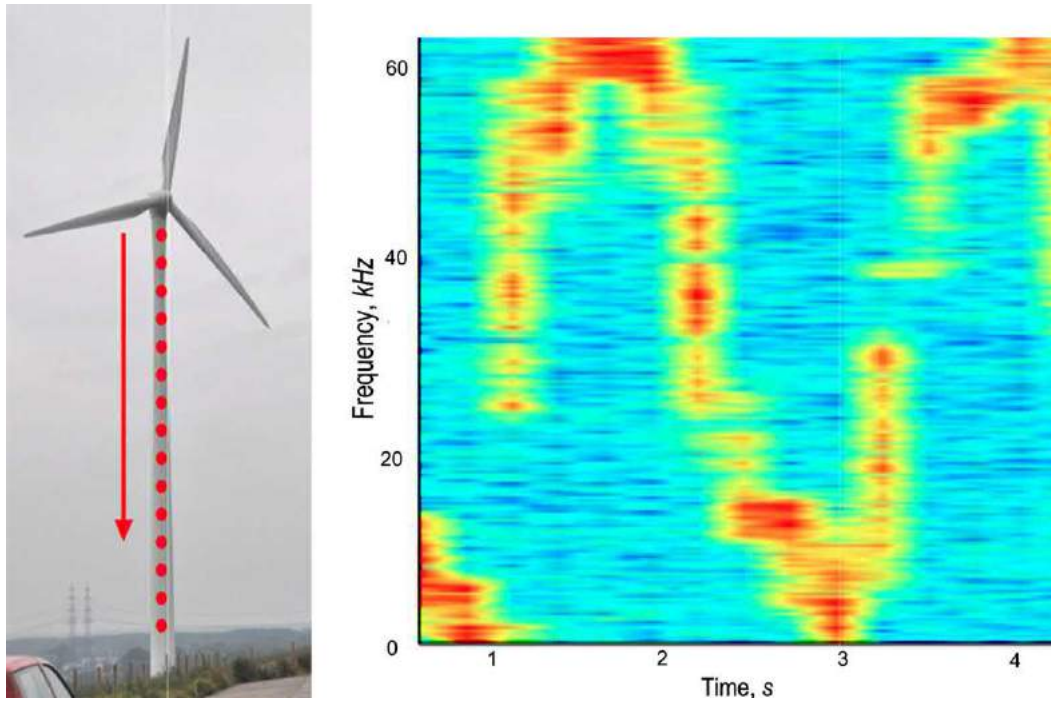


Fig. 16 Measurement of mast modes. Spectrograms of the signal recorded from three selected positions. Reproduced from Lutzmann, P., Göhler, B., Hill, C.A., Putten, F., 2017. Laser vibration sensing at Fraunhofer IOSB: Review and applications. Opt. Eng. 56 (3), 031215.

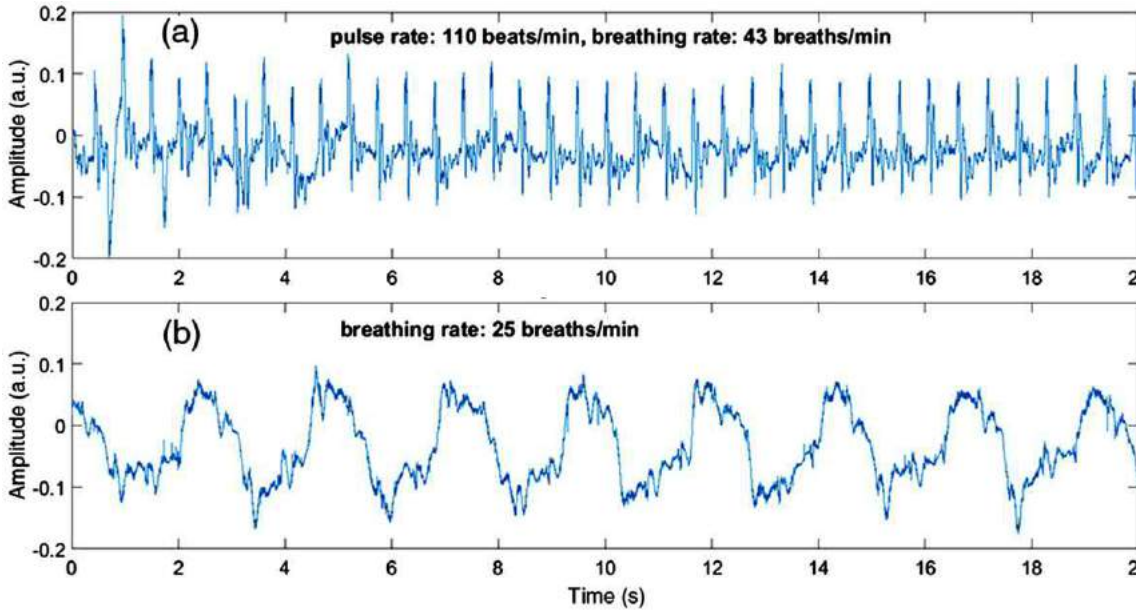


Fig. 17 LDV measurements of pulse beat and breathing. (a) Signal displaying the strong pulse beat spikes immediately after a short run; (b) signal soon after the run, with a deep and slower breathing. Reproduced from Lutzmann, P., Göhler, B., Hill, C.A., Putten, F., 2017. Laser vibration sensing at Fraunhofer IOSB: Review and applications. Opt. Eng. 56 (3), 031215.

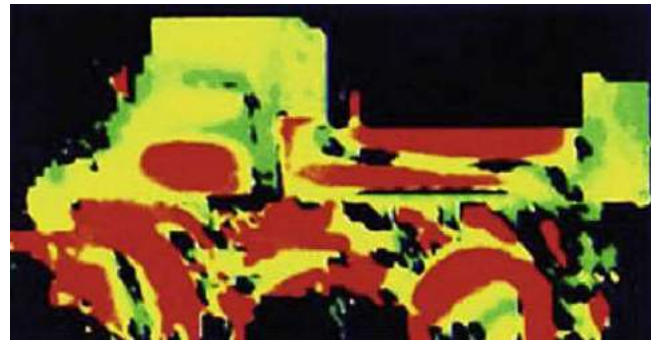


Fig. 18 Vibration intensity map of a functioning engine. Reproduced from Lutzmann, P., Frank, R., Ebert, R., 2000. Laser radar based vibration imaging of remote objects. Proc. SPIE 4035, 436–443.

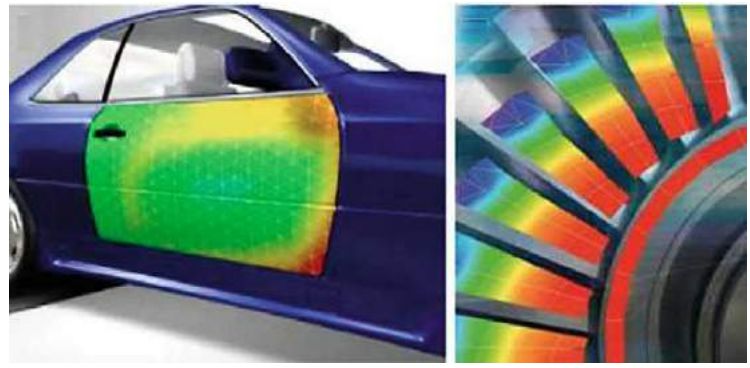


Fig. 19 Vibration intensity maps of a car's door (left) and of a turbine's blades (right). Reproduced from Polytec. Available at: http://www.polytec.com.

large number of measurement points is requested, and allowing measurement of vibrations even on hot surfaces. In some applications, laser vibrometry has also been coupled with laser excitation to develop a complete noncontact and nonintrusive modal analysis procedure.

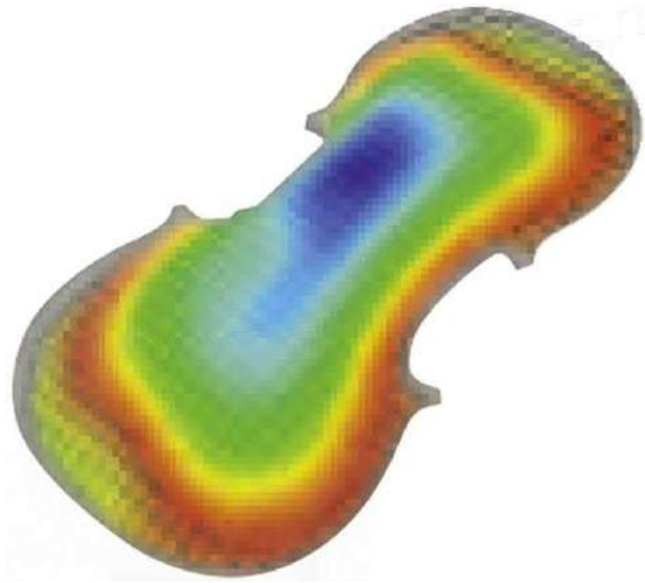


Fig. 20 Acoustic vibrations of a violin acquired with the Polytec vibrometer. Reproduced from Polytec. Available at: http://www.polytec.com.

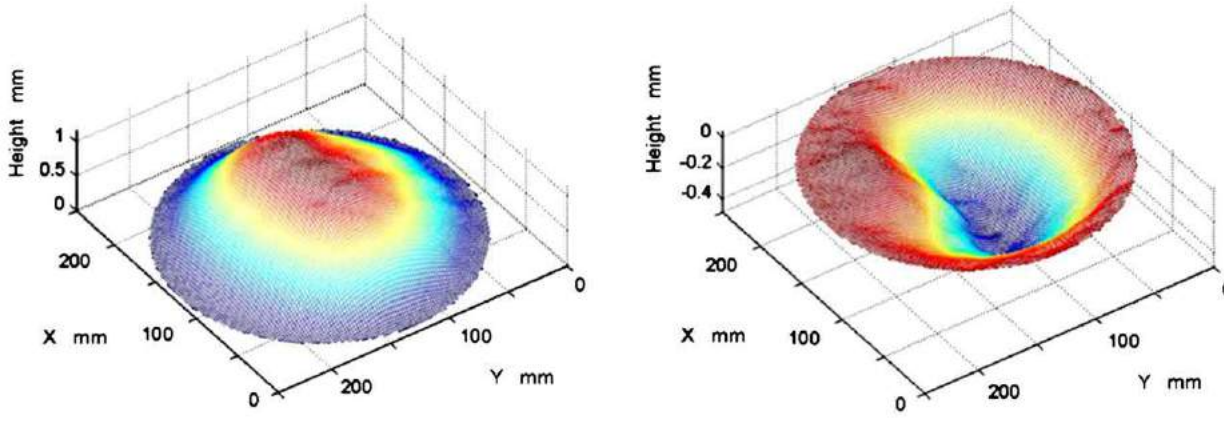


Fig. 21 Different phases of drum vibration. Reproduced from Zhang, Q., Su, X., 2005. High-speed optical measurement for the drumhead vibration. Opt. Express 13, 3110–3116.

Revel *et al.* (2011) investigated the problem of noise in the cabin of Agusta A109MKII. They described the use of a scanning laser Doppler vibrometer to measure vibrations inside the helicopter using its mock-up, and demonstrated the applicability of the technique for in-flight tests. The whole area was 430×315 mm. In the scanning tests, a 30×20 grid (corresponding to a spatial resolution of about 14×15 mm) was used. An additive mass was used in order to lower the center of mass of the vibrometry system and to gain stability. The comparison was also made with the vibrograms measured when the vibrometer was placed outside the mock-up. **Fig. 22** is a summary of the tests at different resonances. They all refer to instantaneous amplitude of the velocity component at a certain frequency orthogonal to the surface with the bandwidth ± 10 Hz for the frequencies up to 1000 Hz and ± 30 Hz for the frequencies up to 5000 Hz.

Perea and Libbey (2016) demonstrated a laboratory system for measurement of three degrees of vibrational freedom simultaneously (Perea and Libbey, 2016). Heterodyne speckle imaging was exploited in which the optical speckle pattern was mixed with a coherent reference. From these data, three dimensions of surface motion are extracted. Axial velocity is measured by demodulating the received time-varying intensity of high amplitude pixels. Tilt, a gradient of surface displacement, is calculated by measuring speckle translation following extraction of the speckle pattern from the mixed signal.

Media Parameters as Dimensions

The diversity of laser radar applications for media investigation is impressive. Many of them use detection and processing of signals scattered in the medium of propagation, that can be not only the atmosphere of planets, but also the water and any other

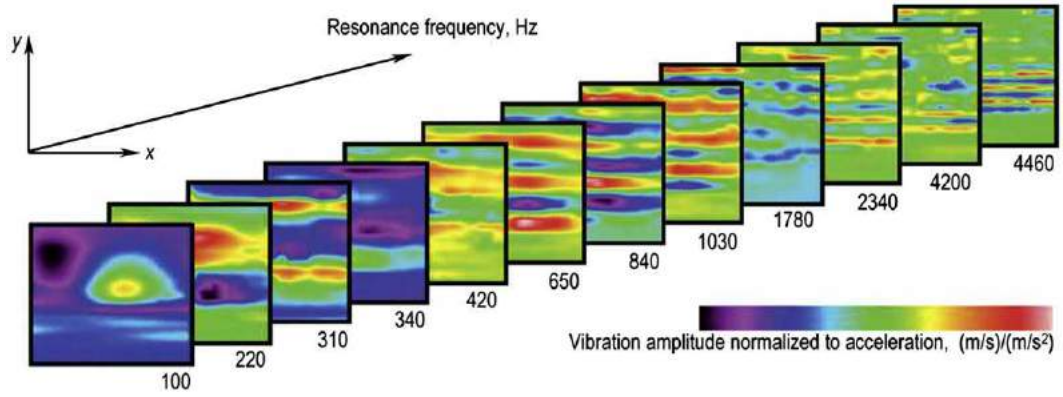


Fig. 22 Resonance vibration frequencies registered in the cabin of the helicopter mock-up. Summary of images reproduced from Revel, G.M., Castellini, P., Chiariotti, P., *et al.*, 2011. Laser vibrometry vibration measurements on vehicle cabins in running conditions: Helicopter mock-up application. Opt. Eng. 50, 101502.

medium of organic or non-organic origin. Examples could be like airborne atmospheric remote sensing (Palm *et al.*, 1994), cloud top remote sensing by airborne lidar (Spinhirne *et al.*, 1982), airborne high spectral resolution lidar that provides measurements of aerosol backscatter and extinction coefficients and aerosol depolarization at two wavelengths (Hair *et al.*, 2008), nanostructure and defect analysis using a simple 3D light-scatter sensor (Herffurth *et al.*, 2013), refraction distribution in the human eye (Molebny, 2013). Even a military rangefinder can be used for atmosphere sensing (Steinvall *et al.*, 2014). There are also many other applications.

Retro-scattering based on interaction of laser light with medium components brings another amount of information. Examples are as follows: Raman lidar for water vapor, temperature, aerosol, and cloud measurement (Reichardt *et al.*, 2012), dual field-of-view Raman lidar measurements for the retrieval of cloud microphysical properties (Schmidt *et al.*, 2013), lidar measurements of the Eyjafjallajökull ash cloud over the Netherlands (Donovan and Apituley, 2013), spectrally resolved Raman lidar measurements of gaseous and liquid water in the atmosphere (Liu and Yi, 2013), using Raman and elastic backscatter for measurement of temperature, density, pressure, moisture, and particle backscatter coefficient (Fraczek *et al.*, 2012).

Philbrick *et al.* (2010) experimentally showed how multi-wavelength backscatter measurements from Rayleigh and Raman lidar techniques can provide signals to be used to profile the properties of the atmospheric column. Rayleigh lidar signals provide backscatter coefficients, and Raman lidar signals backscattered from the major molecular components provide extinction profiles. The ratio of these simultaneous extinction and backscatter measurements are used to classify the aerosol type. In addition, a laser beam can be used to make bistatic and multistatic measurements of the polarization ratio of the scattering phase function. Analysis of multistatic measurements can be used to determine profiles of the aerosol number density, size, size distribution, and type. These parameters can be determined for spherical particles in the size range between about 20 nm and 20 μm. Information on aerosol type and shape can be supposed from determining the approximate refractive index of the scattering aerosols and by measuring the depolarization of the scattered radiation.

In a later publication, Philbrick and Hallen, (2017) discussed the Raman lidar measurements that provide profiles of several different tracers of spatial and temporal variations, which are excellent signatures for studies of dynamical processes in the atmosphere. An examination of Raman lidar data collected during the last four decades clearly showed signatures of atmospheric planetary waves, gravity waves, low-level jets, weather fronts, turbulence from wind shear at surfaces and at the interface of the boundary layer with the free troposphere. Water vapor profiles are important as a tracer of the sources of turbulence eddies associated with thermal convection, pressure waves, and wind shears, which result from surface heating, winds, weather systems, orographic forcing, and regions of reduced atmospheric stability. Examples of these processes were selected to show the influence of turbulence on profiles of atmospheric properties.

Fluorescence is one of physical phenomena successfully used in the today's laser radars for remote detection of oil spills (Sato *et al.*, 1978), for remote sensing of vegetation (Matvienko *et al.*, 2006), for fluorescence imaging of the ocean bottom (Sitter and Gelbart, 2001). High-repetition-rate three-dimensional OH imaging using scanned planar laser-induced fluorescence system allows for multiphase combustion control (Cho *et al.*, 2014). Combinations of several physical phenomena (elastic scattering, fluorescent scattering, and differential absorption) were the basics of the developed airborne lidar for observations of atmospheric tracers (Uthe, 1991). Some examples of acquiring the media parameters and data 3D presentation are given hereinafter.

Wind and Flow Velocity

The need for remote sensing of atmospheric winds is well established, be it from ground-based sites, air-, or space-borne equipment. Ladar systems observe large volumes of atmosphere with high spatial and temporal resolution, making these data important for many applications. Doppler ladars can use incoherent or coherent techniques. In the incoherent system, to isolate

the laser line from the day light influence, a Fabry-Pérot étalons (interference narrow-band filters) are used. The laser is typically Nd-doubled 532 nm, 3 W average power (Fischer *et al.*, 1995). The field of view of the receiving telescope is of the order of 0.5 mrad that is somewhat larger than the laser divergence of 0.2 mrad to collect all of the laser light, which may be outside the divergence angle because of different factors like pointing jitter, scanning instability, or similar. The collected light is focused by the telescope onto an optical fiber with the diameter of 3.5 mm. The spatial scrambling of the collected light is necessary to remove the effects of inhomogeneity in the scattering aerosol volumes and of changes in normalization degree in detectors, and to avoid systematic offset errors in measured wind velocity. **Fig. 23** shows a time series of the horizontal wind field at the altitude of 100 m. The velocity vectors are scaled as shown in the diagram. Complete profiles were generated every 5 min.

Low altitude wind profile measurements with a laser rangefinder were demonstrated using a simple balloon tracking system, with small (0.25 m diameter) lightweight balloons (Wilkerson *et al.*, 2009). Experiments on balloon trajectories demonstrate that laser range detection (± 0.5 m) combined with azimuth and elevation measurements is a simple, accurate, and inexpensive alternative to other wind profiling methods. To increase the maximum detection range to 2200 m, a retroreflector tape was attached to the balloons. Nighttime tracking was facilitated by low power LEDs.

Wind speed and direction results are compared with simultaneous sodar measurements (**Fig. 24**). The profiling resolution is greatly improved using a laser rangefinder with automatic coordinate and time recording. However, balloon tracking is still man-in-the-loop. Improvements should include automation of the tracking system to collect trajectory points automatically at 1 Hz or faster.

Windshear poses the greatest danger to aircraft during takeoff and landing due to its abrupt change of direction (**Fig. 25**) (Targ *et al.*, 1991), when the plane is close to the ground and has little time or room to maneuver (Molebny and Steinvall, 2013) (**Fig. 26**). During landing, the pilot has already reduced engine power and may not have time to increase speed enough to escape the downdraft. During takeoff, an aircraft is near stall speed and thus is very vulnerable to windshear. Microburst windshear often occurs during thunderstorms. But it can also arise in the absence of rain near the ground. Pilots need 10–40 s of warning to avoid windshear.

The first airborne measurements of winds used a pulsed CO_2 laser (Bilbro and Vaughan, 1978; Bilbro, 1984). The lidar was oriented in the fore and aft directions in order to obtain the horizontal vector wind field. Another coherent lidar system Coherent Lidar Airborne Shear Sensor (CLASS) was developed in two versions: with a 10.6 μm CO_2 laser (CLASS-10) and with a solid state 2.02 μm Tm: YAG laser (CLASS-2). Both lidars showed a wind measurement accuracy better than 1 m/s (Targ *et al.*, 1996). Coherent airborne wind lidar system Wind INfrared Doppler lidar (WIND) was developed later in French-German cooperation (Werner *et al.*, 2001).

NASA's Langley Research Center tested three airborne predictive windshear sensor systems (NASA, 2017): microwave radar detecting the raindrop scattering, Doppler laser radar detecting the aerosol scattering, and infrared detector measuring the temperature changes ahead of the airplane. The system was manufactured by Lockheed Corp.'s Missiles and Space Co.; United Technologies Optical Systems Inc., and Lassen Research.

In the wind infrared Doppler lidar (WIND) project, airborne conically scanning CO_2 Doppler lidar was used. Spatial resolution requirements were 250 m in height with a grid size of 10×10 km. The radiation of the local oscillator is mixed with outgoing pulse and with the Doppler-shifted backscattered signal. A locking loop connects both lasers. The laser pulse with a pulse duration

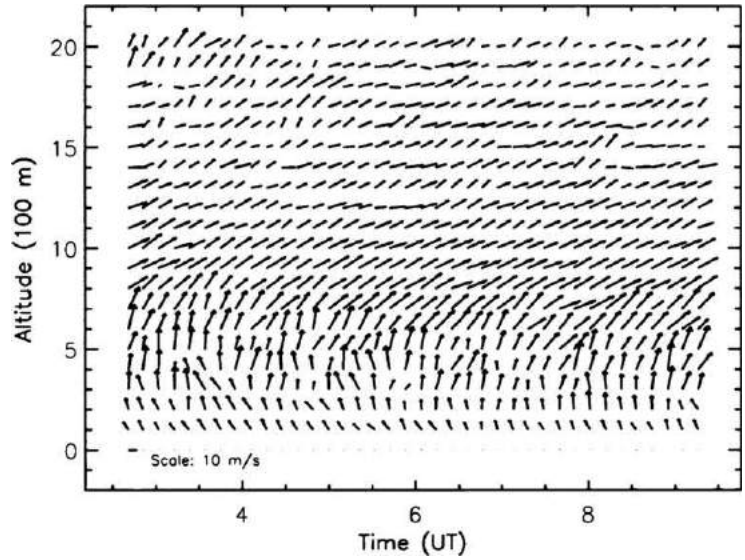


Fig. 23 Examples of displaying the information on wind measurement in the form of vectors vs time, right – vertical distribution in comparison with sonar data. Reproduced from Fischer, K.W., Abreu, V.J., Skinner, W.R., *et al.*, 1995. Visible wavelength Doppler lidar for measurement of wind and aerosol profiles during day and night. Opt. Eng. 34 (2), 499–511.

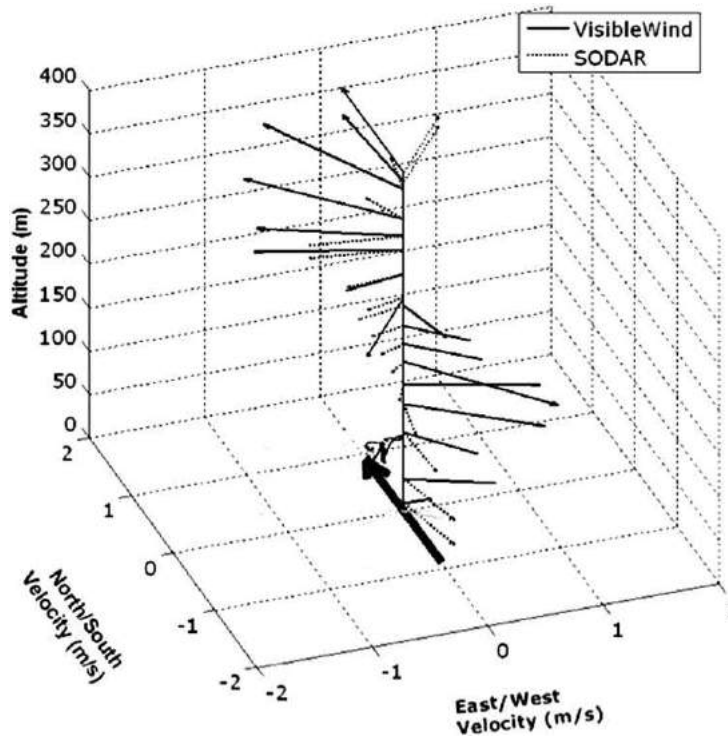


Fig. 24 Wind changing vertically in comparison with sonar data. Reproduced from Wilkerson, T., Bradford, B., Marchant, A., *et al.*, 2009. VisibleWindTM: A rapid-response system for high-resolution wind profiling. Proc. SPIE 7460, 746009.

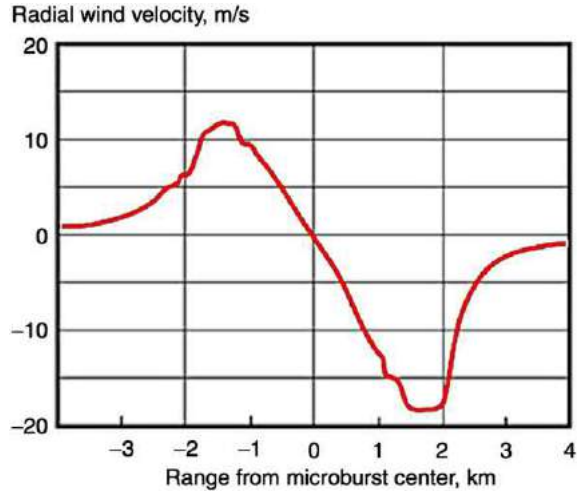


Fig. 25 Velocity distribution within the windshear. Reproduced from Targ, R., Kavaya, M.J., Huffaker, R.M., Bowles, R.L., 1991. Coherent lidar airborne windshear sensor: Performance evaluation. Appl. Opt. 30, 2013–2026.

of 2.5 μs and 100–300 mW of per-pulse power is sent out via the transceiver telescope into the region of investigation with a repetition frequency 10 Hz. The wind-shifted Doppler frequency directly determines the line-of-sight component of the wind vector. At CO_2 laser wavelength 10.6 μm, a velocity component of 1 m/s corresponds to a frequency shift of 189 kHz. Ground-based tests, airborne tests and a validation flight were performed. An example of polar wind map is shown in **Fig. 27**.

Lockheed Martin Coherent Technologies developed a system called WindTracer (Lockheed Martin, 2015), whose specialty is remote sensing of winds in the critical 40–200 m height regime. With no sidelobes, WindTracer can scan the space near the ground, even adjacent to obstacles. It provides a 30–60 min warning of changes in the winds approaching a wind farm. This information is important for grid electric power operators and for turbine operators to optimize the operations. The wavelength is the eye-safe 1.6 mm, the maximum operating range is up to 12 km, range resolution is about 50 m, the coverage area

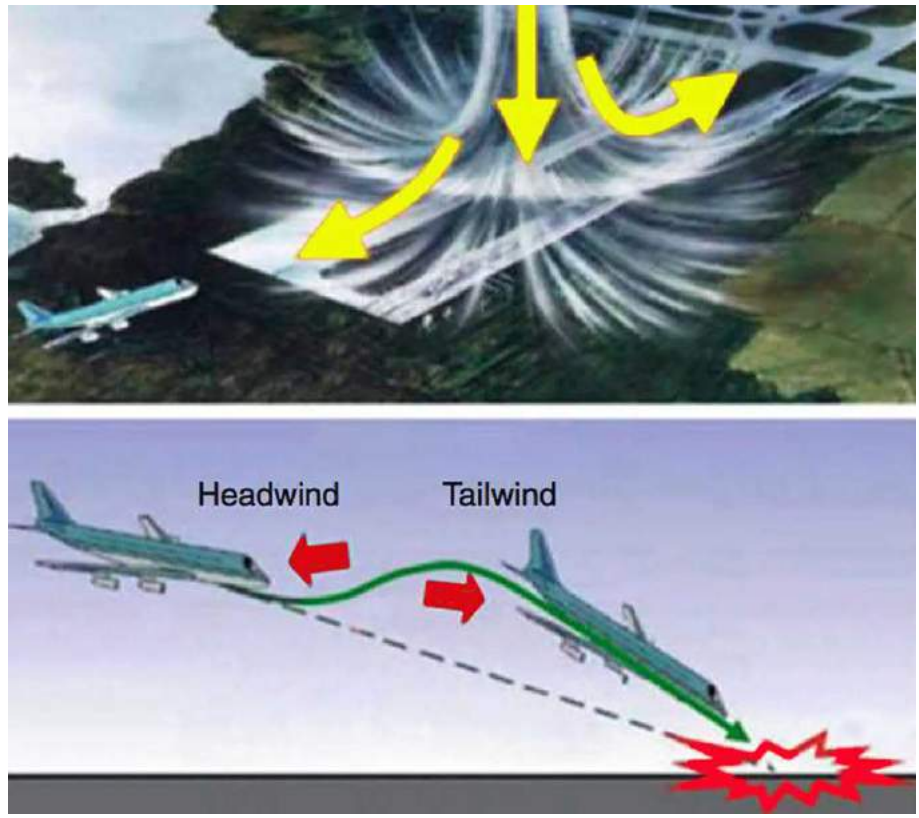


Fig. 26 Windshear and its effect. Reproduced from Molebny, V., Steinvall, O., 2013. Laser remote sensing. Velocimetry based techniques. In: Tuchin, V.V. (Ed.), Handbook of Coherent-Domain Optical Methods. New York: Springer, pp. 363–395.

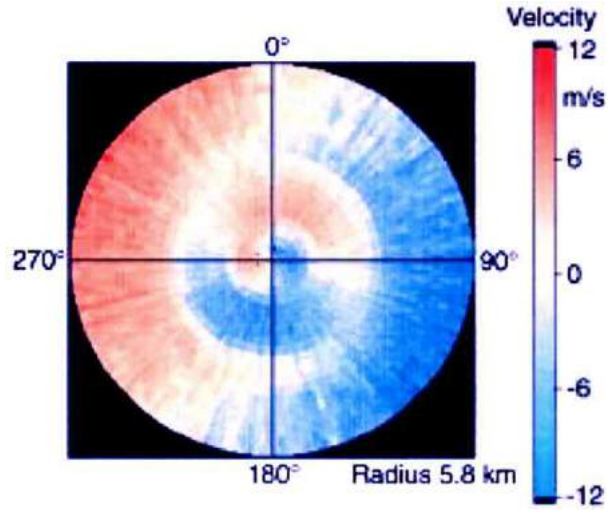


Fig. 27 Polar plot intensities of the wind fields. Reproduced from Werner, C., Flamant, P.H., Reitebuch, O., *et al.*, 2001. Wind infrared Doppler lidar instrument. Opt. Eng. 40 (1), 115–125.

is 100–200 km^2, and a typical vector velocity accuracy is about 1 m/s. The combination of lidar and radar, called WindTracer Terminal Doppler Solution (WTDS) enables wind hazard detection under wider range of weather conditions.

A general view of WindTracer is demonstrated in **Fig. 28**. The screenshot from its display (**Fig. 29**) shows the radial velocity estimates over a 12-km-diameter circular area.

European companies Leosphere (2004) and Halo Photonics specialize in building compact Doppler lidars with emphasis on wind farming applications. Leosphere developed several models of Windcube (100S/200S/400S) of the Doppler lidar. Radial

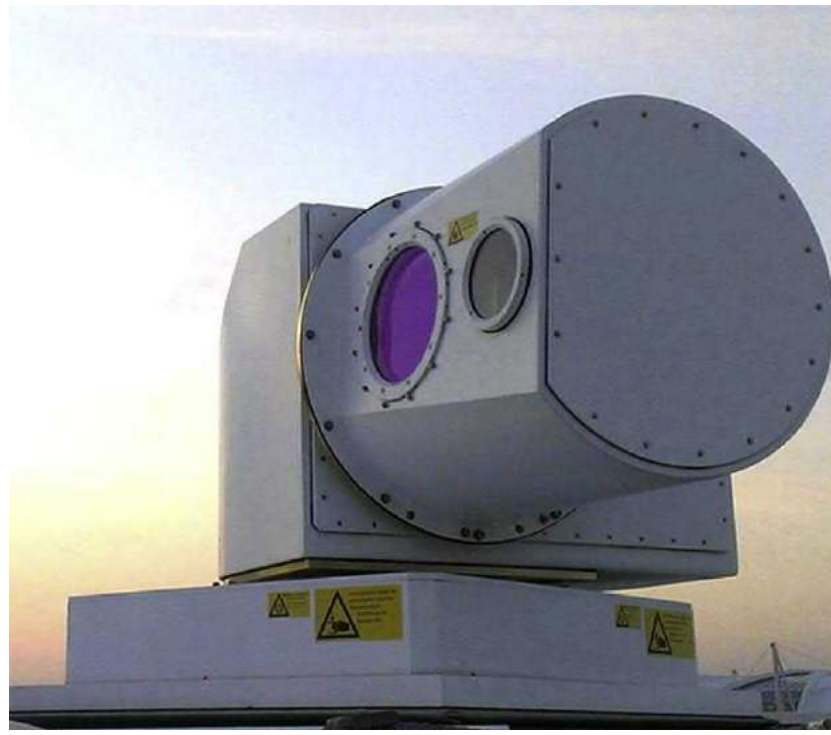


Fig. 28 General view of WindTracer. Reproduced from Lockheed Martin. Available at: http://www.lockheedmartin.com/products/WindTracer/index.html.

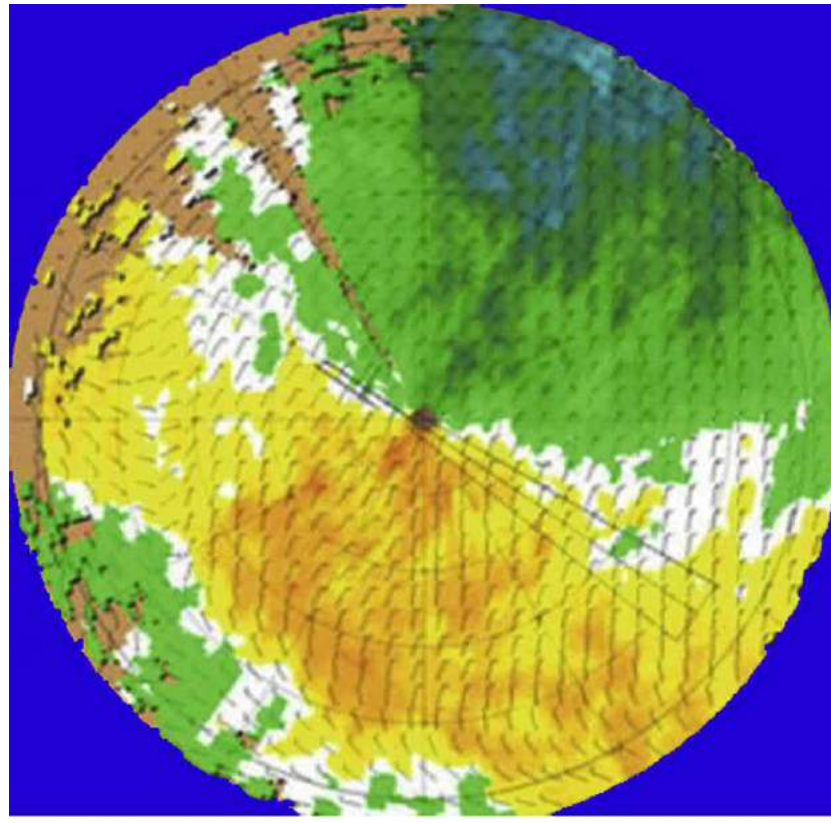


Fig. 29 WindTracer's radial velocity estimates. Reproduced from Lockheed Martin. Available at: http://www.lockheedmartin.com/products/WindTracer/index.html.

velocity display from 200S model overlaid on the land map is shown in Fig. 30. An example of velocity-height dependency on time acquired with Halo Photonics Doppler velocimeter is demonstrated in Fig. 31. As the wind energy industry continues to grow, the need for more accurate and sophisticated wind data grows as well. Such lidars are used to explore the locations like ridgelines, oceans, lakes, and forested areas. Typically, the lidars provide 100–200 m vertical wind profiles with an accuracy down to 0.1 m/s. Maximum range of the Halo Photonics streamline instrument using 1500 nm laser wavelength is 9.0 km, range resolution 15 m, maximum measured velocity 19 m/s, velocity resolution 0.0384 m/s.

The New European Wind Atlas (NEWA) project has a goal to create a freely accessible wind atlas covering Europe and Turkey. Common to all experiments is the use of Doppler lidar systems. Many European and American companies and universities will participate in the project. Two pilot experiments, one in Portugal and one in Germany, showed the value of using multiple synchronized, scanning lidars (Mann *et al.*, 2017). Example of the scans at three different elevation angles are shown in Fig. 32 received from the WindScanner developed at Technical University of Denmark for the purposes of NEWA, within the

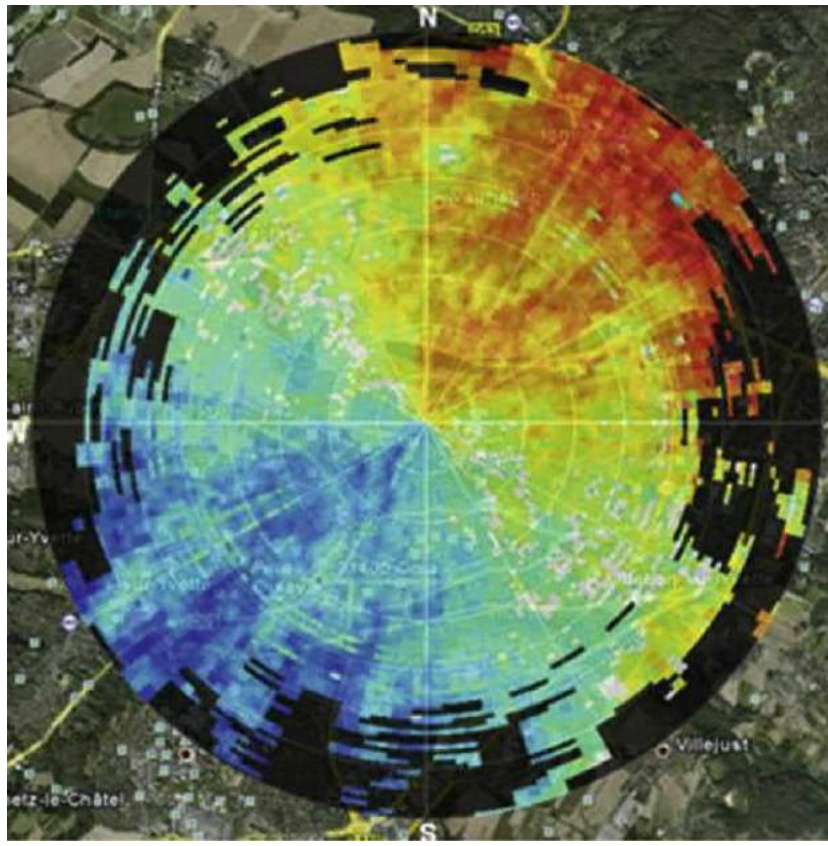


Fig. 30 Radial velocity display from Windcube 200S. Reproduced from Leosphere. Available at: http://www.leosphere.com/8,wind-energy.

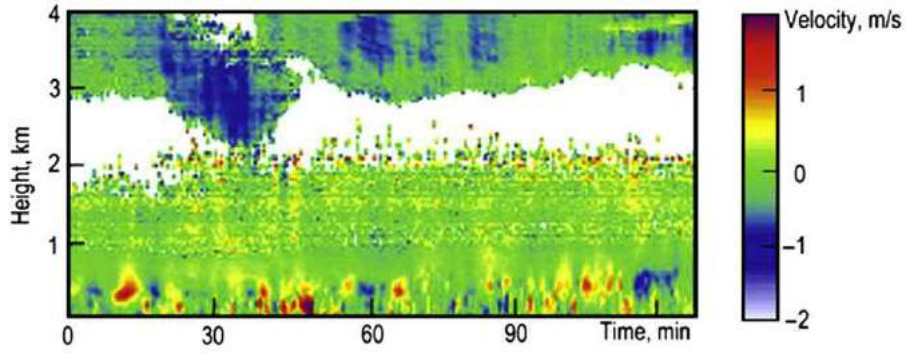


Fig. 31 Velocity (color-coded)-height vs. time acquired with Halo Photonics Doppler velocimeter. Reproduced from Halo Photonics. Available at: http://halo-photonics.com/index.htm.

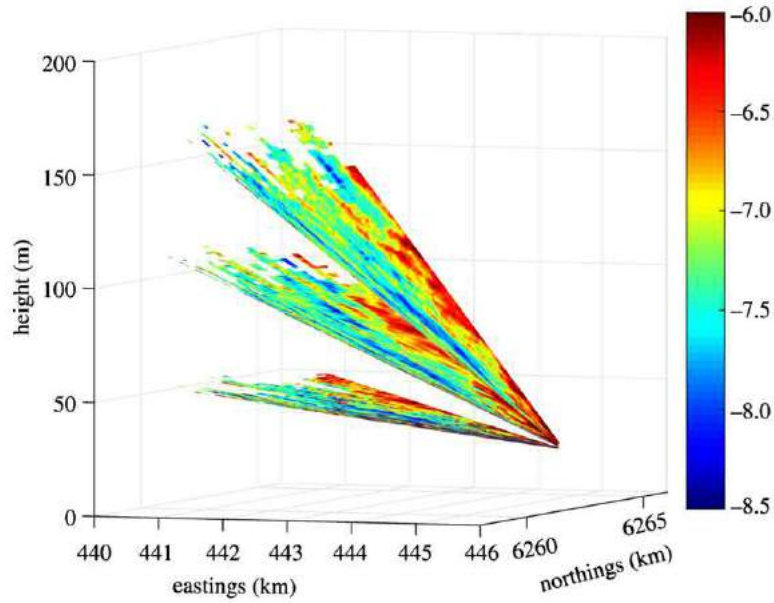


Fig. 32 Wind scans at three different elevation angles. Reproduced from Mann, J., Angelou, N., Arnqvist, J., *et al.*, 2017. Complex terrain experiments in the New European Wind Atlas. Phil. Trans. R. Soc. A 375, 20160101.

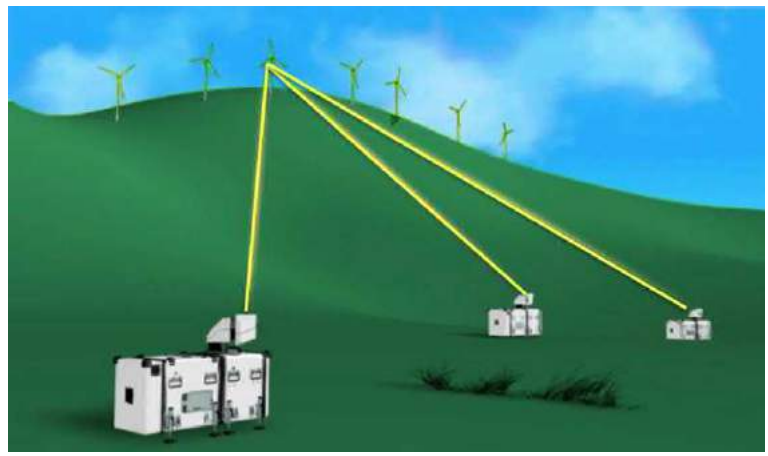


Fig. 33 Three simultaneously operating wind lidars equipped with synchronized steerable-beam scanners. Reproduced from Mikkelsen, T., 2014. Lidar-based research and innovation at DTU wind energy – A review. J. Phys.: Conf. Ser. 524, 012 007.

WindScanner.dk project in long- and short-range versions (Mikkelsen, 2014) with the following main parameters (Vasiljevic *et al.*, 2017). Long-range version: pulse mode, range 50–8000 m, 500 range gates, probe length is 25, 35 or 70 m (fixed with range). Short-range version: CW, range 10–50 m, a single range gate, probe length is 0.2–40 m (evolving with range). These instruments confirmed the potentials not only to economically measure the wind field at a single point located within the rotor swept area, but also to map entire wind fields within the volume occupied by modern wind turbine and wind farms.

By combining space and time synchronized measurements along the line of sight from three simultaneously operated wind lidars equipped with synchronized steerable-beam scanners, true 3D wind velocity and turbulence measurements have become attainable for wind energy industry applications (**Fig. 33**).

WindScanner system consists practically of two or more spatially separated scanning lidars (long- or short-range WindScanners), that are coordinated or controlled by a master computer. The first generation of the long-range WindScanner (LRWS) and short-range WindScanner (SRWS) originate from commercially available vertical profiling lidars Windcube 200 (Werner *et al.*, 2001) and ZephIR, respectively. These profiling lidars were converted into the scanning lidars. The WindScanners have been specifically tailored to perform user-defined and time-controlled scanning trajectories, either independently or in a synchronized mode. A single LRWS can acquire radial velocities simultaneously at a maximum 500 different distances along each line-of-sight at a maximum rate of 10 Hz, whereas a single SRWS can measure radial velocity from one single range at a time but with a maximum rate of 400 Hz.

The lidars in the long-range WindScanner system are usually connected to the master computer using a 3G network. The short-range WindScanner system is formed by connecting the master computer with the lidars via a 300-meter long optic fiber.

The long-range WindScanner system is intended for measurements of mean flow field within a large volume of the atmosphere, while the short-range WindScanner system is typically applied to perform small-scale measurements of turbulent flows around a single wind turbine rotor.

The NEWA project aims to improve wind resource modeling for different site conditions. Areas with steep ridges and forested terrain are of particular interests, since the current engineering (linear) flow models are unable to correctly predict behavior of the flow over the sites with these features. The accompanying projects were focused on optimization of wind generators. To fulfill the project's ambition, it was necessary to investigate how the incoming flow is modified by a wind turbine. Still another objective was to develop numerical tools for wind farm layout optimization in complex terrain.

The site for the tests was chosen in Perdigão, Portugal with the opportunity to measure wind resources along a ridge, occurrences of flow separations on lee sides of hills (i.e., recirculation zone) and valley flows. According to the scientific objectives, the site contains a steep ridge where an isolated wind turbine is operating (**Fig. 34**, top photo).

A quasi two-dimensional, or a long ridge hill was a logical choice. Also, to assure a two-dimensional flow, dominant winds should be perpendicular to the ridge. Land cover, particularly forests would add to the flow complexity, and is considered desirable, since many wind farms are installed near or within forested regions. The presence of a wind turbine at the site gave the opportunity for wake and inflow measurements.

The two WindScanner systems, comprised a total of six scanning lidars: three LRWSs and three SRWSs. The lidar locations were selected according to the aim to sample the flow field along the South ridge and within the transect perpendicular to the ridges entailing the wind turbine.

The simulation of wind flow was done using steady state formulation for the equations, and the model's wall boundary parameters were calibrated to yield wind speeds of 5–6 m/s and other parameters were taken into account with suggested values. The simulations of the flow from the Northeast and Southwest predict large recirculation zone enclosed in the valley (**Fig. 34**) and a high complexity of the flow (**Fig. 35**).

Experimental studies should follow to confirm the simulation with the correctives shown up in the studies. It will be the six-dimensional (6D) maps of wind space, three dimensions of which will be the (x, y, z) coordinates of the space and other three dimensions (V_x, V_y, V_z) will be the velocity components of the wind vectors within that space. If these 6D maps would be supplemented with the temperature in each (x, y, z) point, the information would become seven-dimensional (7D: x, y, z, V_x, V_y, V_z, T). The humidity can be the following candidate for the eight-dimensional (8D: x, y, z, V_x, V_y, V_z, T, h) space. Reconstructing this 8D space of information as a time sequence, one should get the nine-dimensional (9D: x, y, z, V_x, V_y, V_z, T, h, t) space of information. The list of dimension candidates can be changed or extended in correspondence with the tasks of the study.

Non-Doppler, ground-based, scanning aerosol backscatter lidars can measure the wind speed and direction by cross correlation of coherent aerosol structures (Eloranta *et al.*, 1975). The primary assumptions are that the inhomogeneities remain coherent, and that these spatially-coherent turbulent features persist for a time that is long compared to the acquisition time of the lidar system during the advective process.

Non-Doppler incoherent lidars do not require coherent detection, which removes limitations of telescope diameter and relaxes requirements for high-spectral-purity lasers. This lack of restriction becomes more attractive as large lightweight mirrors and small powerful lasers become more affordable and commercially available. In addition to the wind field, these measurements can also be used to compute mean dimensions and lifetimes of turbulence. Unfortunately, the cross-correlation technique does not measure the wind velocity in all conditions. Coherent structures may not be advected with the wind in the scan plane or scan volume. Three-dimensional scan strategies are able to eliminate this possibility but are still susceptible to false velocities from wavelike motions. It is fortunate that irregularities in the aerosol scattering field within turbulent regions are usually oriented in all directions, and, by averaging in space and time, the motion of the coherent structures advecting with the wind dominates the correlation functions.

Based on the aerosol backscatter correlation, a three-beam lidar was developed (Prasad and Mylapore, 2017) to measure wind characteristics for wake vortex and plume tracking applications. This is a direct detection elastic lidar that uses three laser transceivers (**Fig. 36**), operating at 1030-nm wavelength with ~10-kHz pulse repetition frequency and nanosecond class pulse widths, to directly obtain three components of wind velocities. By tracking the motion of aerosol structures along and between three near-parallel laser beams, three-component wind speed profiles along the field-of-view of laser beams can be obtained. With three 8-in. transceiver modules, placed in a near-parallel configuration on a two-axis pan–tilt scanner (**Fig. 37**), the lidar measures wind speeds up to 2 km away. Aerosol density fluctuations are cross-correlated between successive scans to obtain the displacements of the aerosol features along the three axes. Using the range resolved elastic backscatter data from each laser beam, which is scanned over the volume of interest, a three-dimensional map of aerosol density can be generated.

Atmosphere Components and Contaminants

3D structure of the water vapor field was observed using a scanning water vapor differential absorption lidar (DIAL) (Behrendt *et al.*, 2009). The instrument is mobile and was applied successfully in two field campaigns. The data products of the DIAL are profiles of absolute humidity with typical range resolutions of 15–300 m and temporal resolution of 1–10 s. Maximum range is several kilometers at both day and night. Spatial and temporal resolution can be traded off against each other.

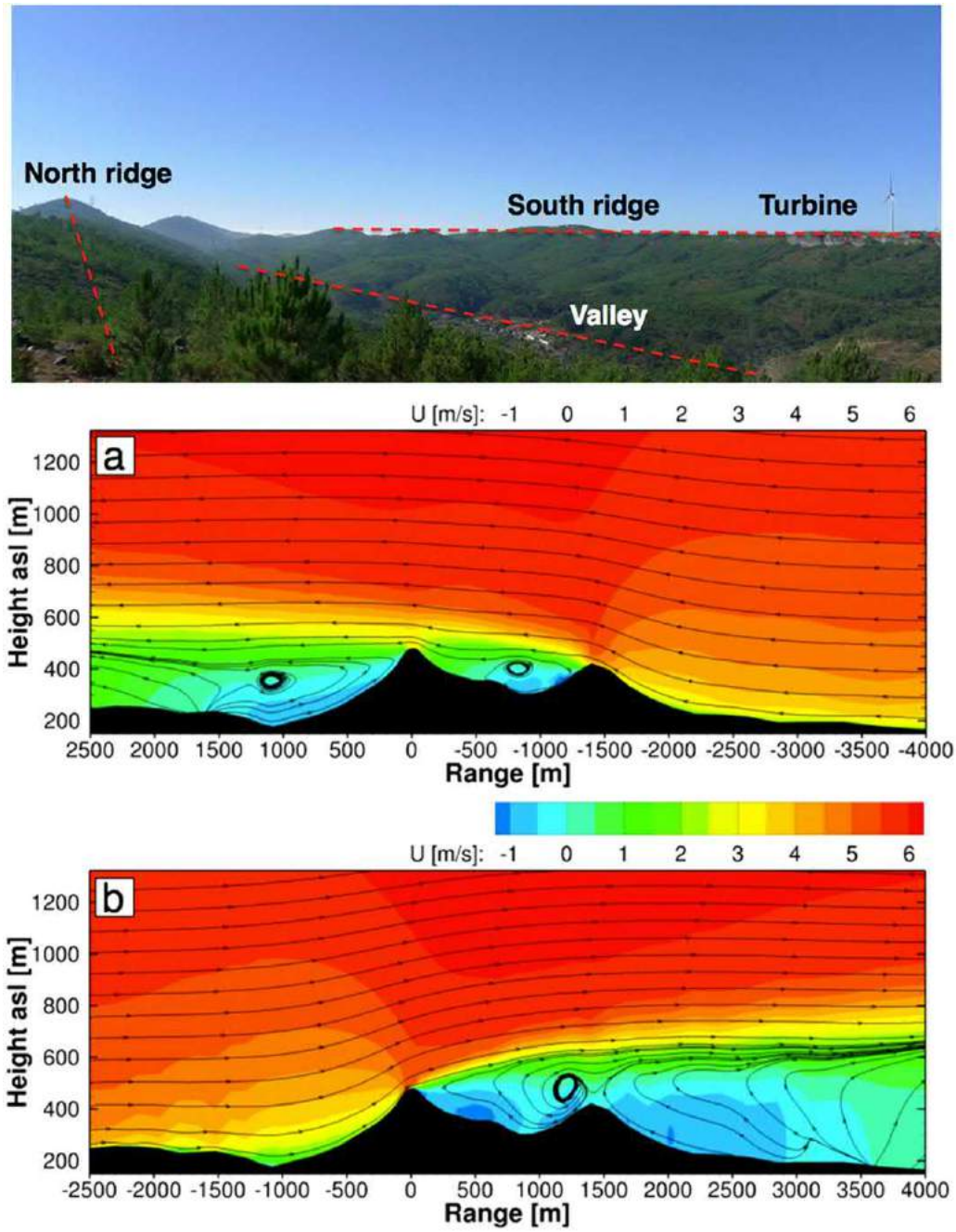


Fig. 34 Profile of the test site (top). Wind flow in a vertical plane: (a) Northeast winds; (b) Southwest winds. Coordinate system origin corresponds to wind turbine base. Reproduced from Vasiljevic, N., Palma, J., Angelou, N., *et al.*, 2017. Perdigão 2015: methodology for atmospheric multi-Doppler lidar experiments. Atmos. Meas. Tech. Discuss. doi:10.5194/amt-2017-18.

Beside humidity, the backscatter field and thus aerosols and clouds can be observed simultaneously. The DIAL transmitter is based on an injection-seeded Ti:Sapphire laser operated at 820 nm which is pumped with a Nd:YAG laser.

A measurement example in vertical pointing mode during the first field deployment of the instrument is shown in **Fig. 38**. Even with a resolution of only 15 m and 1 s, the noise of the data is low; the high variability of the water vapor field in this case is revealed. Comparisons of the data acquired with the instrument and with the radiosondes launched at the lidar site showed a close agreement.

Range-height-intensity scanning measurement examples are presented in **Fig. 39**. The complexity of the humidity field is striking. Several horizontal layers can be seen in addition to turbulent structures with high moisture value from close to the ground up to a height of about 300 m.

To measure gaseous and liquid water in the atmosphere, a spectrally resolved Raman lidar based on a tripled Nd:YAG laser (100 mJ per pulse at 354.7 nm with a repetition rate of 20 Hz and a pulse width of 6 ns) was developed at the Wuhan University

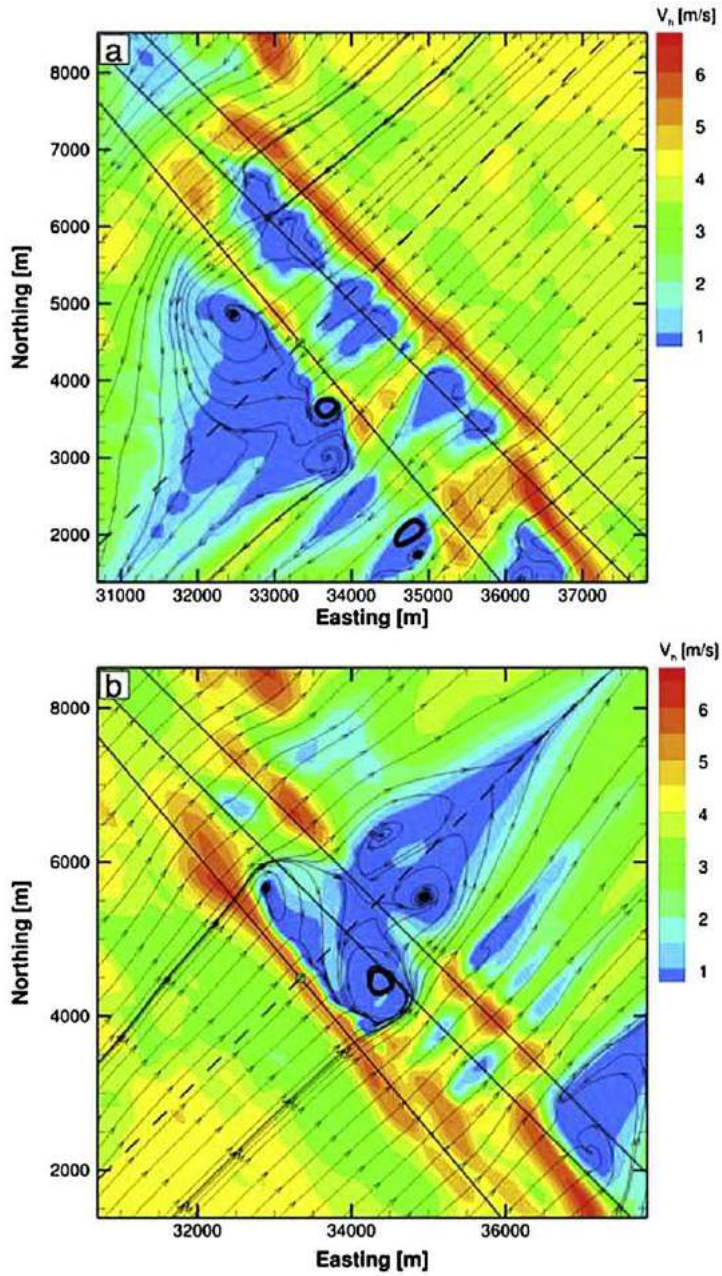


Fig. 35 Wind flow in surface 80 m agl: (a) Northeast winds; (b) Southwest winds. Reproduced from Vasiljevic, N., Palma, J., Angelou, N., *et al.*, 2017. Perdigão 2015: Methodology for atmospheric multi-Doppler lidar experiments. Atmos. Meas. Tech. Discuss. doi:10.5194/amt-2017-18.

(Liu and Yi, 2013). At the receiver side, the backscattered photons are collected by a 450 mm Cassegrain telescope. After the bandpass filtering, the light is directed into a double-grating polychromator to separate the wavelengths and to suppress elastic backscatter. A 32-channel linear-array photomultiplier tube is employed to sample atmospheric Raman water spectrum between 401.65 and 408.99 nm. The 354.7 nm elastic Mie/Rayleigh scattering light in each channel is suppressed by more than 13 orders of magnitude (6 orders by the bandpass filter and 7 orders by the polychromator).

The output photons from each channel are counted with a bin width of 200 ns (corresponding to a range bin length of 30 m). The photon counts from 11,000 laser shots are accumulated to produce one photon count profile at each channel. This yields a time resolution of about 10 min for the raw data. The 32-channel photon count data constitute a spectrum- and altitude-resolved atmospheric Raman water signal. Altitude-resolved atmospheric Raman water spectrum signal is demonstrated in **Fig. 40**. Note that the prominent Raman water vapor signal peak at about 407.57 nm is visible at the altitudes up to 8.0 km.

The lidar-observed Raman water spectrum in the very clear atmosphere is nearly invariable in shape. It is dominated by water vapor, and can serve as a background reference for Raman lidar identification of the phase state of atmospheric water under

Fig. 36 Direct detection elastic lidar with three laser transceivers. Reproduced from Prasad, N.S., Mylapore, A.R., 2017. Three-beam aerosol backscatter correlation lidar for wind profiling. Opt. Eng. 56 (3), 031222.

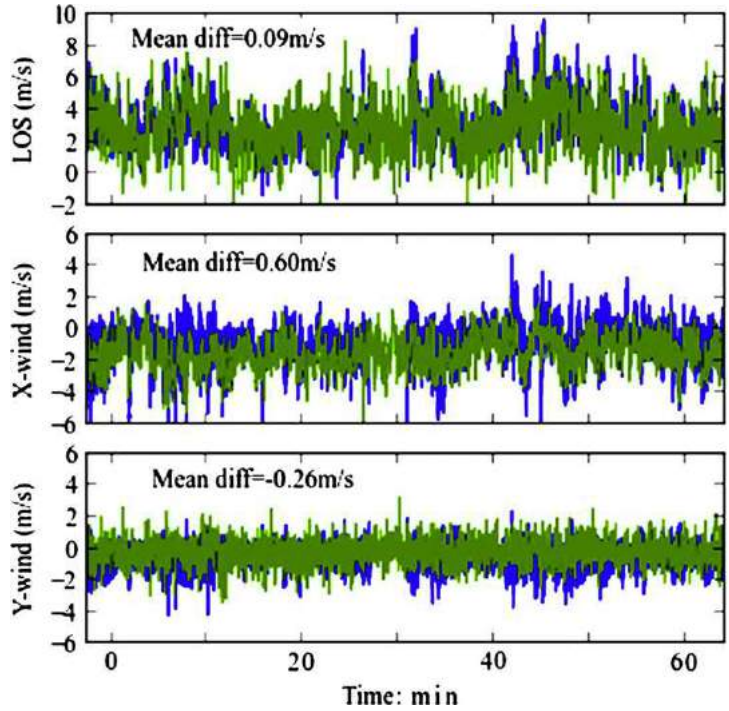


Fig. 37 Three components of lidar wind speed (green line) compared with the sonic anemometer (blue line) from over 2 h of measurements. Reproduced from Prasad, N.S., Mylapore, A.R., 2017. Three-beam aerosol backscatter correlation lidar for wind profiling. Opt. Eng. 56 (3), 031222.

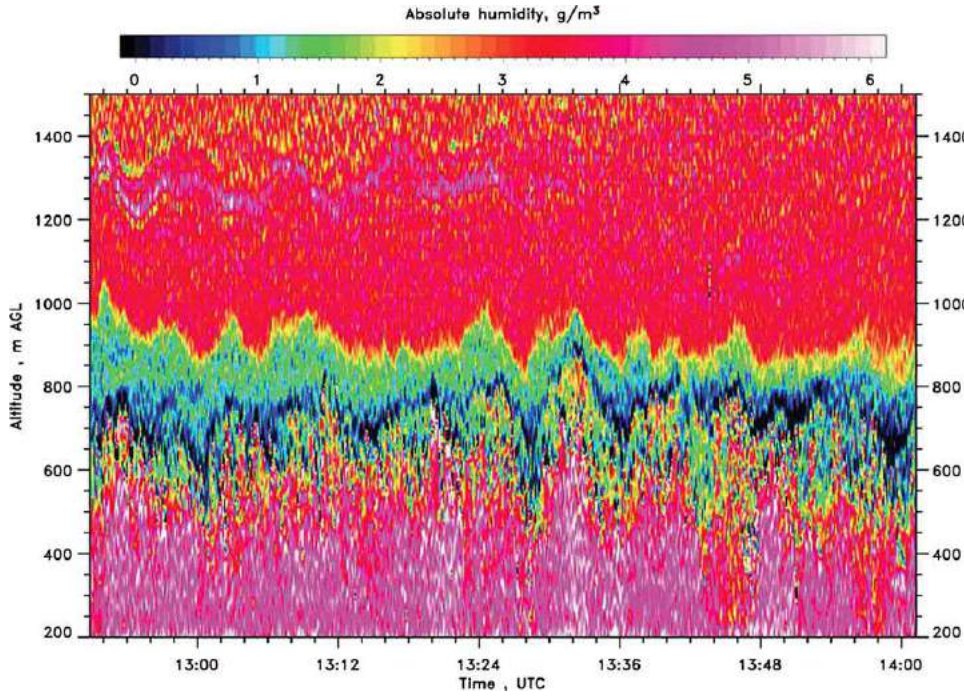


Fig. 38 Vertical-pointing measurement with DIAL. Reproduced from Behrendt, A., Wulfmeyer, V., Riede, A., *et al.*, 2009. 3-Dimensional observations of atmospheric humidity with a scanning differential absorption lidar. Proc. SPIE 7475, 74750L.

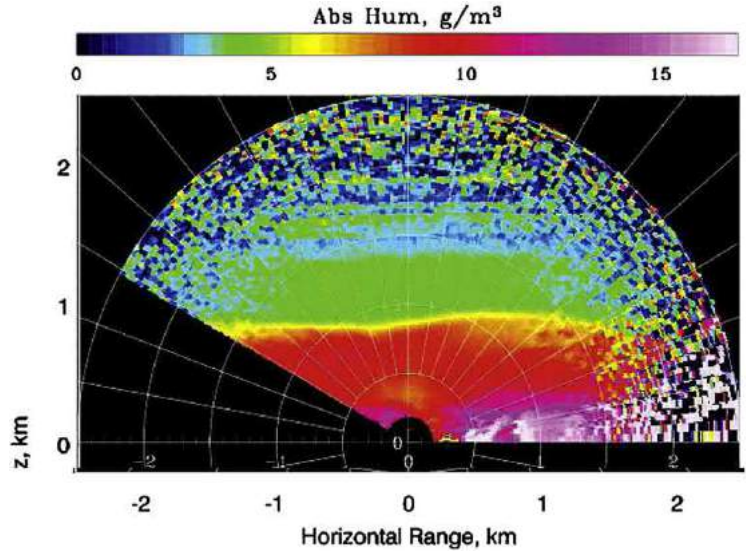


Fig. 39 Range-height indicator scan of the absolute humidity field. Reproduced from Behrendt, A., Wulfmeyer, V., Riede, A., *et al.*, 2009. 3-Dimensional observations of atmospheric humidity with a scanning differential absorption lidar. Proc. SPIE 7475, 74750L.

various weather conditions. The lidar has measured also the Raman water spectrum of an aerosol/liquid water layer. It was noted that under clear weather conditions the Raman water spectrum intensity in the 401.6–404.7 nm range is permanently at a very low level.

The authors used also a 532-nm polarization/Raman lidar to observe the aerosol/liquid water layers in an apparently cloudless atmosphere (**Fig. 41**). These layers are characterized by elevated backscattering ratio R and low depolarization ratio δ (generally $\delta < 0.03$) in a narrow altitude range (with a width generally less than 1 km). The layers usually persist for a few to tens of hours.

It was noted by Mukherjee *et al.* (2010) that a Raman-scattering based instrument is expected to be less sensitive than an instrument based on IR absorption. One of the principal reasons for that is a 1012 times weaker Raman scattering cross-section than an IR absorption cross-section. The wavelength dependence of a Raman-scattering cross-section necessitates the use of short

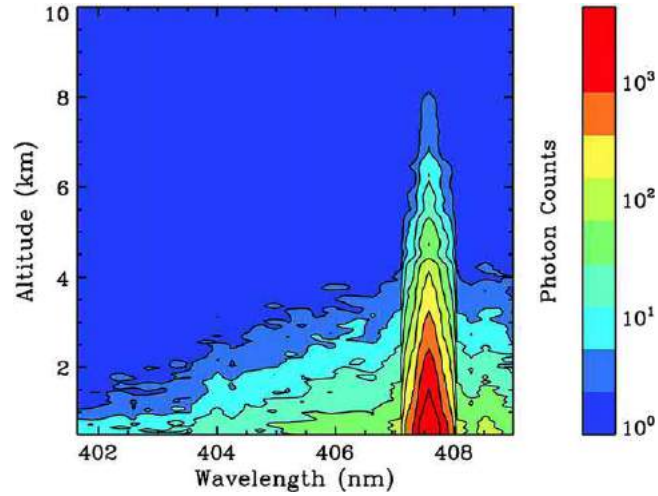


Fig. 40 Altitude-resolved atmospheric Raman water spectrum signal. Reproduced from Liu, F., Yi, F., 2013. Spectrally resolved Raman lidar measurements of gaseous and liquid water in the atmosphere. Appl. Opt. 52, 6884–6895.

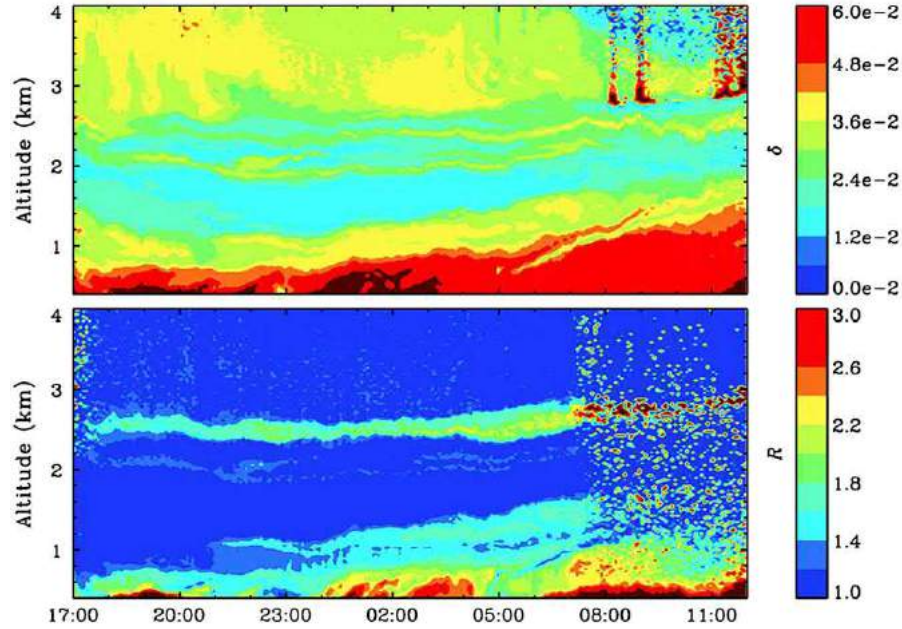


Fig. 41 Time-altitude diagrams of lidar backscatter ratio R and depolarization ratio δ measured by a Raman polarization lidar. Reproduced from Liu, F., Yi, F., 2013. Spectrally resolved Raman lidar measurements of gaseous and liquid water in the atmosphere. Appl. Opt. 52, 6884–6895.

wavelengths, often non-eye-safe, that are not appropriate for atmospheric propagation over long distances. On the other hand, strong optical absorption features of many explosives occur in the long wavelength atmospheric window in 8–12 μm.

For a target sample with absorption characteristics at about 10.6 μm, a wavelength of 10.653 μm (10P26 line from the CO_2 laser) was chosen, which is at the peak of the absorption feature. A spatial x-y scan of the local temperature was undertaken. A temperature rise was recorded only where the sample is located. Thus, the spatial dependence of the temperature rise shows the location of the target residue and its wavelength dependence allows identification of the target substance, even in the presence of other absorbers. Detection and identification of trace amounts of TNT at distances of 150 m were demonstrated. The technique should eventually permit real-time detection of explosives at standoff distances approaching a kilometer with high degrees of specificity and confidence.

To enhance the detection sensitivity, Kamerman (2012) proposed to measure the spectral correlation, i.e., the correlation between the sets of spectral lines, not a single one thus defining the degree of similarity between the received spectrum of unknown contaminants and the spectrum of known components which are to be detected. This technique implemented on the pre-detection stage of optical processing overcomes the problem of small cross-section of target chemicals and makes Raman lidar practical for

many important applications. In particular, the short-range stand-off detection of explosive residue or drugs on the exterior of packages, luggage or vehicles now appear to be possible.

Temperature

Temperature remote sensing using the laser radar techniques is based on measurement either of the Raman scattering or of the differential absorption called differential absorption lidar (DIAL). Murray *et al.* (1980) applied a 10 μm DIAL system that utilizes absorption by naturally occurring CO_2 to make path-averaged measurements of atmospheric temperature. Two widely spaced vibrational-rotational lines, the P(38) and P(20) transitions of the 10 μm band, were chosen for their relative high sensitivity and relative freedom from spectral interference. The transmission of one wavelength relative to the other was shown to be an effective measure of the average temperature over a 5 km path. Using CO_2 lidar and scattering from the foothills, relative transmission over that path was measured. A good correlation was obtained between the lidar-measured temperatures and thermometer-measured values.

The first instruments for rotational Raman temperature measurements were proposed by Cooney (1972). Philbrick (1994) measured the temperature structure of the atmosphere based on the rotational Raman scattering in the region between 526 and 532 nm. Arshinov *et al.* (1983) exploited the double-grating spectrometers to clean the signals from the elastic backscatter. Behrendt and Reichardt (2000) presented a filter polychromator in which the interference filters are mounted sequentially under small angles of incidence. High suppression of the elastic backscatter signal in the rotational Raman detection channels allowed temperature measurements independent of the presence of thin clouds or aerosol layers. Behrendt and Reichardt measured the atmospheric temperature in the northern Sweden up to 40 km heights.

Based on a similar double-grating polychromator approach, Jia and Yi (2014) designed a pure rotational Raman lidar in which the wanted pure Raman rotational signals were extracted, and elastically backscattered light was suppressed. A second-harmonic generation Nd:YAG laser was built with the optical head presented in **Fig. 42** (D – dichroic mirror; RM- reflecting mirror; BP – bandpass filter; L – lens; FMH – fiber mode homogenizer; F – fiber; FA – fiber bundle array; G – grating; PMT – photo-multiplier tube). The instrument allowed to measure the atmospheric temperature at altitudes of 5–30 km. **Fig. 43** compares the results of Behrendt and Reichardt (2000) above acquired in January 1998 in Northern Sweden (Kiruna) with the results of Jia and Yi (2014), acquired in August 2000 in Southern China (Wuhan). Pay attention that at the heights of about 16–17 km the temperatures are equal for both cases.

Fluctuation as a Dimension (Fluctuation Vision)

Atmospheric turbulence provokes much trouble in many applications. Nonetheless, the specificity of fluctuations of the signals propagating in the fluctuating atmosphere and reflected back from specular targets (like optical devices) and from diffuse targets originated the idea of selection of specular targets in the diffuse background, since specular and diffuse targets fluctuate differently when illuminated by the coherent laser light (Molebny, 1993).

It was shown, that these spectra are functions of spatial coherence of laser illumination and relative amplitude of random scanning due to atmosphere turbulence (Molebny *et al.*, 1989). For the diffuse background, the spectra depend also on relative

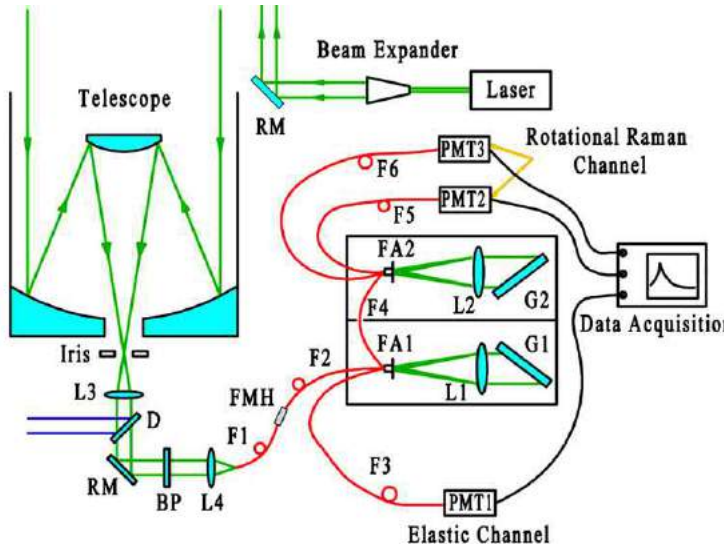


Fig. 42 Optical layout of the rotational Raman lidar (Wuhan University) for atmospheric temperature measurement. Reproduced from Jia, J., Yi, F., 2014. Atmospheric temperature measurements at altitudes of 5–30 km with a double-grating-based pure rotational Raman lidar. Appl. Opt. 53, 5330–5343.

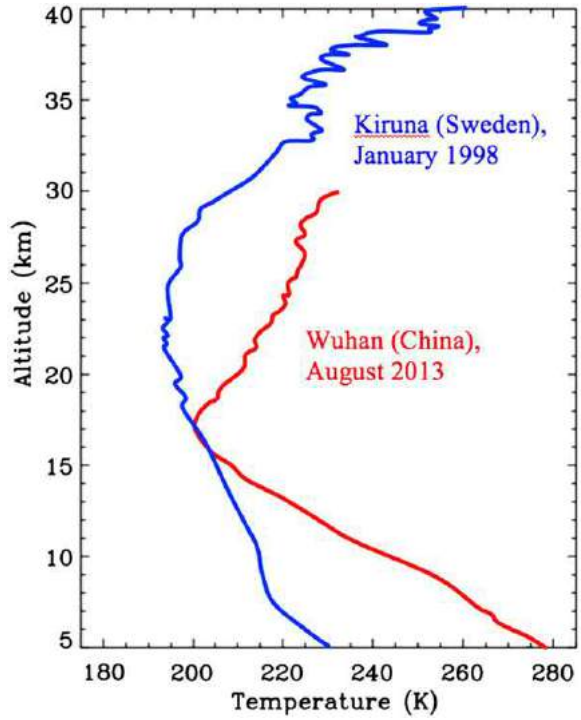


Fig. 43 Comparison of temperature structure for different sites and different seasons. Reproduced from Behrendt, A., Reichardt, J., 2000. Atmospheric temperature profiling in the presence of clouds with a pure rotational Raman lidar by use of an interference-filter-based polychromator. Appl. Opt. 39, 1372–1378 and Jia, J., Yi, F., 2014. Atmospheric temperature measurements at altitudes of 5–30 km with a double-grating-based pure rotational Raman lidar. Appl. Opt. 53, 5330–5343.

scale of background non-homogeneity. Recursive filtration is one of the ways to implement the fluctuation vision using the atmospheric turbulence. The digitized image of the scene is accumulated at the output of the processing unit in such a way that earlier pixels have less weight. For each pixel in the current frame, the difference is calculated between the current value and the value accumulated from the previous frame. Varying the value of the weight factor, the weight of the preceding history can be adjusted to optimize the signal to noise ratio.

The dependence of spectra on random scanning due to atmosphere turbulence can be avoided due to angular micro-scanning of the beam, or by a scanning of the speckle structure inside the beam. This angular scanning or any other kind of sweeping of the internal speckle structure can be implemented at high frequencies thus simplifying the processing of ladar signals.

Dimension Combinations

Velocity and Temperature

Simultaneous measurements of velocity and scalar fields like temperature involve the combination of different techniques, such as particle image velocimetry (PIV) and planar laser induced fluorescence (PLIF). Tsurikov *et al.* (1999) reported measurements of velocity and conserved scalar fields combining PIV and acetone PLIF imaging in a gaseous turbulent flow. In liquids, there are reports of simultaneous velocity and temperature measurements using PIV and PLIF in a water impinging jet (Sakakibara *et al.*, 1997). Laser induced fluorescence of specially involved (seeded) molecules-tracers (tagged velocimetry) allows for simultaneous velocity and temperature measurements in gaseous flow fields (Hsu *et al.*, 2009; Sánchez-González *et al.*, 2012). A transient NO grid in the flow is "written" using the 355 nm photolysis of NO_2, which is subsequently probed by planar laser induced fluorescence imaging to reconstruct velocity maps.

Experimental setup is shown schematically in **Fig. 44**. It contains a photodissociation 355 nm Nd:YAG laser operated at 10 Hz producing a total power of 150 mJ/pulse. The 9 mm diameter laser beam is expanded with a 2.5 × beam expander and directed into the chamber perpendicular to the flow axis through sheeting optics. A 50:50 beam splitter is used to optionally split the beam to direct a second "write" laser sheet parallel to the flow axis to obtain the two-component velocity measurements of the flow field. The photo dissociation laser beams are directed through an aluminum mesh or (in another experiment) a micro-cylindrical lens array to produce a periodic modulated pattern of NO photoproducts in the flow.

Each of the two identical fluorescence (probe) laser systems consists of an injection (seeding) Nd:YAG laser operated at 10 Hz. The 532 nm output is used to pump a pulsed dye laser to produce a tunable output beam in a range from 600 to 630 nm using a

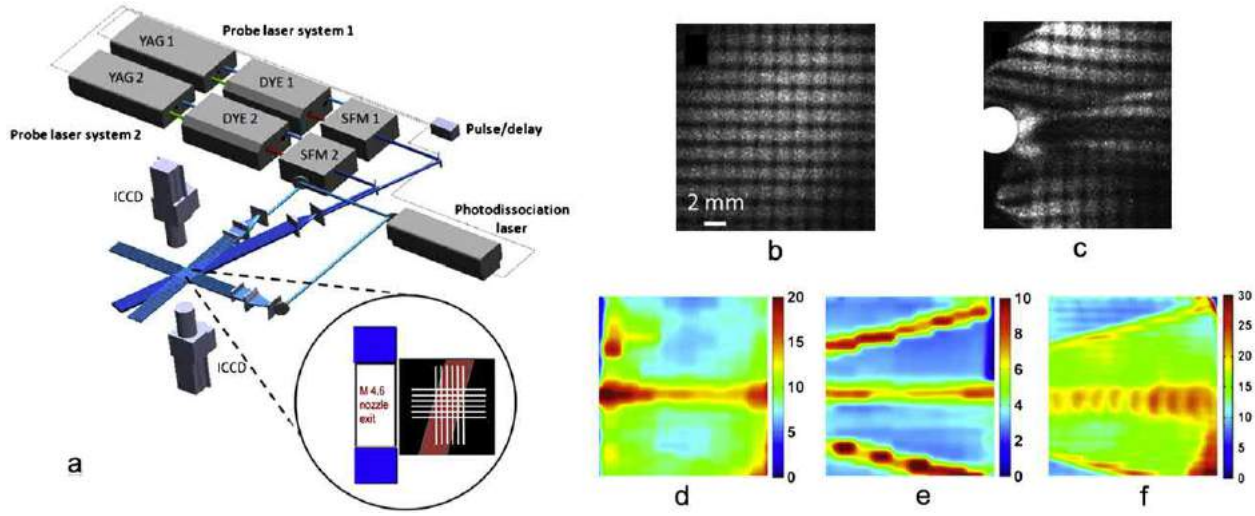


Fig. 44 Schematic of the experiments in the repetitively pulsed hypersonic flow (a). Instantaneous images probing the transition of the vibrationally excited NO in the free stream (b) and in the flow over a sphere (c). Components of velocity u/u_0 and v/v_0, (d, e) and temperature T/T_0 (f) fluctuations normalized to the free stream values (in percentage). Reconstructed from Sánchez-González, R., Bowersox, R.D.W., North, S.W., 2012. Simultaneous velocity and temperature measurements in gaseous flow fields using the vibrationally excited nitric oxide monitoring technique: A comprehensive study. Appl. Opt. 51, 1216–1228.

solution of Rhodamine 610 and Rhodamine 640 in methanol. The dye laser output is mixed with the residual 355 nm beam in a frequency mixing unit to produce approximately 10 mJ/pulse in a range from 223 to 227 nm. This output wavelength range permits the probing of fluorescence transitions. For the experiments, the probe beams are directed into the chamber at an angle of 70° from the streamwise flow direction to avoid the aluminum mesh or the microlens array used for the "write" beam. The fluorescence images were acquired with two ICCD cameras mounted on either side of the chamber perpendicular to the laser sheets.

The measurements are performed during the steady flow phase of the pulsed flow. The basic timing sequence consists of an initial laser pulse, the "write" pulse, that photodissociates NO_2 using 355 nm. After a time delay, a first "read" dye laser excites a specific rotational state of the NO photoproduct, and an associated ICCD camera captures a fluorescence image of the flow. After a second time delay, the second "read" system excites a different NO rotational state to capture a second fluorescence image to get a 2D distribution.

A true simultaneous velocity/temperature measurement is possible when the velocity is determined using the instantaneous images generated by sequential "read" pulses. In **Fig. 44**, diagrams to the right demonstrate measurements in both the Mach 4.6 free stream (b) and in the wake of the Mach 4.6 flow over a cylinder (c). The images (b) and (c) are instantaneous fluorescence shots. The flow movement is from left to right. The cylinder position is shown as a white circle.

Conclusion

Every new dimension of the acquired information can be compared to a new window in the house. The more windows, the brighter the room is. We outlined the main stages of information extension: from a simple single-dimension range measurement to determining the three-coordinate object location in the space, adding velocity and vibrations as dimensions, determining the parameters of the media in which the laser radiation propagates, combining several dimensions from the same laser radar or from several laser radar systems. Velocity and vibrations can be described as single-dimensional (radial component), two-dimensional (lateral x-y directions), or full three-dimensional. Vibrations can have a fourth (spectral) dimension. Fluctuations of the signal reflected from the target can provide still another dimension of the information describing the target. Polarization is a source of information in laser radars that can be used as a dimension in different applications, e.g., remote measurement of object orientation (Molebny, 1975) or its shape (García-Arellano *et al.*, 2017), clutter suppression (Sassen, 2003), object or its material identification (Raghavan *et al.*, 2008), etc. We sketched a hypothetical multi-dimensional laser radar system which is the combination of 3D space coordinates plus 3D velocity in each point of the space + time dimension. This dimensionality can be even broader if temperature or another medium parameter is added. It is very productive to combine the information from laser radar with information from other sources: acoustic (Donzier and Cadavid, 2005; Lindgren *et al.*, 2011), microwave (Duckworth *et al.*, 1995; Yang and Qianqian, 2009), infrared (Tong *et al.*, 1987; Bartlett *et al.*, 2017), etc.

References

AGM-129 ACM. Directory of U.S. Military Rockets and Missiles. Raytheon (General Dynamics) AGM-129 ACM. Available at: http://www.designation-systems.net/dusrm/m-129.html.

Albota, M.A., Aull, B.F., Fouche, D.G., *et al.*, 2002b. Three-dimensional imaging laser radars with Geiger-mode avalanche photodiode arrays. Lincoln Lab. J. 13 (2), 351–370.

Albota, M.A., Heinrichs, R.M., Kocher, D.G., *et al.*, 2002a. Three-dimensional imaging laser radar with a photon-counting avalanche photodiode array and microchip laser. Appl. Opt. 41, 7671–7678.

Amzajerdian, F., Petway, L., Hines, G., *et al.*, 2012. Fiber Doppler lidar for precision navigation of space vehicles. Lasers, Sources, and Related Photonic Devices Tech. Digest. doi:10.1364/FILAS.2012.FTh3.1A.2.

Arshinov, Y.F., Bobrovnikov, S.M., Zuev, V.E., Mitev, V.M., 1983. Atmospheric temperature measurements using a pure rotational Raman lidar. Appl. Opt. 22, 2984–2990.

Baker, I., Owton, D., Trundle, K., *et al.*, 2008. Advanced infrared detectors for multimode active and passive imaging applications. Proc. SPIE 6940, 69402L.

Bartlett, P.W., Coblenza, L., Sherwina, G., *et al.*, 2017. A custom, multi-modal sensor suite and data analysis pipeline for aerial field phenotyping. Proc. SPIE 10218, 1021804.

Behrendt, A., Reichardt, J., 2000. Atmospheric temperature profiling in the presence of clouds with a pure rotational Raman lidar by use of an interference-filter-based polychromator. Appl. Opt. 39, 1372–1378.

Behrendt, A., Wulfmeyer, V., Riede, A., *et al.*, 2009. 3-Dimensional observations of atmospheric humidity with a scanning differential absorption lidar. Proc. SPIE 7475, 74750L.

Bilbro, J., 1984. Airborne Doppler lidar wind field measurements. Bull. Am. Meteorol. Soc. 65, 348–359.

Bilbro, J.W., Vaughan, W.W., 1978. Wind field measurement in the non-precipitous regions surrounding severe storms by an airborne pulsed Doppler lidar system. Bull. Am. Meteorol. Soc. 59, 1095–1100.

Bovenkamp, E.G.P., Schutte, K., 2010. Laser gated viewing: An enabler for automatic target recognition. Proc. SPIE 7684, 768435.

Bradbury, S.M., Mirzoyan, R., Gebauer, J., *et al.*, 1997. Test of the new hybrid INTEVAC intensified photocell for the use in air Cherenkov telescopes. Nucl. Instrum. Methods Phys. Res. A 387, 45–49.

Brigety, R.E., 2007. Ethics, Technology, and the American Way of War. Cruise Missiles and US Security Policy. London, New York: Routledge.

Busck, J., Heiselberg, H., 2004. Gated viewing and high-accuracy three-dimensional laser radar. Appl. Opt. 43, 4705–4710.

Cho, K.Y., Satija, A., Pourpoint, T.L., Son, S.F., Lucht, R.P., 2014. High-repetition-rate three-dimensional OH imaging using scanned planar laser-induced fluorescence system for multiphase combustion. Appl. Opt. 53, 316–326.

Cho, P., Anderson, H., Hatch, R., Ramaswami, P., 2006. Real-time 3D ladar imaging. Lincoln Lab. J. 16 (1), 147–164.

Cooney, J., 1972. Measurement of atmospheric temperature profiles by Raman backscatter. J. Appl. Meteorol. 11, 108–112.

Donovan, D.P., Apituley, A., 2013. Practical depolarization-ratio-based inversion procedure: Lidar measurements of the Eyjafjallajökull ash cloud over the Netherlands. Appl. Opt. 52, 2394–2415.

Donzier, A., Cadavid, S., 2005. Small arm fire acoustic detection and localization systems: Gunfire detection system. Proc. SPIE 5778, 245–253.

Duckworth, G.L., Frey, M.L., Remer, C.E., *et al.*, 1995. Comparative study of nonintrusive traffic monitoring sensors. Proc. SPIE 2344, 16–29.

Electron Bombarded Active Pixel Sensor (EBAPS). Available at: http://www.intevac.com.

Eloranta, E.W., King, J.M., Weinman, J.A., 1975. The determination of wind speeds in the boundary layer by monostatic lidar. J. Appl. Meteorol. 14, 1485–1489.

Fischer, K.W., Abreu, V.J., Skinner, W.R., *et al.*, 1995. Visible wavelength Doppler lidar for measurement of wind and aerosol profiles during day and night. Opt. Eng. 34 (2), 499–511.

Flom, T., 1972. Spaceborne laser radar. Appl. Opt. 11, 291–299.

Fraczek, M., Behrendt, A., Schmitt, N., 2012. Laser-based air data system for aircraft control using Raman and elastic backscatter for the measurement of temperature, density, pressure, moisture, and particle backscatter coefficient. Appl. Opt. 51, 148–166.

García-Arellano, A., Cruz-Santos, W., García-Arellano, G., Juvenal Rueda-Pazb, J., 2017. One-shot shape measurement of small objects with a pulsed laser and modulation of polarization. Opt. Eng. 56 (6), 064102.

Hair, J.W., Hostetler, C.A., Cook, A.L., *et al.*, 2008. Airborne high spectral resolution lidar for profiling aerosol optical properties. Appl. Opt. 47, 6734–6753.

Halo Photonics. Available at: http://halo-photonics.com/index.htm.

Henderson, S.W., Hale, C.P., Huffaker, R.M., *et al.*, 1993. Eyesafe coherent laser radar for velocity and position measurements, US Pat. 5,237,331 (Aug. 17, 1993).

Herffurth, T., Schröder, S., Trost, M., *et al.*, 2013. Comprehensive nanostructure and defect analysis using a simple 3D light-scatter sensor. Appl. Opt. 52, 3279–3287.

Heuvel van den, J.C., Pace, P., Bekman, H.H., *et al.*, 2008. Experimental validation of ship identification with a laser range profiler. Proc. SPIE 6950, 69500V.

Hsu, A., Srinivasan, R., Bowersox, R., North, S., 2009. Two-component molecular tagging velocimetry utilizing NO fluorescence lifetime and NO_2 photodissociation techniques in an under expanded jet flow field. Appl. Opt. 48, 4414–4423.

Jia, J., Yi, F., 2014. Atmospheric temperature measurements at altitudes of 5–30 km with a double-grating-based pure rotational Raman lidar. Appl. Opt. 53, 5330–5343.

Kamerman, G.W., 2012. Optical correlation spectroscopy for remote contaminant detection. Bul. (Visnyk) Natl. Tech. Univ. Ukraine "KPI", Instum. Making, Kiev 43, 61–71.

Lockheed Martin. Available at: http://www.lockheedmartin.com/products/WindTracer/index.html.

Leosphere. Available at: http://www.leosphere.com/8,wind-energy.

Lindgren, D., Bank, D., Carlsson, L., *et al.*, 2011. Multisensor configurations for early sniper detection. Proc. SPIE 8186, 81860D.

Liu, F., Yi, F., 2013. Spectrally resolved Raman lidar measurements of gaseous and liquid water in the atmosphere. Appl. Opt. 52, 6884–6895.

Lutzmann, P., Frank, R., Ebert, R., 2000. Laser radar based vibration imaging of remote objects. Proc. SPIE 4035, 436–443.

Lutzmann, P., Göhler, B., Hill, C.A., Putten, F., 2017. Laser vibration sensing at Fraunhofer IOSB: Review and applications. Opt. Eng. 56 (3), 031215.

Lutzmann, P., Göhler, B., Putten, F., Hill, C.A., 2011. Laser vibration sensing: Overview and applications. Proc. SPIE 8186, 818602.

Mann, J., Angelou, N., Arnqvist, J., *et al.*, 2017. Complex terrain experiments in the New European Wind Atlas. Phil. Trans. R. Soc. A 375, 20160101.

Matvienko, G.G., Grishin, A.I., Kharchenko, O.V., Romanovskii, O.A., 2006. Application of laser-induced fluorescence for remote sensing of vegetation. Opt. Eng. 45, 056201.

Morpheus lander. Available at: https://morpheuslander.jsc.nasa.gov.

Mikkelsen, T., 2014. Lidar-based research and innovation at DTU wind energy – A review. J. Phys.: Conf. Ser. 524, 012 007.

Molebny, V., Steinvall, O., 2013. Laser remote sensing. Velocimetry based techniques. In: Tuchin, V.V. (Ed.), Handbook of Coherent-Domain Optical Methods. New York: Springer, pp. 363–395.

Molebny, V., Steinvall, O., 2014. Multi-dimensional laser radars. Proc. SPIE 9080, 908002.

Molebny, V., 2013. Wavefront measurement in ophthalmology. Aberrometry through the eyes of an engineer. [Chapter 9] In: Tuchin, V.V. (Ed.), Handbook of Coherent-Domain Optical Methods. New York: Springer Science + Business Media, pp. 315–361.

Molebny, V., Zarubin, P., Kamerman, G., 2010. The dawn of optical radar: A story from another side of the globe. Proc. SPIE 7684, 76840B.

Molebny, V.V., 1975. Remote sensing of the orientation of the facets of sea surface, USSR Patent 433339, Bulletin # 23, 20.06.1975.

Molebny, V.V., 1993. Image synthesis from fluctuations: Fluctuation vision. Proc. SPIE 2029, 68–76.

Molebny, V.V., Protasov, V.G., Steba, A.M., 1989. Fluctuation spectra of light reflected from specular and diffuse surfaces. Herald (Visnyk) Kiev Univ., Physics 30, 92–100.

Mukherjee, A., Von der Porten, S., Patel, C.K.N., 2010. Standoff detection of explosive substances at distances of up to 150 m. Appl. Opt. 49, 2072–2078.

Murray, E.R., Powell, D.D., van der Laan, J.E., 1980. Measurement of average atmospheric temperature using a CO_2 laser radar. Appl. Opt. 19, 1794–1797.

NASA. Making the skies safe from windshear: Langley-developed sensors will help improve air safety. Available at: https://www.nasa.gov/centers/langley/news/factsheets/Windshear.htm.

Palm, S.P., Melfi, S.H., Carter, D.L., 1994. New airborne scanning lidar system: Applications for atmospheric remote sensing. Appl. Opt. 33, 5674–5681.

Perea, J., Libbey, B., 2016. Development of a heterodyne speckle imager to measure 3 degrees of vibrational freedom. Opt. Express 24, 8253–8265.

Philbrick, C.R., Hallen, H.D., 2017. Signatures of dynamical processes in Raman lidar profiles of the atmosphere. Proc. SPIE 10191, 101910E.

Philbrick, C.R., 1994. Raman lidar measurements of atmospheric properties. Proc. SPIE 2222, 922–931.

Philbrick, R., Hallen, H., Wyant, A., *et al*., 2010. Optical remote sensing techniques characterize the properties of atmospheric aerosols. Proc. SPIE 7684, 76840J.
Pierrottet, D., Amzajerdian, F., Petway, L., *et al*., 2009. Flight test performance of a high precision navigation Doppler lidar. Proc. SPIE 7323, 732311.
Polytec. 2006. Polytec's vibrometers are indispensable tools to optimize parts and goods and to investigate natural dynamic processes. Available at: http://www.polytec.com.
Prasad, N.S., Mylapore, A.R., 2017. Three-beam aerosol backscatter correlation lidar for wind profiling. Opt. Eng. 56 (3), 031222.
Raghavan, M., Morris, M.D., Sahar, N.D., Kohn, D.H., 2008. Polarized Raman spectroscopy: Application to bone biomechanics. Proc. SPIE 6853, 68530W.
Reichardt, J., Wandinger, U., Klein, V., *et al*., 2012. RAMSES: German meteorological service autonomous Raman lidar for water vapor, temperature, aerosol, and cloud measurements. Appl. Opt. 51, 8111–8131.
Revel, G.M., Castellini, P., Chiariotti, P., *et al*., 2011. Laser vibrometry vibration measurements on vehicle cabins in running conditions: Helicopter mock-up application. Opt. Eng. 50, 101502.
Sakakibara, J., Hishida, K., Maeda, M., 1997. Vortex structure and heat transfer in the stagnation region of an impinging plane jet (simultaneous measurement of velocity and temperature fields by digital particle image velocimetry and laser-induced fluorescence). Int. J. Heat Mass Transfer 40, 3163–3176.
Sánchez-González, R., Bowersox, R.D.W., North, S.W., 2012. Simultaneous velocity and temperature measurements in gaseous flow fields using the vibrationally excited nitric oxide monitoring technique: A comprehensive study. Appl. Opt. 51, 1216–1228.
Sassen, K., 2003. Polarization in lidar: A review. Proc. SPIE 5158, 151–160.
Sato, T., Suzuki, Y., Kashiwagi, H., Nanjo, M., Kakui, Y., 1978. Laser radar for remote detection of oil spills. Appl. Opt. 17, 3798–3803.
Schmidt, J., Wandinger, U., Malinka, A., 2013. Dual-field-of-view Raman lidar measurements for the retrieval of cloud microphysical properties. Appl. Opt. 52, 2235–2247.
Sitter, D.N., Gelbart, A., 2001. Laser-induced fluorescence imaging of the ocean bottom. Opt. Eng. 40 (8), 1545–1553.
Spinhirne, J.D., Hansen, M.Z., Caudill, L.O., 1982. Cloud top remote sensing by airborne lidar. Appl. Opt. 21, 1564–1571.
Steinvall, O., Tulldahl, M., 2017. Laser range profiling for small target recognition. Opt. Eng. 56 (3), 031206.
Steinvall, O., Persson, R., Berglund, F., Gustafsson, O., Gustafsson, F., 2014. Using an eyesafe military laser range finder for atmospheric sensing. Proc. SPIE 9080.[9080-32].
Stettner, R., 2010. Compact 3D flash lidar video cameras and applications. Proc. SPIE 7684, 768405.
Takeda, M., Motoh, K., 1983. Fourier transform profilometry for the automatic measurement of 3D object shapes. Appl. Opt. 22, 3977–3982.
Targ, R., Kavaya, M.J., Huffaker, R.M., Bowles, R.L., 1991. Coherent lidar airborne windshear sensor: Performance evaluation. Appl. Opt. 30, 2013–2026.
Targ, R., Steakley, B.C., Hawley, J.G., *et al*., 1996. Coherent lidar airborne wind sensor II: Flight-test results at 2 and 10 μm. Appl. Opt. 35, 7117–7127.
Tong, C.W., Rogers, S.K., Mils, J.P., Kabrisk, M.K., 1987. Multisensor data fusion of laser radar and forward looking infrared (FLIR) for target segmentation and enhancement. Proc. SPIE 782, 10–19.
Tsurikov, M.S., Rehm, J.E., Clemens, N.T., 1999. High-resolution PIV/PLIF measurements of a gas-phase turbulent jet. In: Proceedings of the 37th Aerospace Sciences Meeting, AIAA 99-0930.
Uthe, E.E., 1991. Elastic scattering, fluorescent scattering, and differential absorption airborne lidar observations of atmospheric tracers. Opt. Eng. 3 (1), 66–71.
Vasiljevic, N., Palma, J., Angelou, N., *et al*., 2017. Perdigão 2015: Methodology for atmospheric multi-Doppler lidar experiments. Atmos. Meas. Tech. Discuss. 1–28. doi:10.5194/amt-2017-18.
Verghese, S., McIntosh, K.A., Liau, Z.L., *et al*., 2009. Arrays of 128×32 InP-based Geiger-mode avalanche photodiodes. Proc. SPIE 7320, 73200M.
Wang, J.Y., Bartholomew, B.J., Streiff, M.L., Starr, E.F., 1984. Imaging CO_2 laser radar field tests. Appl. Opt. 23, 2565–2571.
Werner, C., Flamant, P.H., Reitebuch, O., *et al*., 2001. Wind infrared Doppler lidar instrument. Opt. Eng. 40 (1), 115–125.
Wilkerson, T., Bradford, B., Marchant, A., *et al*., 2009. VisibleWindTM: A rapid-response system for high-resolution wind profiling. Proc. SPIE 7460, 746009.
Wong, C.M., Logan, J.E., Bracikowski, C., Baldauf, B.K., 2010. Automated in-track and cross-track airborne flash LADAR image registration for wide-area mapping. Proc. SPIE 7684, 768427.
Yang, W., Qianqian, W., 2009. The application of lidar in detecting the space debris. Proc. SPIE 7160, 71601S.
ZephIR. Available at: https://www.zephirlidar.com.
Zhang, Q., Su, X., 2005. High-speed optical measurement for the drumhead vibration. Opt. Express 13, 3110–3116.

A Review of Laser Range Profiling for Target Recognition

Ove Steinvall, Swedish Defense Research Agency (FOI), Linköping, Sweden

Introduction

The classification and identification of small target, and/or targets at long range is a prime need for defense and security. Examples involve targets such as aircraft, unmanned aerial vehicles (UAVs) and missiles as well as small boats at sea and in littoral waters which have become common in illegal immigration, piracy, drug trafficking, and asymmetric threats.

Radar is the dominant sensor for long-range target detection and also offers model- and feature-based classification techniques using high resolution range profiles (HRRPs), target micro-Doppler spectra, range-Doppler imaging, and inverse synthetic aperture radar (ISAR) imaging. Techniques based on target tracking and target behavior are also studied for recognition.

Laser radar (or ladar) can also accomplish small target, or long range, target classification. Laser radar can provide narrow beams resolving multiple targets and can also provide recognition capabilities based on vibration, 1D range profiling, as well as 2D or 3D imaging. In this paper, we will concentrate on range profiling. Compared with its radar counterpart (HRRP) laser radars can offer more stable signals where fluctuations of the range-profile amplitude due to speckle often causes problems in classification.

The speckle problem is much less pronounced for a direct detection optical system, and very high range resolution in the centimeter range can be obtained. This makes a profiling laser radar an interesting candidate for long-range target, or small target recognition, where the transverse resolution for imaging sensors is limited to a single or a few pixels. A laser range finder is today a standard sensor in most EO systems and can have a profiling capability, or be modified to be more suitable for this task.

1D laser techniques for target recognition include vibrometry (Lutzmann *et al.*, 2011) and laser range profiling (LRP) including its extension to tomography. LRP is attractive because the maximum range can be substantial, especially for a small beam width while still preserving a high range resolution. LRP will usually work together with an IR and radar sensor which acquires the target. A laser profiler can complement a jammed radar, or a silent radar wanting to limit the effectiveness of anti-radiation missiles. It can also extend the capability of an IR-sensor to include small and long-range target classification. LRP can be used in a scanning mode to detect targets within a certain sector. The same laser can be used for illumination in active imaging when the target comes closer and is angular resolved.

LRP has been investigated for both search and ID of small surface targets (Kunz *et al.*, 2005; van den Heuvel *et al.*, 2009a; Schoemaker and Benoist, 2011) as well as for aircraft ID (van den Heuvel *et al.*, 2008; Dierking *et al.*, 1998). Recently, we showed by simulations and laboratory measurements that LRP is a promising method of identification of small sea surface targets at long ranges (Steinvall *et al.*, 2012b, 2014; Steinvall and Tulldahl, 2016). A series of laser radar range profiles collected at various aspect angles allows for a tomographic reflective reconstruction of the target (Knight *et al.*, 1989b). Many examples of this are found from simulations and laboratory measurements (Knight *et al.*, 1989a,b; Jin and Levine, 2009; Parker *et al.*, 1988; Henriksson *et al.*, 2012a,b) and other demonstrate field results (Murray *et al.*, 2010). Many of the applications seem to be devoted to satellite imaging (Lasche *et al.*, 2009; Matson and Mosley, 2001). Using time correlated single-photon counting, very high range resolution (millimeter–centimeter) profiling and tomography can be demonstrated (Sjöqvist *et al.*, 2014).

This review article starts with reviewing the theoretical basis of LRP, and relating this theoretical basis to the corresponding techniques for microwave radar. Examples of simulations, and related uncertainty in building a realistic library, to compare measured with stored data, will be discussed. Experimental results will be given for different types of targets and compared with simulated data. Furthermore, examples in signal processing techniques will be discussed. Finally, conclusion and future prospects for the technique will be presented.

Background for LRP

LRP is based on sending out a short pulse, or modulating the laser emission, so that high range resolution is obtained. Depending on the target depth structure the resolution may be high enough to profile this structure, typically meaning at least 10 resolution cells over a target, such as an aircraft.

Direct detection systems, such as present military range finders, offer target maximum ranges in the 10–20 km range and a range resolution in the region 0.2–1 m. The number of speckle lobes N_{sp} falling within the receiver aperture D_{ap} from the target plane from a certain range cell is at a first approximation

$$N_{sp} \sim \left[\frac{D_{ap} D_{target}}{\lambda R}\right]^2 \tag{1}$$

where D_{target} is the illuminated target area, λ the laser wavelength, and R the target range. For $D_{ap}=0.1$ and $D_{target}=1$ m the range for which $N_{sp}=1$ is about 100 km and for $D_{target}=0.1$ m, $R=10$ km. Thus, for very small targets, and long ranges, the speckle noise might be present from a diffuse target. On the other hand, for a glint target turbulence fluctuations will dominate. However, pulse

 doi:10.1016/B978-0-12-803581-8.09429-7

averaging over 1–3 s will average out most of the turbulence and target speckle fluctuations. This has been demonstrated in both measurements and simulation when range profiling of small surface vessels (Steinvall *et al.*, 2012b; Steinvall and Tulldahl, 2016).

In order to estimate the range performance of a laser profiling system indicative of a laser range finder type we will start by expressing the detected laser power (P_{det}) written as

$$P_{\mathrm{det}} = T_{\mathrm{optics}} \cdot P_0 \cdot \frac{\rho_{\mathrm{target}}}{\pi} \cdot \frac{A_{\mathrm{rec}}}{R^2} \cdot \frac{4 \cdot A_{\mathrm{target}}}{\pi(\phi R)^2} \cdot T^2_{\mathrm{atm.laser}} \tag{2}$$

where P_0 is the transmitted peak power, R the target range, ϕ the laser divergence, ρ_{target} the target reflectance, and $T_{\mathrm{atm\cdot laser}} = \exp(-\sigma R)$ the one-way atmospheric laser transmission. The signal-to-noise ratio (SNR) is given by $SNR = P_{\mathrm{det}}/NEP$. The maximum range for the system as a function of atmospheric extinction can easily be obtained with the help of the Lambert-W function (Steinvall, 2009).

We will choose some sensor data indicative of an eye-safe 1550 nm laser range finder. The laser pulse energy is set to 30 mJ and the pulse length 6 ns, giving a peak power of about 10 MW. The receiver area is assumed to be 0.02 m^2 corresponding to a diameter of 16 cm, and the receiver noise equivalent power noise equivalent power (NEP) was 3 nW. The optical background is in general lower than this figure. The relation between the atmospheric extinction coefficient σ at 1550 nm wavelength and the visibility V was assumed to be $\sigma = 0.996(3/V)^{1.2}$ according to Hutt (Hutt *et al.*, 1994).

Fig. 1 shows the maximum range vs. visibility for different combinations of target strength $A_{\mathrm{target}} \times \rho_{\mathrm{target}}$ m^2 and the case of an unresolved target. As seen the range can be substantial during a wide range of visibilities. The threshold for the SNR was set to 7. Averaging over 10–100 pulses will improve the SNR approximately as the square root of the number of pulses provided that the individual pulses are summed up with the same "start" bin.

Time correlated single photon counting (TCSPC) is a special form of a direct detection laser radar where a single detected photon will cause the detector to saturate. This techniques enables millimeter–centimeter range resolution. In TCSPC, this is accomplished by carefully measuring the time between a laser pulse sync signal and registration of a single-photon event of photons reflected from a target. The measurement is performed multiple times and a histogram of arrival times is computed to gain information about surfaces at different distances within the field of view (FOV) and to exclude spurious detections from detector and background noise. Systems using moderate pulse repetition rates and a limited number of acquisitions are usually called Geiger-mode APD (GMAPD) laser radar, whereas systems using laser pulse repetition rates in the megahertz range, that is, more acquisitions and lower detection probabilities, are termed TCSPC laser radar with single-photon avalanche photodiode (SPAD) detectors. We refer to Sjöqvist *et al.* (2014) for more detailed discussion of these techniques for range profiling and reflectance tomographic imaging.

For a range profile using GMAPDs the number of detected photons per pulse must be kept low, typically 0.1 photon. Otherwise the first photon will block pulses from close lying structures due to the dead time in this form of detection (Geiger mode).

To estimate the maximum range performance the SNR expression is not quite useful. Instead, the detection and false alarm probability ($P_{\mathrm{det}}/P_{\mathrm{fa}}$) should be used. An example of such an analysis is made in Steinvall *et al.* (2012a) where $P_{\mathrm{det}}/P_{\mathrm{fa}}$ is plotted versus range for a visibility $V = 10$ km and a 1 m^2 diffuse 10% reflectivity target. The average power was 1 mW for the 200 kHz laser with 6 ps pulse length and a 10 cm receiver aperture. The dwelltime T_{dwell} was assumed to be 1 ms which is enough to get a good detection probability. On the other hand, 10–100 ms may be more suitable to collect the range waveform details. We can see from **Fig. 2** that both 1060 and 1550 nm laser radars have a good range capability using photon counting with only 1 mW of average power and an integration time $T_{\mathrm{dwell}} = 1$ ms.

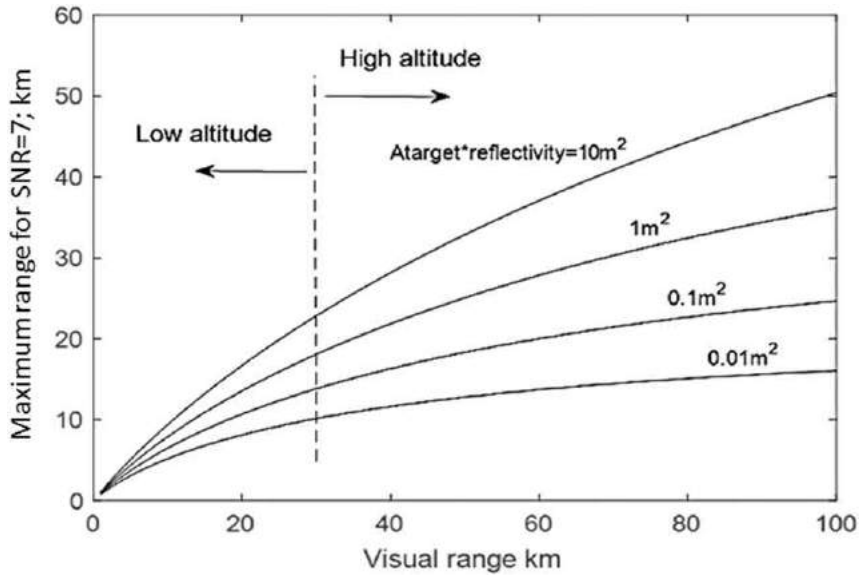


Fig. 1 Range performance for a typical profiling system corresponding to a laser range finder at 1550 nm.

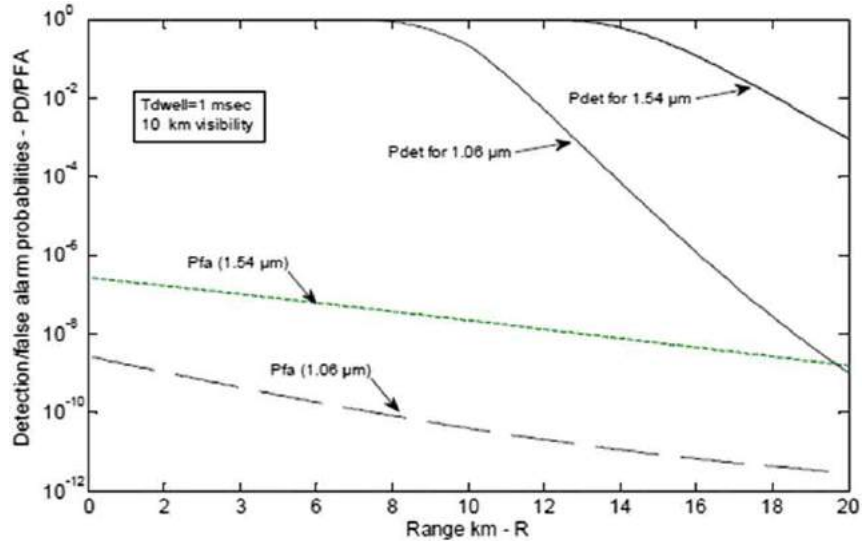


Fig. 2 Comparison between 1.06 and 1.5 μm ladar performance using the parameters according to the text above and in Steinvall *et al.* (2012) for a diffuse 10% reflectivity 1 m^2 target. Reproduced from Steinvall, O., Sjöqvist, L., Henriksson, M., 2012. Photon counting ladar work at FOI, Sweden. Proceedings of SPIE 8375, 83750C.

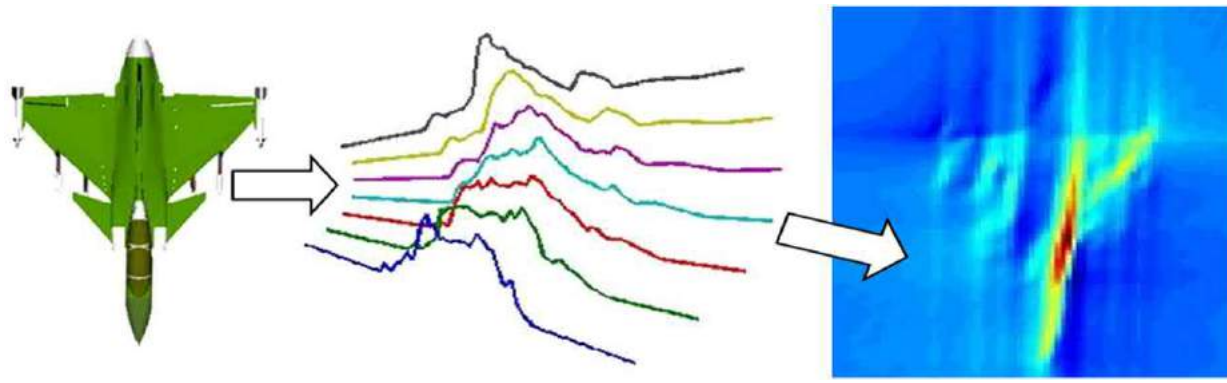


Fig. 3 Simulated laser radar waveforms from an aircraft and reconstruction of the tomographic image using these waveforms. Reproduced from Steinvall, O., Tulldahl, M., 2016. Laser range profiling for small target recognition. Optical Engineering Vol. 56 (3), 031206.

An attractive feature of LRP is the possibility to fuse several range profiles taken at different angles of incidence to reconstruct an image of the object (Murray *et al.*, 2010; Parker *et al.*, 1988). A tomographic image can be obtained by using the Radon transform to reconstruct the shape of the target. Satellite imaging by tomography has been demonstrated with a pulsed, mode locked, coherent system (Matson and Mosley, 2001). Frequency chirp techniques for obtaining high range resolution are also a well known method in coherent laser radar. Other techniques like range correlation methods (pulse compression) have been exploited by others using pseudo-random pulse codes. This was demonstrated in Murray *et al.* (2010) for image reconstruction of a meter sized target at 22.4 km range with 15 cm resolution.

Coherent detection also enables the use of the target rotation, and the generated Doppler spectrum, to determine the object size and rotation speed from 1D measurements related to range profiling.

Fig. 3 shows an example from simulations using several laser range rofiles from different aspects to form a tomographic image. However, the tomographic technique has its limitations in practical systems and is probably most useful for a fixed sensor and a rotation target around its own axis (Henriksson *et al.*, 2012a,b).

Simulating Laser Profiling Data

In order to enable target recognition from range profiling the measured range waveform has to be compared with that from a library containing a number of potential targets. The stored waveforms should contain all possible aspect angles of interest. This fact will widen the library content. Coupled to this data storage is also the question if the whole waveform,

Fig. 4 Left: a point cloud model of aircraft SK37 Viggen using the Riegl VZ-400 laser scanner. Right: the polygon model built with the point cloud as a reference. Images FOI.

or only its features, like individual peak positions and the relative amplitude of peaks should be stored. The best is of course if the whole waveform in addition to the characteristic features can be stored because that will enable a more robust signal processing.

As direct measurements of the target profiles over all aspects are unrealistic for many reasons, the modeling of different target waveforms is inevitable. This can be accomplished if a good 3D CAD model of the target is available. The surface angular reflectivity characteristics at the laser wavelength of interest is given from the bi-directional reflection distribution function (BRDF). This function is hard to measure in practice and often has to be modeled based on known or assumed target surface properties. Examples of studies of the BRDF influence on range pulse waveforms are discussed by Steinvall (2000) and the BRDF effects on ladar-based reflection tomography by Jin and Levine (2009). In general, the more glossy the surface is, the more accentuated will those parts be which are normal to the angle of incidence while other surfaces with a slope toward the beam will be reduced in amplitude. Strong glints from a target can mask the following range information if the dynamic range of the receiver is limited. In order to simulate range profiling data a CAD model of the target is desired. 3D CAD models can in many cases be found on the internet to a varying degree of details. These CAD models seldom contain any information about reflectivity and material properties so this has to be assumed. The most straightforward assumption may be to suppose that the surface is diffuse with a 10–20% reflectivity. If the real target is available for measurements this can be scanned using a commercial laser scanner, for example, the Riegl VZ-400 which has the ability to scan a position within an elevation angle of 160 degrees in a full revolution and with a pitch of 0.005 degrees using a 1550 nm laser. Using calibration charts the correct reflectivity in each scan point will automatically be obtained.

Fig. 4 shows a point cloud of an aircraft generated from a scan using the RIEGL scanner and also the model built with the point cloud as a reference. This polygon model including each surface reflectivity will be the base for simulating the laser radar profiling waveforms. Then a sensor model (in our case called the FOI-LadarSim (Chevalier, 2011)) is applied to simulate waveforms from each pixel at the target (pixel density is selectable) and in that way 1D, 2D, and 3D data can be simulated. Beside sensor parameters and their effects, the atmospheric and target speckle are included in the signal generation.

Fig. 5 shows an example of simulated and measured range profiles. The measurements are made against a physical scaled down model of the aircraft JAS Gripen (length about 1.5 m). The measurements were made at short ranges with a 1.5 ns pulsed low power laser mimicking a typical laser range finder with about 10 ns pulse length against a full-sized aircraft (Pace *et al.*, 2008).

Fig. 6 illustrate range profiling of a small boat complemented with a tube and a box to alter the profile. The dashed line is the simulated pulse return. Deconvolution of the second echo with the transmitter pulse reveals two peaks interpreted as indicated in the figure.

There are many uncertainties that could result in differences between the measured and the simulated data. Examples include incorrect geometry in the CAD models and unknown reflectance of the surface material. Roughly, the outer shape of the models may be correct but because the simulations take into account even small features there might be errors. One example is the inside of an aircraft engine or air inlet which might be a source of error. To solve this, further validation with access to the physical models would be required.

Other error sources include incorrect sensor and environmental effects modeling. For example, the exact beam distribution and beam pointing error at the target are hard to predict in detail. There are also some uncertainties in the system parameters (noise, receiver electronics, laser power, etc.).

Examples of Range Profiles Studies

Naval Targets

TNO Defense, Security, and Safety in the Netherlands have shown results on laser profiling mainly from naval ships (Schoemaker and Benoist, 2011; van den Heuvel *et al.*, 2009a,b). One example is illustrated in **Fig. 7**.

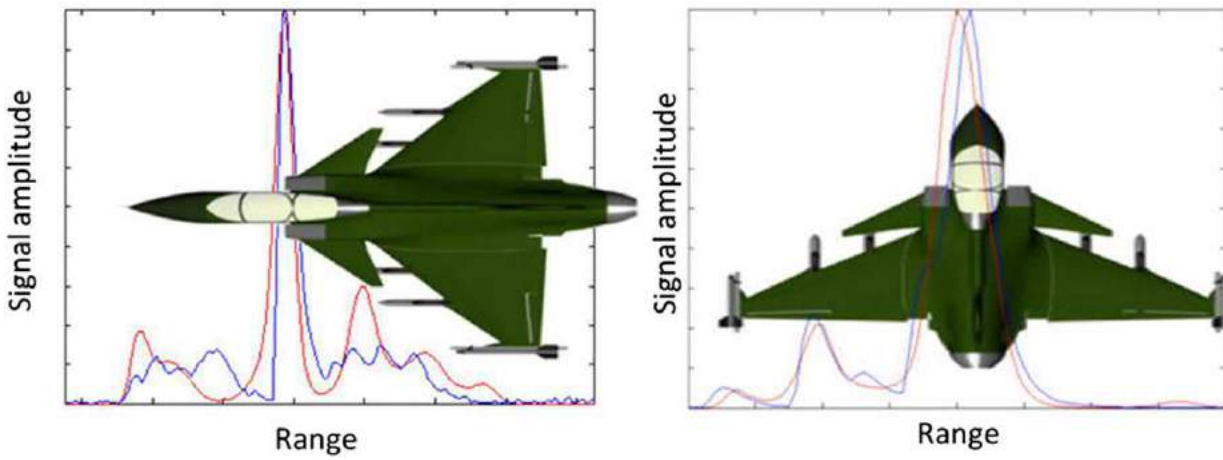


Fig. 5 An example of simulated and measured data. Left: a model of the JAS 39 Gripen seen from the front. Measurement data from a scaled down physical model of the aircraft is given by the red curve, while the simulated of the blue. The superimposed planes' picture shows the structures that correspond to every part of the signal. Right: shows the same comparison seen straight at 90 degrees from the main direction. Reproduced from Pace, P., Steinvall, O., Chevalier, T., 2008. Automatic target classification using one dimensional laser profiling. Technical Memorandum DRDC, Ottawa, TN 2007-000.

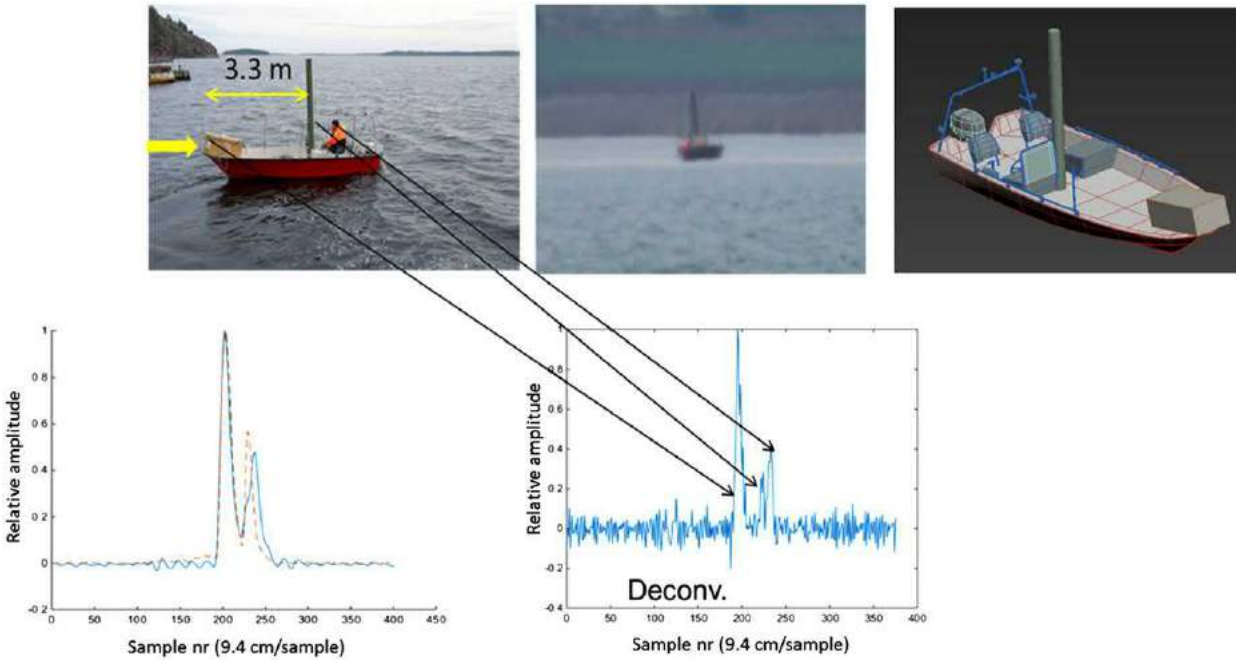


Fig. 6 Upper image shows the small boat complemented with a tube and a box to alter the range profile together with the model for waveform simulation. The first profile (dashed) is the model generated pulse. Deconvolution of the second echo with the transmitter pulse reveals two peaks interpreted as indicated in the figure. Reproduced from Steinvall, O., Tulldahl, M., 2016. Laser range profiling for small target recognition. Optical Engineering 56 (3), 031206.

Fig. 7 shows the measured and simulated range waveforms from a frigate at 10 km range measured in the head on direction. As seen the correspondence is quite good with the exception from two extra peaks present in the measured waveform. These can be associated with the two extra antenna spheres. In the 3D model these structures above the bridge are absent. The system used in these experiments had a transmitter output at 1570 nm around 30 mJ. The receiver used a 750 mm telescope, with an optical aperture of 125 mm diameter. The receiver bandwidth was 200 MHz (van den Heuvel *et al.*, 2009a,b).

A NATO trial in San Diego under the name "NATO Maritime Target ID" was carried out outside San Diego. Several laser and IR systems were used to investigate small surface classification/identification using both 1D, 2D, and 3D laser radars

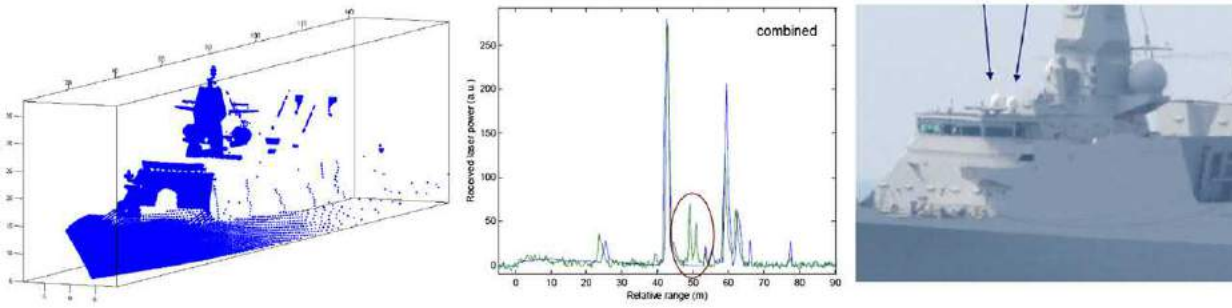


Fig. 7 Left: the point cloud of a ship is used to generate simulated range waveforms. Middle: the simulated (blue) and measured range waveforms from 10 km range. Reproduced from van den Heuvel, J.C., Schoemaker, R.M., Schleijpen, R.M.A., 2009a. Identification of air and sea-surface targets with a laser range profiler. Proceedings of SPIE 7323, 73230Y.

(Schoemaker and Benoist, 2011; Steinvall *et al.*, 2010; Armbruster and Hammer, 2012). The authors in Schoemaker and Benoist (2011) examine the discrimination between three different small boats all between 5–6 m in length. A simple and rather crude model of the boats was used to generate the simulated waveforms to compare with the measured. The result shows that the desired accuracy (95% or better) which was achieved in investigating different simulated waveforms for different noise conditions was not achieved for the experimental data.

In a later experiment in Sweden, FOI demonstrated rather good results with a profiling system which had almost the same characteristics as the TNO system (laser pulse width 6 ns, 10 cm optical aperture, 15 mJ /pulse, and 1570 nm laser wavelength at 40 Hz pulse repetition frequency – PRF). The paper Steinvall *et al.* (2015) shows both experimental and simulated results for LRP of small boats out to 6–7 km. **Fig. 8** shows the small boats used in the test. These are more different from each other than the boats studied in the San Diego trial mentioned above. In the figure the CAD models of the boats are also shown based on laser scanning.

Fig. 9 shows the structure and the corresponding distances as well as an example of a range waveform. Some interpretation is formulated in the figure text. **Fig. 10** shows the echo amplitude variation for the three observed peak during a 10 s data collection at a PRF of 40 Hz. Manual tracking was used to follow the small target at 6 km range which was a demanding task. Also note the high resolution sensors in the SWIR and mid-IR gave hardly any possibility for target classification without a support from the range profile. The SWIR camera had a FOV=0.7 degree $\times$ 0.6 degree fully zoomed. The illuminator laser had a maximum power of 1.6 W and the beam divergence could be varied between 10 and 50 mrad. The camera had a spectral response between 0.4 and 1.7 μm and the detector array had 640 $\times$ 512 pixels. The mid-IR camera had a 1280 $\times$ 720 array of 12 μm pixels responding in the spectral range between 3.7 and 4.95 μm with a typical sensitivity of 17 mK and a minimum narrow FOV of 0.9 degree $\times$ 0.5 degree.

Airborne Targets

Long-range laser profiling measurements of airborne targets are hard to find in the open literature. Measurements on static targets and results from simulation are reported in open sources. Heuvel *et al.* (van den Heuvel *et al.*, 2009a,b) describes an identification algorithm used to distinguish three aircraft from their simulated range profile with good results. The authors use simulated profiles obtained using 3D computer models for three aircraft: an F15, an F16, and an F117. They assume a lidar with 30 cm range resolution.

Heuvel *et al* investigate the relation between identification and aspect angle. They assume that all aircraft have the same diffuse reflectivity. The identification range depends both on range and target aspect angle. An analysis as described in van den Heuvel *et al.* (2009a,b) will yield a (Bayes) identification probability for the various aspect angles. The SNR for a given identification probability can be associated with a range for a given system and target. **Fig. 11** shows the radial plots of the required range (in arbitrary units) against the aspect angle for an identification probability of 90%. According to the authors it is clear from these plots that the required range is shorter for front view (aspect angle of zero degrees). This may be understood if we consider the aerodynamic shape of the aircraft that is responsible for a low laser return. We see a similar effect at 180-degree aspect angle.

It is also informative to compare the radial plots of the three aircraft. The F16 has a smaller maximum identification range than the F15 which has again a smaller range than the F117. Thus, the F117 can be identified at considerably longer ranges than the F16 on average. This is most likely due to the characteristic shape of the F117 (van den Heuvel *et al.*, 2009a,b).

Example of full scale laser profiling measurements on a real aircraft (SK-60 veteran aircraft) is shown in **Fig. 12**. Note that the range profiles change shape for rather small changes in azimuth. The measurements were made at 1 km range using a 1570 nm laser.

Fig. 13 shows a profile map from a horizontal scan for two different aircraft. The signature is quite different as can be seen. In **Fig. 14** the measured profile covering the whole, aircraft mimicking a long-range profiling is compared using a simple 3D cad

Fig. 8 Test vehicles for the trial in Sweden. The upper boats were slightly modified by placing a stove pipe and a box to change the signature. Boat A (upper): length 4.73 m, width 1.92 m. Boat B (middle): length 6.18 m, width 2.15 m (total length incl. the rear flat plate = 6.57 m). Boat C (lower): length 8.64 m, width 3.15 m (total length incl. the rear flat plate = 9.55 m). The last row shows the 3D models based on laser scanning using the Riegl scanner. Reproduced from Steinvall, O., *et al.*, 2015. Passive and active EO sensing of small surface vessels. Proceedings of SPIE 9649, 96490I.

model from the internet and assuming a complete diffuse surface for the whole aircraft. Many main features compare well but there are also some difference due to the incomplete aircraft model.

Small targets of interest for detection and classification include UAVs and missiles. Laser profiling of these targets may be done but need a much higher range resolution probably 10 cm or less. TCSPC offers a range resolution in the sub-centimeter level and we have demonstrated this against a UAV mockup according to **Fig. 15** (Steinvall and Tulldahl, 2016).

Fig. 16 shows the UAV at the 1.3 km distance and the scanning the pattern from the single detector device color-coded in reflected laser intensity. The size of the pixels was about 15 cm in square. Since we had a weak laser we integrated the counting during 5 s/pixel. The laser PRF was 4 MHz and with a pulse duration of 25 ± 5 ps. The laser's average power was 20 mW. The laser divergence was 280 μrad and the detector FOV = 150 μrad (which gave some "spillover"). **Fig. 17** shows the sum of the range waveforms over the entire scan area (corresponding to a fictitious beam divergence of about 1.7 mrad). We can clearly see echoes of the various structures including that from the cone-like front (provide a linear amplitude change with time). The distances derived from the pulse response are consistent with the actual distances within 1 cm.

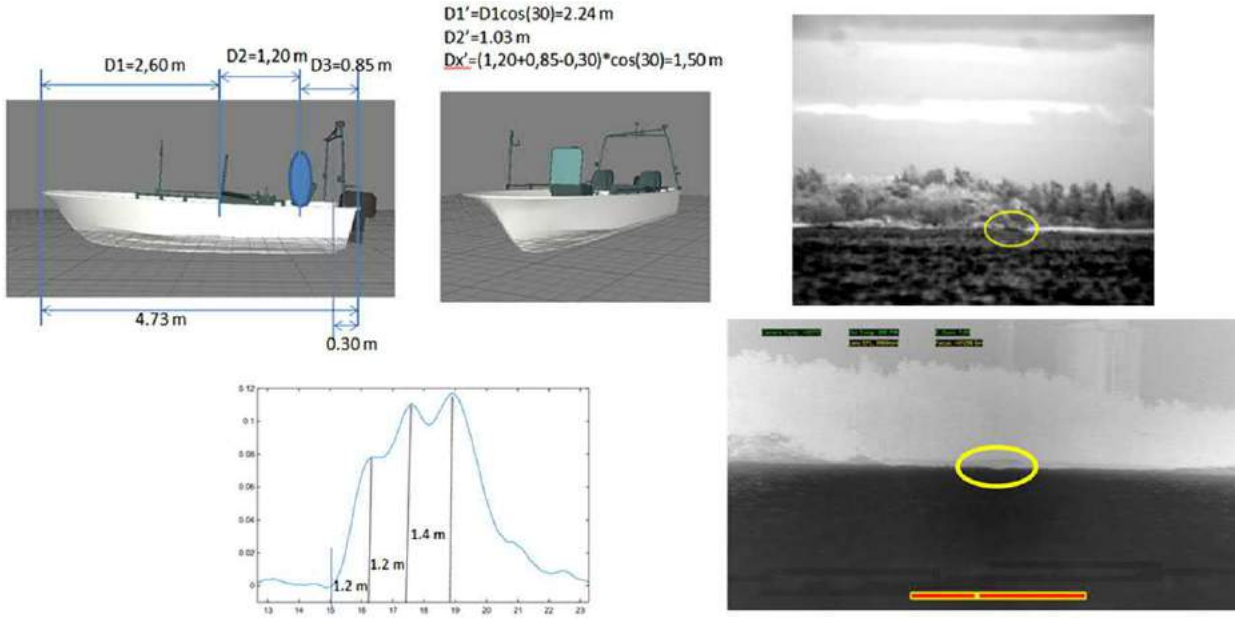


Fig. 9 The dimensions and structure of a the small boat. The 3D model was obtained from the laser scanning with a Riegl scanner. Looking at the waveform (an example given in the figure) we estimate the (1.2 + 1.2) m = 2.4 m as representative to D1′ (= 2.24 m) seen at a 30-degree azimuth angle. The peak separation (1.2 + 1.4)m = 2.6 m is also observed and may be representative to the pulse form from the keel at 30 degrees and the driver. The peak separation of 1.2 m could correspond to the peak from the shield in front of the driver and the driver. Right shows two images from high resolution passive sensors (SWIR at the top and mid-IR at the bottom). Reproduced from Steinvall, O., *et al.*, 2015. Passive and active EO sensing of small surface vessels. Proceedings of SPIE 9649, 96490I.

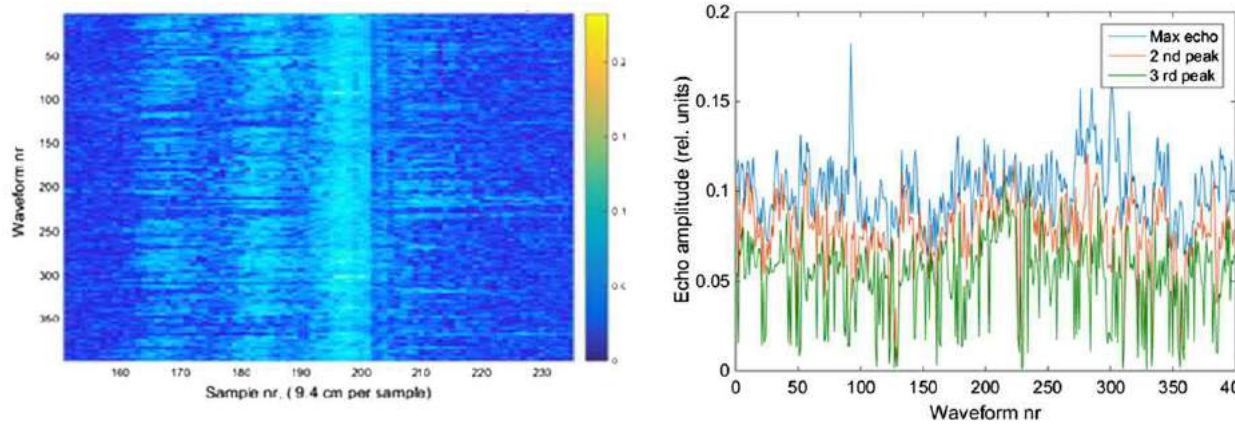


Fig. 10 Left: the intensity (*x*-axis) of 400 waveforms during a 10 s tracking period for the small boat illustrated in Fig. 9. Right: echo amplitudes for the maximum, 2nd and 3rd peak. One can see correlation between the amplitudes and amplitude fluctuations due to beam pointing variations. Reproduced from Steinvall, O., *et al.*, 2015. Passive and active EO sensing of small surface vessels. Proceedings of SPIE 9649, 96490I.

In these examples, we used long integration times to collect the range waveforms. If a stronger laser is used we can anticipate shorter integration times and still achieve maximum ranges out to 10 km (Steinvall *et al.*, 2012b). But even for integration times in the millisecond range a target may have moved over a distance exceeding the range resolution and thus smearing out the information in the range waveforms. Recently, we have developed a method (Jonsson *et al.*, 2015) that can correct for this. The method is based on the autocorrelation of the waveform so that in retrospect one can reconstruct the waveform that would correspond to a stationary target. One assumption is that the target is relatively stable in angular motion during the integration time so that the instantaneous waveform retains its shape.

Fig. 18 shows the results from simulation of the UAV azimuth angles between 0 and 35 degrees. The figure illustrates the maximum amplitude from the wing echo for target azimuth = 0 degrees which, at about 5 degrees, splits into two echoes and separates further with increasing azimuth angle. At about 25–30-degree azimuth angle, the second wing echo (later in ToF) almost

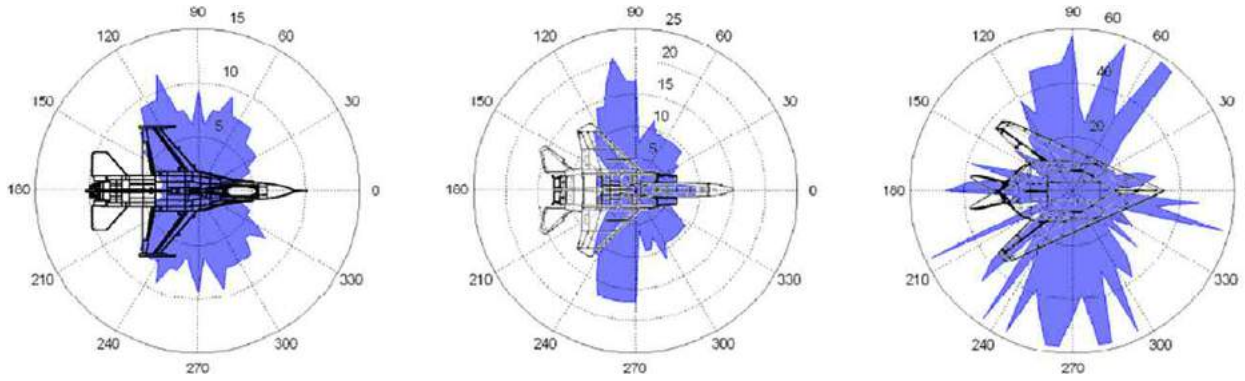


Fig. 11 Radial plot corresponding to an identification probability of 90% for an F16, F15, and F117, left to right respectively. After van den Heuvel, J.C., Schoemaker, R.M., and Schleijpen, R.M.A., 2009a. Identification of air and sea-surface targets with a laser range profiler. Proceedings of SPIE 7323, 73230Y.

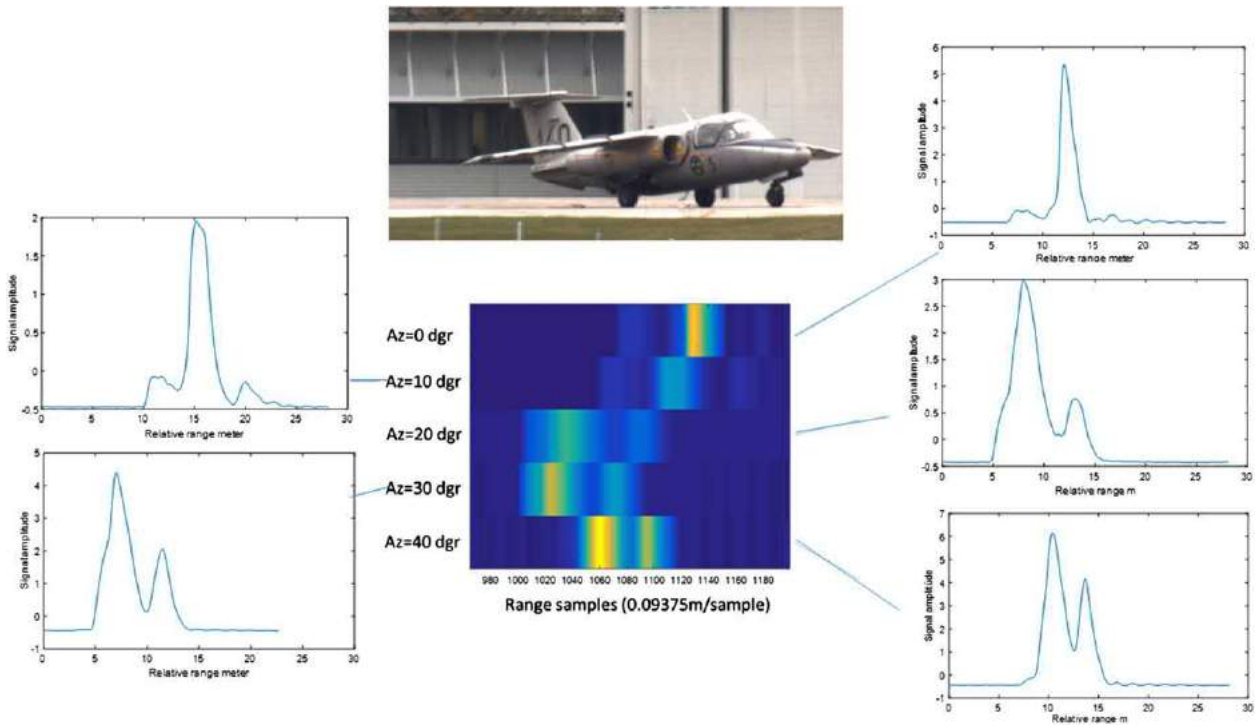


Fig. 12 Range profiles for the different azimuth positions of the SK-60 aircraft. The range profiles are summed over all measurement points on the aircraft for a specific azimuth (=0 for head on) and this mimicking a long-range profile when the laser lobe is covering the whole aircraft. Image FOI.

completely disappears, shadowed by the UAV body, and the first wing echo is spread out over time and combined with reflections from the UAV body.

Land Targets and Mapping

Results on LRP of land targets are limited concerning horizontal paths. On the other hand, vertical laser scanning from an airborne platform is well established especially for terrain and underwater sensing.

The civilian market has been leading this field for the last 20 years. Both space and airborne laser radar systems, as well as terrestrial laser radars, have been developed by several different vendors. The large number of publications and applications, as well as the rapid development of hardware, shows the large interest in this technology. A book giving a good overview of topographic laser scanning techniques was published by Shan and Toth (2009). **Fig. 19** illustrate some developments of airborne laser scanning system for both terrain and depth sounding applications.

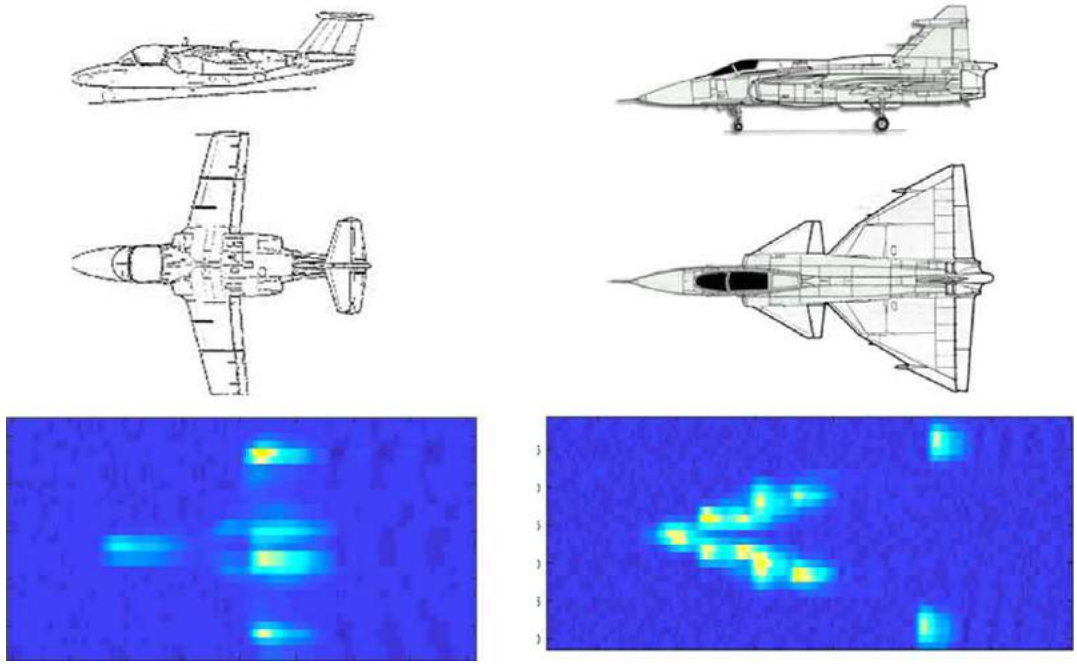


Fig. 13 Laser scan using a beam with 1 mrad beam divergence with the aircraft at 1 km range. Two veteran aircraft SK-60 (left) and Viggen (right) from the Swedish Air Force. Images FOI.

Laser scanners are used on UGVs for navigation and obstacle avoidance. Systems from Teledyne Optech and Velodyne Lidar are the examples of commercial road mapping laser radars with many potential military applications (e.g., generation of synthetic environments, detection of improvised explosive devices, and so on). The accuracy is said to be on the order of centimeters at a maximum range of 100–250 m. Measurement rates are found between 1 and 2 million points per second. This allows a 6-cm point spacing at 10-m range for a vehicle velocity of 43 km/h.

Vehicle borne laser radar is becoming a very important sensor for self-driving cars. This may be the first widespread commercial application of laser radar technology. It will force the technology to be small and compact at a cost of a few hundred dollars per system.

Scanning laser radars have been intensely studied in the United States for missile applications (Gustavson and Davis, 1992). A civilian counterpart is traffic monitoring classifying different vehicles (SICK-Sensor Intelligence, 2017). Target detection and homing were also studied for air to air seekers, and in space for the SDI program. The largest investments in laser radar seeker technology have been made for air to ground seekers. The Low Cost Autonomous Attack System (LOCAAS) was a demonstrator program (Ladar Seeker/ATA Algoritm Captive Flight Test Results, 2017) driven by the Air Force (Eglin Florida). The AFRL's LOCAAS program terminated in mid-2006 without going to production. The seeker was based on scanning laser radar generating 3D imagery of the targets with ATR and aim point selection built into the system. Loitering Attack Missile (Signal, 2017) was a project relying on the laser radar technology developed for LOCAAS. The paper from Andressen *et al.* (2005) gives some insight to this technology.

Laser scanning systems with a profiling capability can also be used for both target search and classification. If the return has a specific signature this could be indicative of a target that differs from the surrounding terrain. In principle, this could include the profiles from land vehicles which probably differs a lot from those from the terrain background. One example of a specific strong and deviating signature is the glint return from optics which can be detected from a scanning system placed on a combat vehicle looking for optical threats (Sjöqvist *et al.*, 2016). One of the problems in these types of systems is to discriminate between real targets like optical sights from other strong reflectors like traffic signs, etc. One technique to overcome this might be to use high resolution range (HRR) profiling based on photon counting (Sjöqvist *et al.*, 2013). See **Figs. 20** and **21**.

Atmospheric Profiling

One property for a profiling laser may also be to probe the atmospheric transmission, for example, along slant paths. Probing of the density of aerosols and thus the attenuation can be done using the backscatter signal from the atmosphere. Cloud mapping including their height distribution and density (up to a certain level) are also relevant for this type of operation. As an example it is interesting to probe the atmosphere in the dark (e.g., from a ship) when visual references disappear. Atmospheric laser radar or lidar is well investigated in the research community (Molebny *et al.* 2016; Fujii and Fukuchi, 2005; Steinvall *et al.*, 2015b).

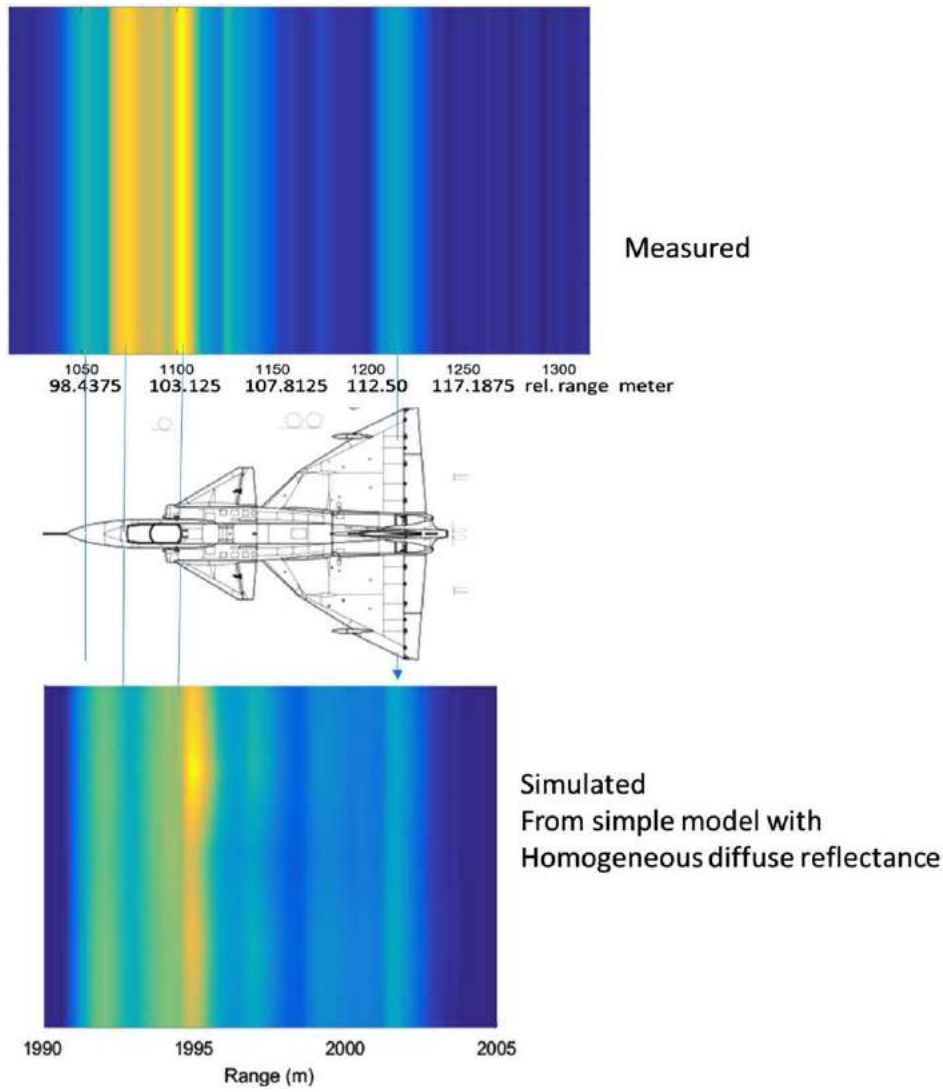


Fig. 14 Comparison between the measured and simulated range profiles in the head on direction for the veteran Viggen aircraft. Images FOI.

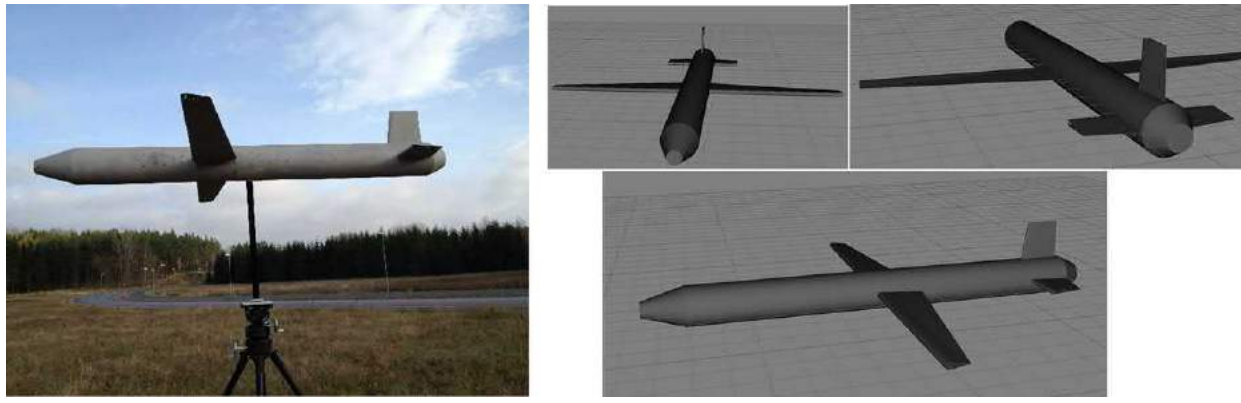


Fig. 15 Left: unmanned aerial vehicle (UAV) mockup made from sandblasted aluminum. Length 3450 mm, diameter 300 mm wing span 3000 mm. The distance from nose to the large wings is 1300 mm. Right: a CAD model for laser radar modeling. Reproduced from Steinvall, O., Tulldahl, M. 2016. Laser range profiling for small target recognition. Optical Engineering 56 (3), 031206.

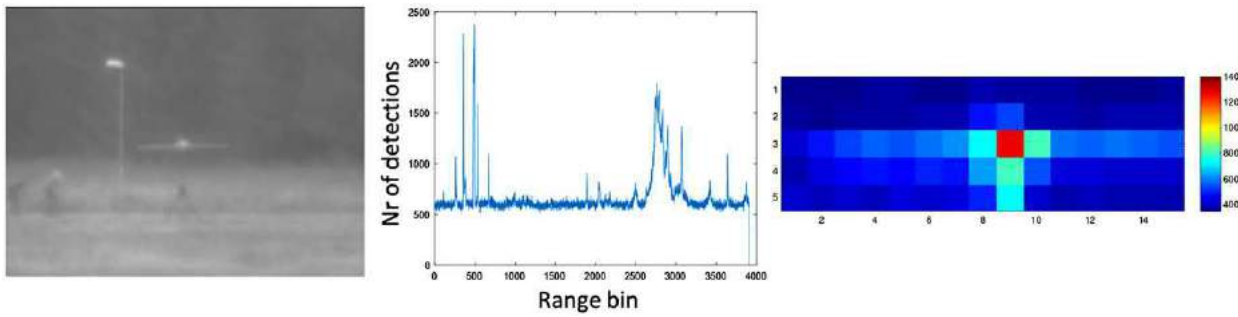


Fig. 16 Measurement with the photon-counting system on the UAV mock-up at 1.3 km distance. In the middle a waveform contains the target and terrain background. Right shows the resulting intensity of each pixel (approximately 15 cm square) in the photon-counting system when it was scanned over the target. Reproduced from Steinvall, O., Tulldahl, M. 2016. Laser range profiling for small target recognition. Optical Engineering 56 (3), 031206.

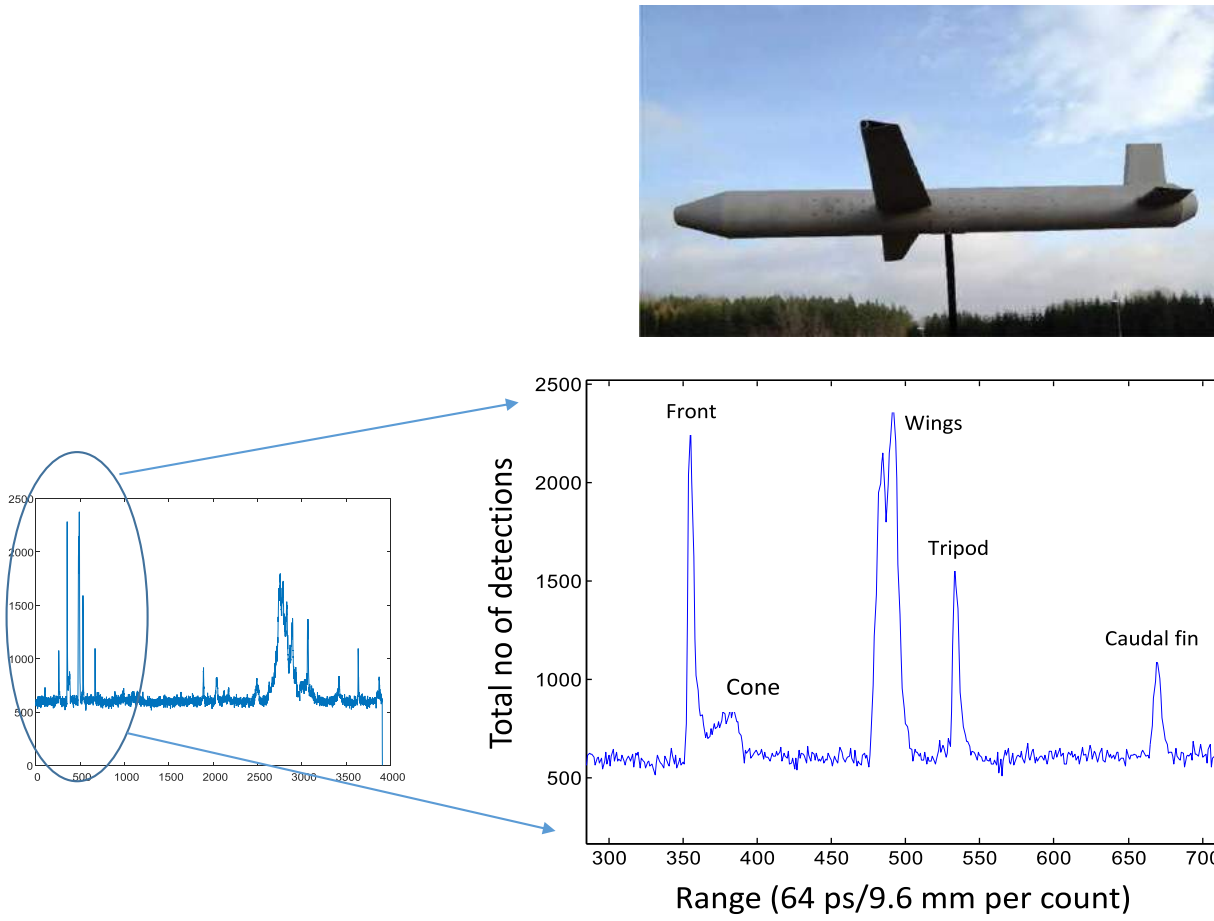


Fig. 17 Summed range waveforms from all 3 × 15 pixels above the target (UAV). The rear pulses belong to the terrain while the first part is enlarged and shows the range structure of the UAV with very high precision. The wings provide two close echoes because the UAV was not exactly perpendicular to the beam. Reproduced from Steinvall, O., Tulldahl, M. 2016. Laser range profiling for small target recognition. Optical Engineering 56 (3), 031206.

Examples of Tomography Based on Laser Profiling

The aim of reflective tomography is to estimate object surface features based on a set of reflective projections that are measured in angular increments around the object either by rotating the sensor or the object. The basics of laser based reflective tomography are well described by Knight *et al.* (1989b).

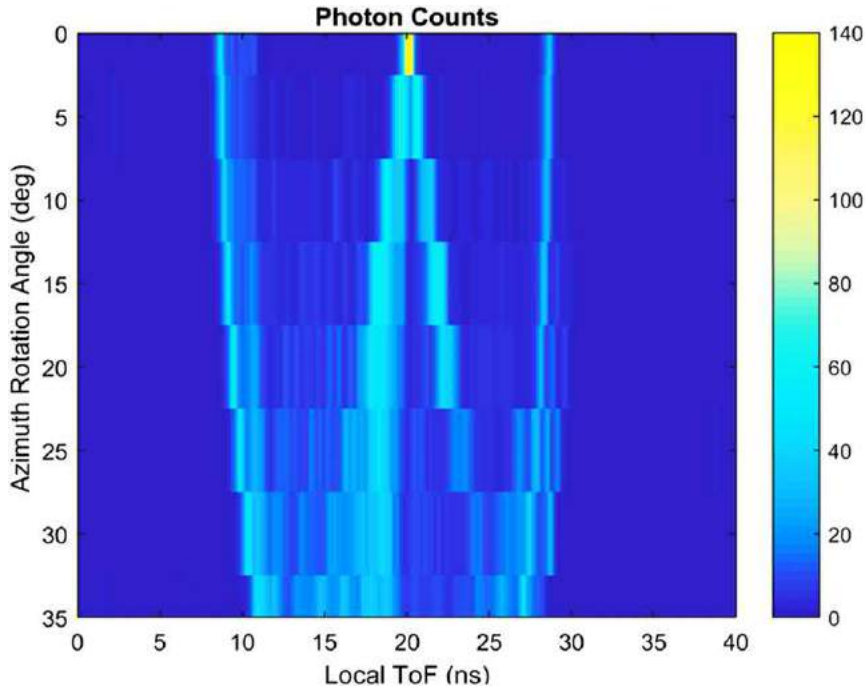


Fig. 18 Illustration of simulated waveforms over the target. The number of detected photons are shown in color scale, and the simulated target aspect angles are presented for the interval from 0 to 35 degrees in azimuth (in steps of 5 degrees). Reproduced from Steinvall, O., Tulldahl, M. 2016. Laser range profiling for small target recognition. Optical Engineering 56 (3), 031206.

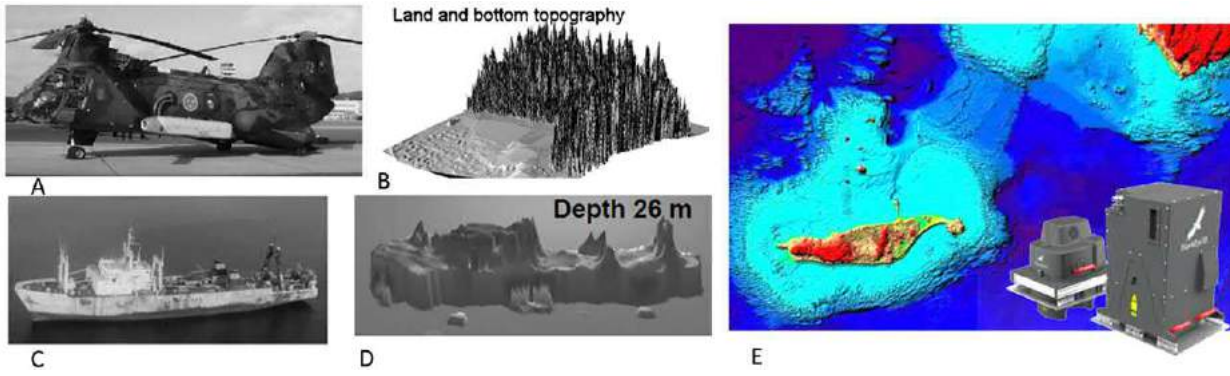


Fig. 19 (A) The original Hawk Eye – a system which developed further can combine topographic mapping on land and water depth sounding (now made by AHAB Leica Geosystems). The figure (B) shows one example of real data. Below (C and D): a ship before and right after it sunk at 26 m depth. Images AHAB, FOI and the Swedish Maritime Administration. (E) Example of a Hawk Eye III, an airborne multi-sensor deep water bathymetric and topographic LiDAR from Leica Geosystems. The green to blue color is indicative of the water depth.

Laser tomography using high range resolution profiling based on TCSPC enables small models mimicking real targets to be studied in detail. Examples are given in Steinvall *et al.* (2014), Henriksson *et al.* (2012a,b), Sjöqvist *et al.* (2014), and Steinvall *et al.* (2012b). The principle is depicted in **Fig. 21**.

Based on the scheme for filtered backprojection (FBP) using the inverse Radon transform, the tomographic image can be reconstructed. Mathematically, the 2D object can be expressed as (Henriksson *et al.*, 2012a,b):

$$g(x,y) = \sum_{i=1}^{m} q(z\cos\theta_i + x\sin\theta_i)\Delta\theta \tag{3}$$

where

$$q(r,\theta) = \mathcal{F}^{-1}(f\mathcal{F}(p(r,\theta)) \tag{4}$$

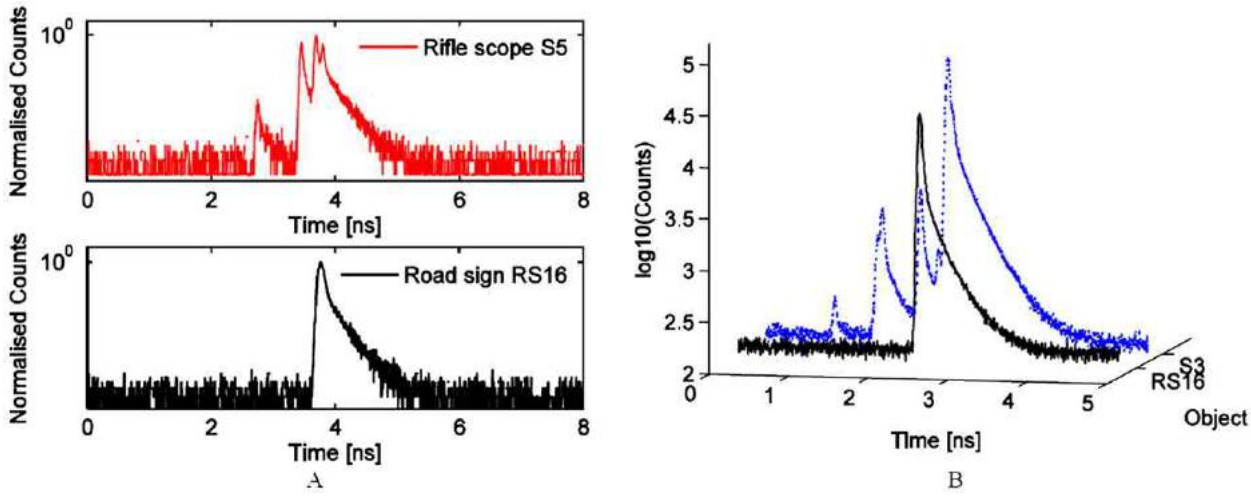


Fig. 20 (A) Example showing range-profile signal from a rifle scope S5 (upper) and a road sign RS19 (lower) at 750 m distance and 100 ms integration time. (B) Comparison between lineshape of a rifle scope (S3) and a road sign (RS16). Reproduced from Sjöqvist, L., Allard, L., Henriksson, M., Jonsson, P., Pettersson, M., 2013. Target discrimination strategies in optics detection. Proceedings of SPIE 8898, 88980K.

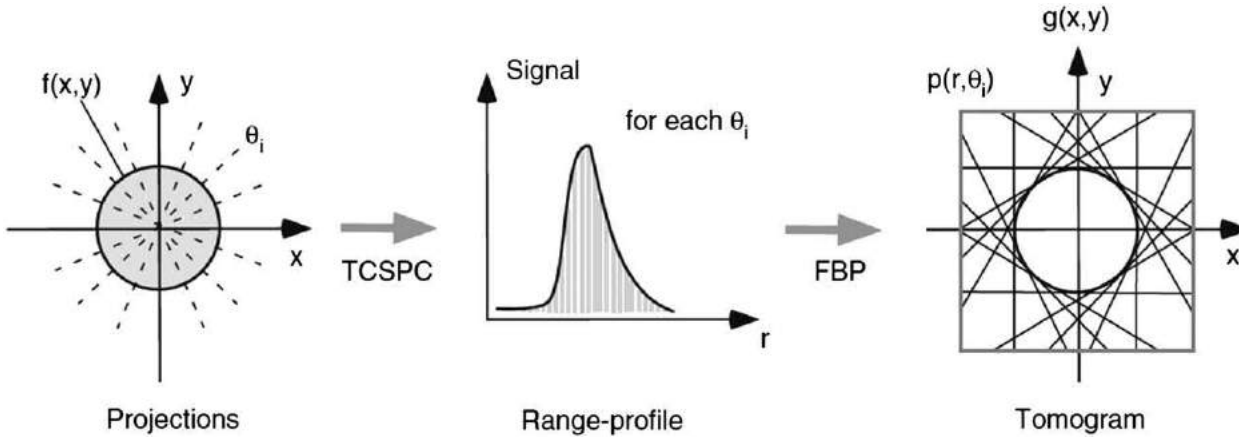


Fig. 21 Schematic showing the principle of reflectance TCSPC tomography with projections and range profiles for each angle of incidence θ_i, and the corresponding tomogram after the filtered backprojection (FBP) transform. Reproduced from Sjöqvist, L., Henriksson, M., Jonsson, P., Steinvall, O., 2014. Time-correlated single-photon counting range profiling and reflectance tomographic imaging. Advanced Optical Technologies 3, 187–197.

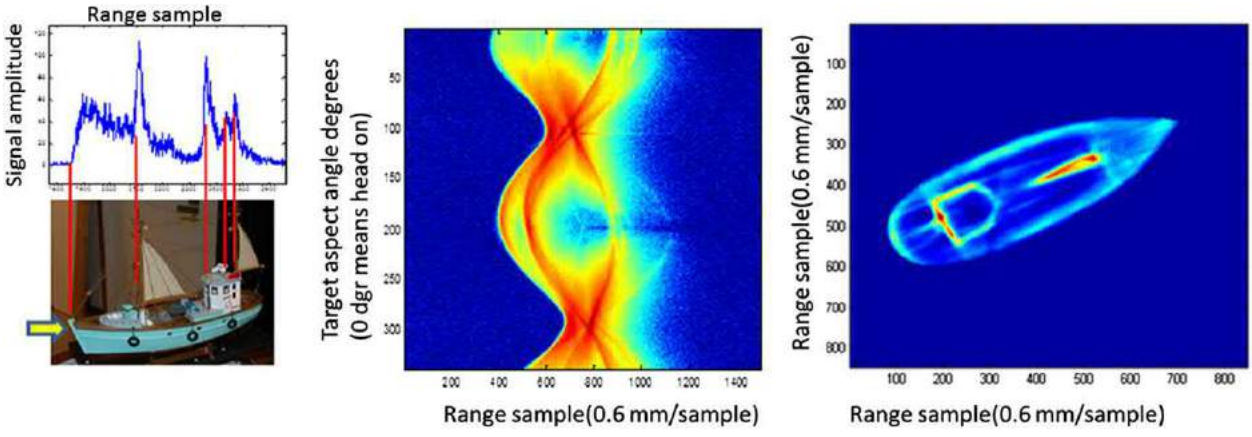


Fig. 22 Example of tomography of a small model boat, 43 cm in length. Our laboratory TCSPC ladar system had a time response of 54 ps corresponding to a range resolution of 0.8 cm. Reproduced from Steinvall, O., Chevalier, T., Grönwall, C., 2014. Simulation and modeling of laser range profiling and imaging of small surface vessels. Optical Engineering 53 (2), 029801.

and $\mathcal{F}$, $\mathcal{F}^{-1}$ denote the forward and inverse Fourier transform, respectively, r is the distance along the measurement direction with zero at the rotation center of the studied object (not a polar coordinate), $p(r, \theta_i)$ is the range profile for each rotation angle θ_i, and f is a filter function. Here it was assumed that the rotation axis is parallel to the y-axis. The filter function, f, was chosen to be a generalized ramp function according to:

$$f(\omega) = |\omega| \cdot e^{-\xi|\omega|^a} \qquad 0 \le |\omega| \le \pi \tag{5}$$

and $\xi = |1/\omega_c|^a$ with ω_c defining the cut-off frequency and a is an adjustable parameter. Typical values used in our work with a photon counting systems (Sjöqvist *et al.*, 2014) to reduce the high-frequency components were $a = 3.4$ and $\omega_c = \pi/4$. The values need to be optimized based on the IRF of the measurement system.

FOI laboratory TCSPC ladar system had a time response of 54 ps corresponding to a range resolution of 0.8 cm. Realistic range resolution from conventional high resolution direct detection range finders of laser radars is in the region 0.2–1 m, which correspond to scale factors between 25 and 125. The extreme range resolution of TCSPC ladar makes it possible to perform experiments using laboratory scale models relevant for scaling to data expected with linear-mode laser radar systems for the identification of large-scale objects, allowing testing of possible system performance in a simple and inexpensive way (Steinvall *et al.*, 2014).

Fig. 22 shows examples from the measurements using the 43-cm long fishing boat as a target. Left is an example of the observed waveform looking straight at the fore of the boat, middle, and range profile for 341 degree in steps of 1 degree. Right is a tomographic image obtained by using the Radon transform to reconstruct the shape of the boat. The measurement range in the laboratory was about 40 m and the beam covered the whole FOV of the receiver, which was 10.2 mrad ($1/e^2$) corresponding to 40 cm at the target.

There are many factors affecting the quality of the tomographic image, for example, strong reflections which cause linear features stretching through the reconstructed image. Jin *et al.* has investigated the BRDF effects in ladar-based reflection tomography (Jin and Levine, 2009). The influence from these artefacts can be partly removed by using the "convex hull" method where the first response (opaque surface) at each angle from the object is used to create an image mask outside which the response is known to be zero (Sjöqvist *et al.*, 2014). Critical issues regarding the quality of the reconstructed tomogram beside SNR include defining the center of rotation, applied angular resolution, and the used angular projection sector. The way of presenting the tomogram also has a large influence on the image quality. Often the log intensity including a threshold give a good quality (see **Fig. 23**) paired with a type of filtering like the "convex hull" mentioned above.

So far only laboratory measurements examples have been presented. There are, however, also several long-range demonstrations reported in the literature. The use of 1D Doppler-resolved projections to form a 2D tomographic image of a rotating object was demonstrated using the MIT Lincoln Labs 10.6 μm Firepond laser radar. With a model of a Thor-Delta rocket target at 5.4 km ground range, Doppler-time-intensity was demonstrated (Knight *et al.*, 1989b). Images of satellites at ranges up to 1500 km were also collected by the Firepond (Gschwendtner and Keicher, 2000) (**Fig. 24**).

Matson and Mosley (2001) report on the first satellite feature reconstruction, by use of range-resolved reflective tomography techniques collected on an orbiting satellite. The reconstructed features were two retroreflectors mounted on a satellite. The data were collected with a coherent laser radar system located at the Maui Space Surveillance Site in Maui, Hawaii. The laser system called HI-CLASS was a pulse coherent laser radar emitting 30 J/pulse at 30 Hz repetition rate. The acquisition waveform was a gain-switched pulse with a duration of the order of 10 ms, whereas the imaging waveform is a mode-locked version of the acquisition waveform. The mode-locked waveform was a series of micropulses modulated by the gain-switched pulse waveform in which the micropulses are less than 1.5 ns in width and are separated from each other by 40 ns. **Fig. 25** shows some results. A key step in the projection creation process was to align the projections to the center of rotation of the satellite.

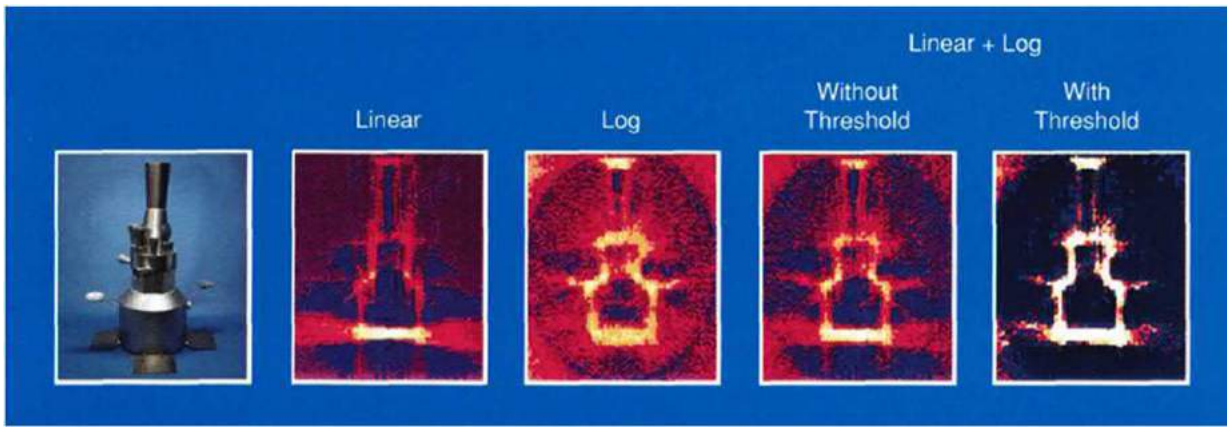


Fig. 23 Photograph and reconstructions of a model of a DSP satellite. Note the better image quality using log intensity and thresholding. Reproduced from Knight, F.K., Kulkarni, S.R. Marino, R.M., Parker, J.K., 1989. Tomographic techniques applied to laser radar reflective measurements. Lincoln Laboratory Journal 2 (2), 143–158.

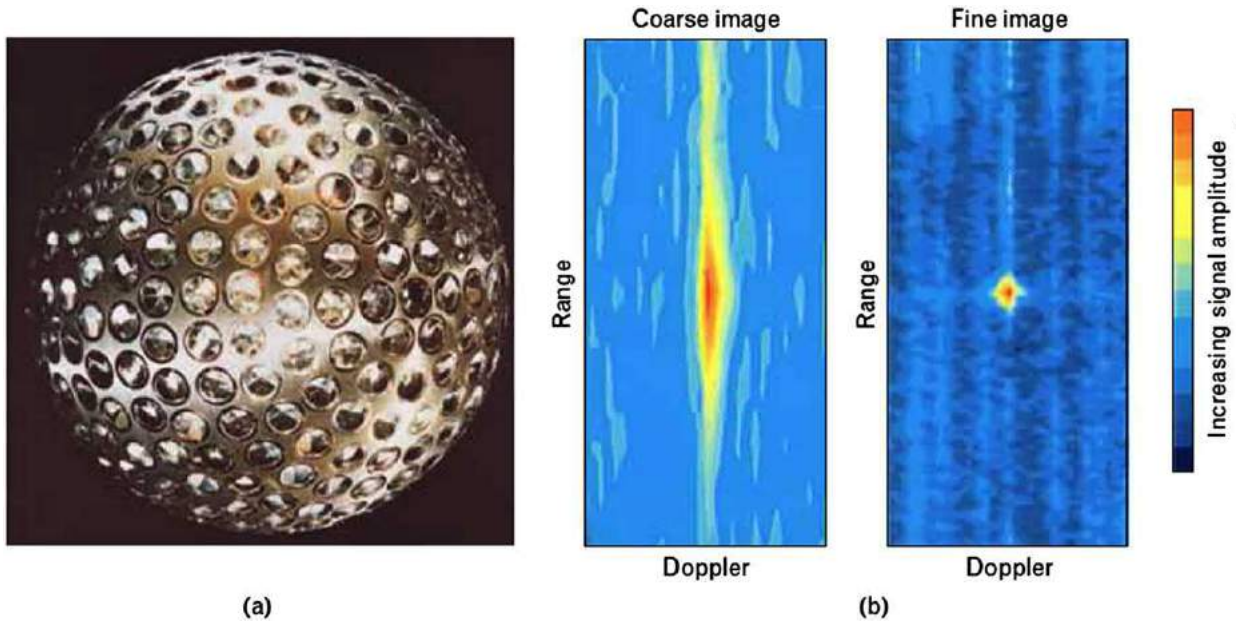


Fig. 24 (a) The Laser Geodynamics Satellite (LAGEOS), a 60-cm aluminum sphere with a brass core. The satellite has a total mass of 406 kg and is in a near circular orbit at an altitude of approximately 5900 km. LAGEOS has 426 silica retroreflectors and 4 germanium retroreflectors to serve as a laser-radar target. (b) Range-Doppler image of the LAGEOS satellite collected by the wideband CO_2 laser radar at Firepond. The coarse image was made with a signal bandwidth of 150 MHz, while the fine image was made with a bandwidth of 1 GHz. Doppler velocity resolution is approximately 30 cm/s. Color in the image represents relative signal amplitude. The range and Doppler sidelobes are better than 20 dB below the main lobe. The range-Doppler images are the point-target responses of the imaging laser radar produced by different bandwidths. Figure text and figure from Gschwendtner, A.B., Keicher, W.E., 2000. Development of coherent laser radar at Lincoln Laboratory. Lincoln Laboratory Journal 12 (2), 383–396.

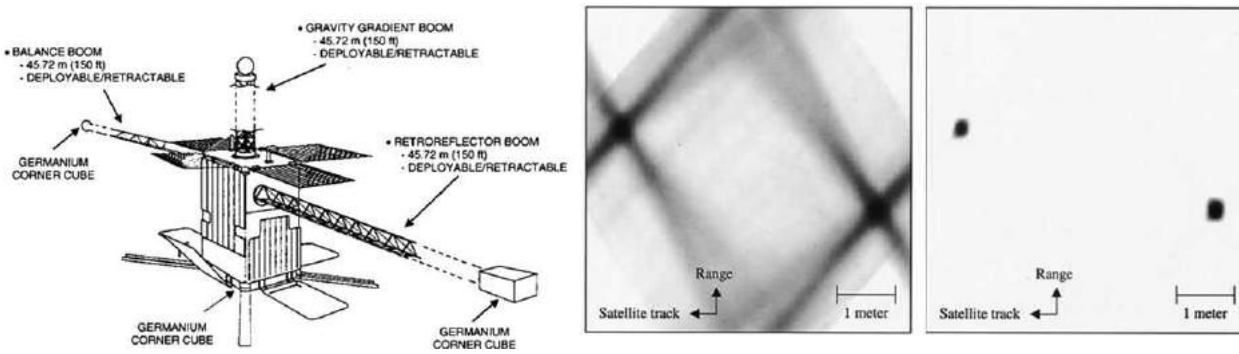


Fig. 25 Left: representation of the LACE satellite. Middle: tomographic reconstruction of the two retroreflectors at the satellites arms. Right: thresholded image to remove limited-view artifacts. Reproduced from Matson, C.L., Mosley, D.E., 2001. Reflective tomography reconstruction of satellite features – field results. Applied Optics 40 (14), 2290–2296.

Another long-range example is given by Murray *et al.* (2010). They use pulse coded ladar waveforms that have desirable side-lobe-free autocorrelation properties. In opposite to short high pulse energy techniques this will enable high bandwidth (>GHz), low-peak power, compact and efficient fiber laser transmitters and photonic receivers to be used in compact systems. See **Fig. 26** for an example of the result.

A theoretical analysis of including turbulence effects for lidar reflective tomography imaging for space objects is presented by Qu *et al.* (2011). The development of laser range profile, Doppler spectra, and range-resolved Doppler imaging technologies are reviewed in the paper (Wu *et al.*, 2014).

Reflective tomography can also include making 3D images from a series of 2D images. Different techniques have been published (Xinwei *et al.*, 2016; Andersson, 2006; Laurenzis and Woiselle, 2014) for this reconstruction but they all rely on evaluating the range from the pixel intensity increase or decrease with time. The convolution of the camera gate function and the laser pulse shape form this time dependence to be used in the 3D reconstruction. The 3D quality depends on the single frame SNR and the number of frames used for the reconstruction. In **Fig. 27** we illustrated how thermal and laser imaging can be combined for target detection and classification using a 2D to 3D tomographic technique.

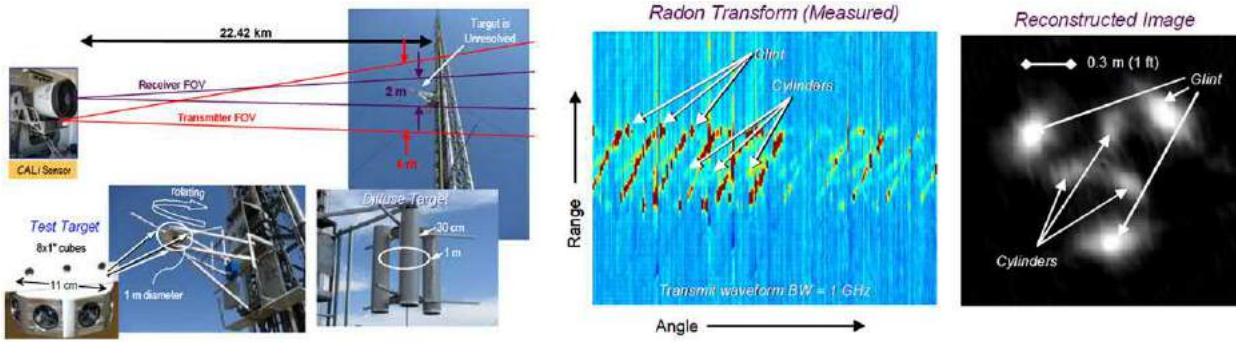


Fig. 26 Left: long-range field test configuration. An unresolved rotating 1 m diameter target consisting of both retro-reflective and diffuse scattering structures is placed on a tower 22 km from the laser platform. The ladar field of view (FOV) is 2 m and the transmitter FOV is 4 m at range. Right: first tomographic image reconstruction of an unresolved target located 22 km from the ladar sensor. Both the retro-reflective (glint) and diffuse cylindrical targets appear in the image. Reproduced from Murray, J., Triscari, J., Fetzer, G., *et al.*, 2010. Tomographic lidar. In: Applications of Lasers for Sensing and Free Space Communications, San Diego, CA, OSA Conference Paper.

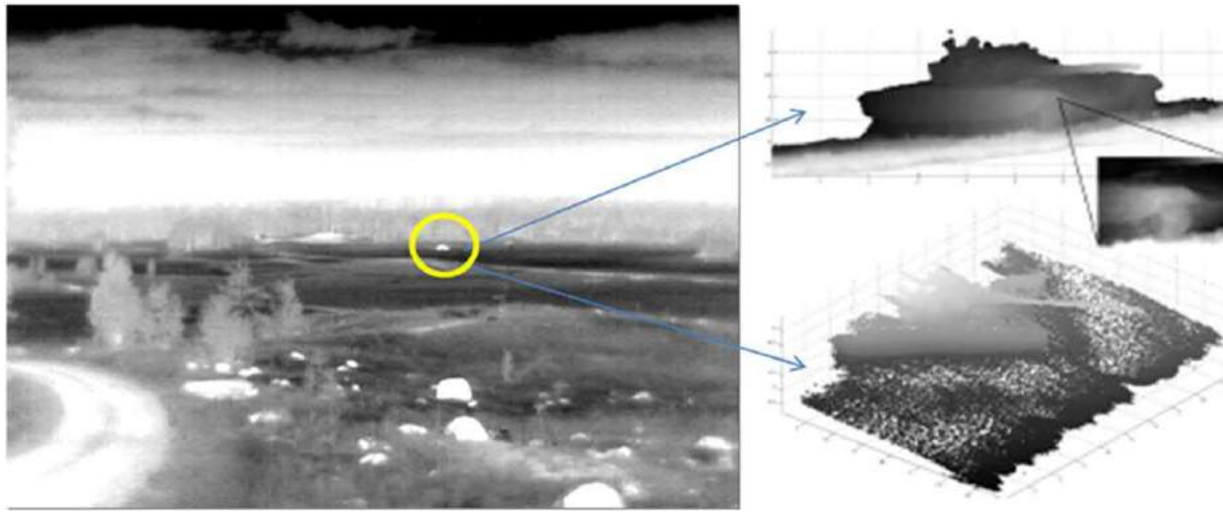


Fig. 27 Illustration of target detection with a thermal imaging camera and the subsequent target classification based on 3D target generation from a series of sliding gates over the target. Reproduced from Andersson, P., 2006. Long range 3D imaging using range gated laser radar images. Optical Engineering 45 (3), 034301–034310.

Signal Processing Techniques for Range Profiling and Tomograhy

In the microwave radar field signal processing techniques for target classification have been studied more extensively than for laser radar. This should motivate a closer look at the radar signal processing methods to see if they are applicable to laser radar while noting that there are difference in the type of signals they generate.

In the radar field techniques based on sequential classification tracking as well as model and feature-based methods have been developed. The sequential technique is based on successive target measurements performed over long time intervals, typically several seconds or minutes. During this period target maneuvering will be indicative of the type of target. This information from radar (or an optical sensor) will limit the search target search space and allow for other model or feature-based methods to be added for a more robust recognition.

Model-based methods employ physical and/or empirical target models to predict the measured target signals. The measured and predicted (model) signals are compared and the model which gives the smallest difference to the measured signal is chosen.

Feature-based methods do not use a model of the measured signals but rely on extracting a limited number of target features from the measurements. Subsequently, the extracted feature vectors are compared with feature vectors in a training set extracted from measurements.

It seems that model-based methods are most applicable to LRP because real measurements may be hard to achieve at least for military targets. On the other hand, real measured profiles may be collected over time using targets of opportunity.

HRR profiling for microwave radar has been investigated rather extensively (see e.g., Shaw *et al.* (2013), Li and Yang (1993) and references therein). One of the most critical issues about the range profile signature in the microwave domain is its extreme variability as a function of aspect angle. As little as a 0.1-degree aspect angle change can cause such a severe change that the signature is no longer recognizable. This is understood as the radar signal is composed of a number of strong scatter center contribution and a small change of the scattering center positions to a quarter of a wavelength will completely change the phase of the return. This fact motivated that radar signal processing of HRR often contains statistical modeling to capture these variations.

The most direct processing technique for range profiling may be correlation. In this method, the profile is convolved with a reference profile from the model library. It is convenient to normalize the waveforms to the highest peak for example. The method has been applied to real radar HRR data from 24 aircraft with (Hudson and Psaltis, 1993) success provided aspect information is

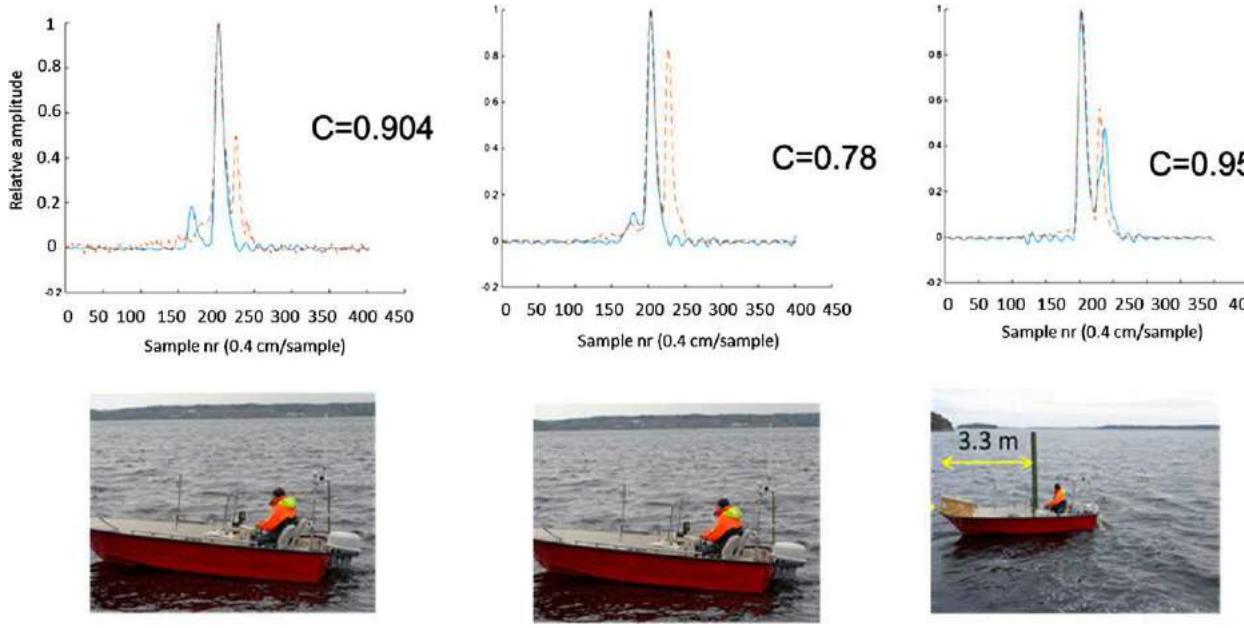


Fig. 28 Example of measured (full line) and simulated (dotted) waveforms together with the correlation coefficients obtained after matching the largest peak. One can observe how the waveforms both match and mismatch probably due to different angular boat positions and beam jitter. Reproduced from Steinvall, O., *et.al.*, 2015. Passive and active EO sensing of small surface vessels. Proceedings of SPIE 9649, 96490I.

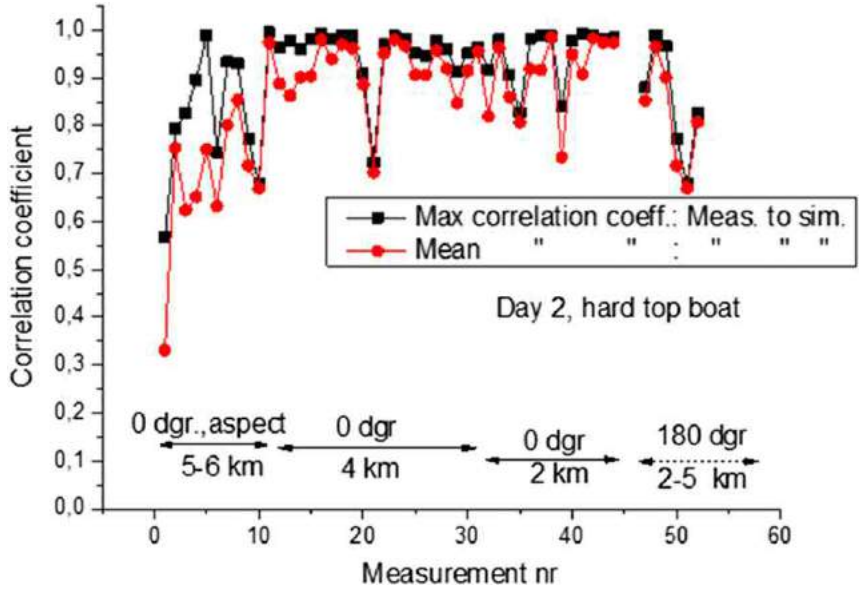


Fig. 29 Correlation coefficient between measured and simulated waveforms from boat B at 0- and 180-degree aspect. Different target ranges are indicated. Reproduced from Steinvall, O., Tulldahl, M., 2016. Laser range profiling for small target recognition. Optical Engineering 56 (3), 031206.

used, and provided identifications are based on sufficient number of profiles so as to decrease errors due to speckle fluctuations. Li and Yang (1993) used correlation to investigate target classification using experimental data from physical aircraft model with sizes in the 50–80 cm range measured in the frequency range 6–16 GHz.

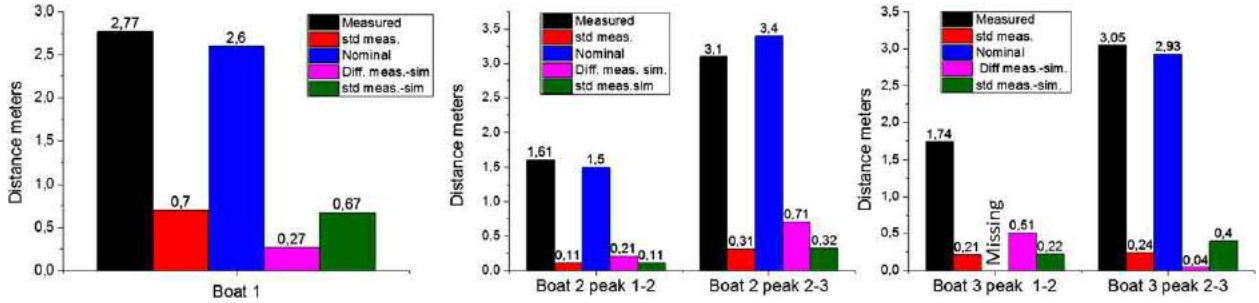


Fig. 30 The measured distances vs. the nominal together with the standard deviations of the measurements and the difference between measurements and simulations. The aspect angle was 0 degree (boats approaching toward the sensor). Reproduced from Steinvall, O., Tulldahl, M., 2016. Laser range profiling for small target recognition. Optical Engineering 56 (3), 031206.

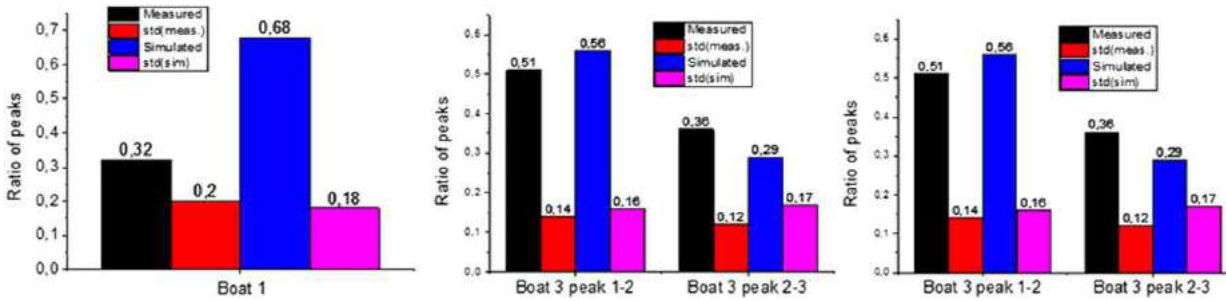


Fig. 31 The ratio of peaks in the waveform for the unmodified boats. Boat 1 = A, boat 2 = B and boat 3 = C. Reproduced from Steinvall, O., Tulldahl, M., 2016. Laser range profiling for small target recognition. Optical Engineering 56 (3), 031206.

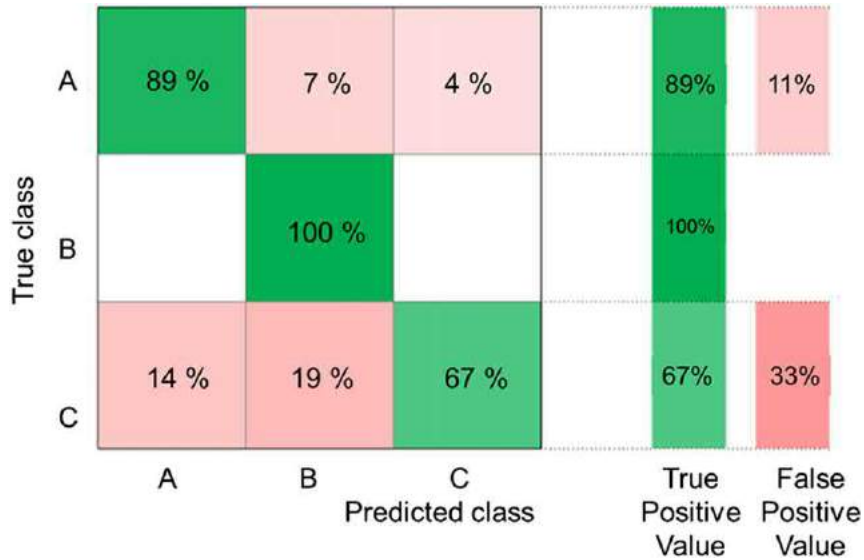


Fig. 32 Confusion matrix for classification using distance and ratio of the three peaks in each range waveform. Right: results for true positive and false negative. The classification accuracy for a decision tree classifier was 89%. Reproduced from Steinvall, O., Tulldahl, M., 2016. Laser range profiling for small target recognition. Optical Engineering 56 (3), 031206.

For radar it has been argued that the magnitude of the HRR profiles is too uncertain to use for target recognition, and one should instead use the positions of peaks in the HRR profile. In Zwart (2004), Zwart applies this approach using the same in-flight measured radar data of five aircrafts with promising results. Furthermore, Eigen template based recognition have been investigated in a number of publications, see Xinwei *et al.* (2016) and references therein.

The laser range waveforms are less variable with respect to small aspect changes as the reflection is more or less diffuse in character. Thus, a more parameter-based processing technique seems appropriate to investigate beside correlation.

The correlation for LRP methods has also been tested on small boats (Steinvall *et al.*, 2015a). When the main peak matches a correlation coefficient as exemplified in **Fig. 28** was achieved. Note that the correlation value can be rather high even if other the peaks mismatches. In **Fig. 29** we exemplify the correlation coefficient for all measurements against the boat B of **Fig. 8** and at 0-degree aspect (approaching course) and a few at 180-degree aspect. Most of the 0-degree aspect correlation data show high values, above 0.9 for both maximum and mean values. Lower values can most probably be explained by poor pointing and beam jitter.

As seen from the correlation examples (also see Steinvall *et al.* (2015a)) the discrimination capability only based on correlation analysis may not be sufficient for a robust recognition. It is therefore of interest to add other parameters into the decision process. These can be related to the distance between peaks and the ratio of peak amplitudes.

The absolute distances as shown in **Fig. 30** are derived from the peak separation for both simulated and collected range profiles for 0-degree aspect. In general, we can see that the correspondence between the observed and simulated distances is rather good.

The mean peak amplitude ratios for the first and second pair of peaks for every waveforms are shown in **Fig. 31**, estimated from all measurements at 0 degree and all target types without modifications with the extra box and the pipe. From this we can see that the difference between measured and simulated ratios is rather good irrespective of the lack of detailed reflectivity information and information on how the beam was actually pointed.

Using all available measures that is two distances between peaks and the corresponding ratios of peaks both from measurements and simulations we arrive at a classification accuracy of 89% (**Fig. 32**). The true positive classification probabilities are now even better, namely 89% for boat A, 100 for boat B, and 67% for boat C with false rates as 11, 0 and 33%, respectively, for A, B, and C. The corresponding values for the positive prediction and false discovery rates are 89, 88, 83% and 11, 12, and 7% respectively.

Heuvel *et al.* (van den Heuvel *et al.*, 2009a) describes a full scenario processing LRP for target classification using a likelihood test. Two aircraft, of which the one to classify is assumed to be a F16, are approaching each other with a speed of about 0.9 Mach at high altitude. The simulated aircraft profiles belong to F117, F15, and F16 obtained from 3D models. An LRP sensor with 30 cm range resolution is mounted on the observation aircraft. Noise has been added so that the signal at distances above 10 km is dominated by noise. The two aircraft are approaching at an angle of about 90 degrees. Atmospheric effects are neglected which is motivated that the scenario is at high altitude. Detection and tracking of the aircraft is based on radar or IRST. **Fig. 33** shows the scenario and the result of the Bayesian identification. The F16 is identified at shorter ranges with a maximum probability of 95%. In **Fig. 34** the time axis corresponds to absolute ranges between the aircraft and the relative range within the profile. In this figure 250 s correspond to 30 km range and at 320 s the range between the two aircraft is at its minimum, about 5.6 km. This

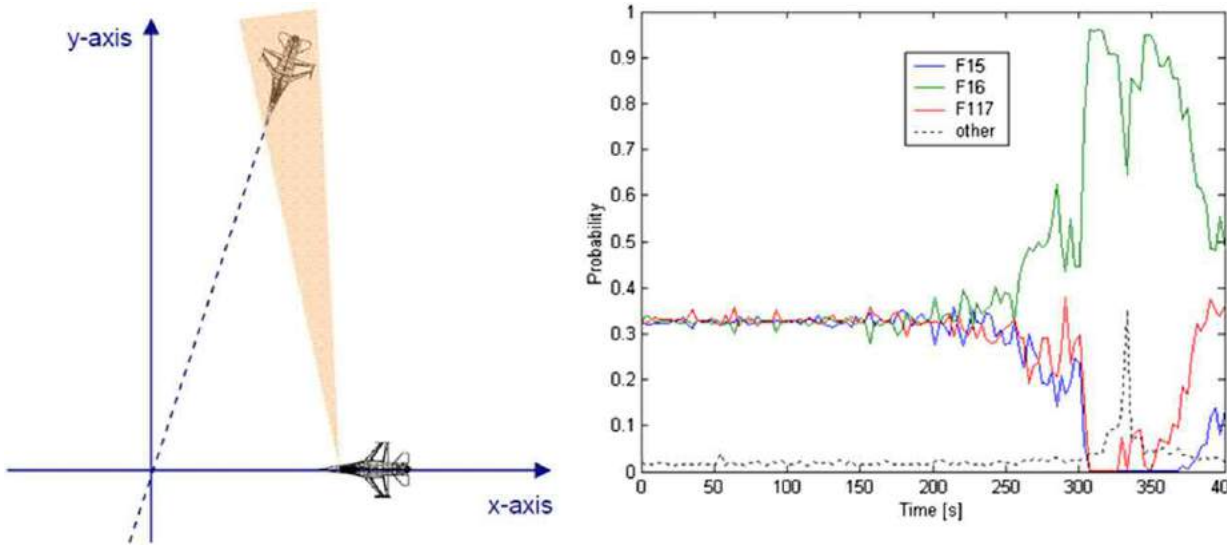


Fig. 33 Left: the scenario with two aircraft, of which the one to classify is assumed to be a F16, are approaching each other with a speed of about 0.9 Mach at high altitude. Right: the Bayes probabilities for an approaching F16 in the scenario. Reproduced from van den Heuvel, J.C., Schoemaker, R.M., Schleijpen, R.M.A., 2009a. Identification of air and sea-surface targets with a laser range profiler. Proceedings of SPIE 7323, 73230Y.

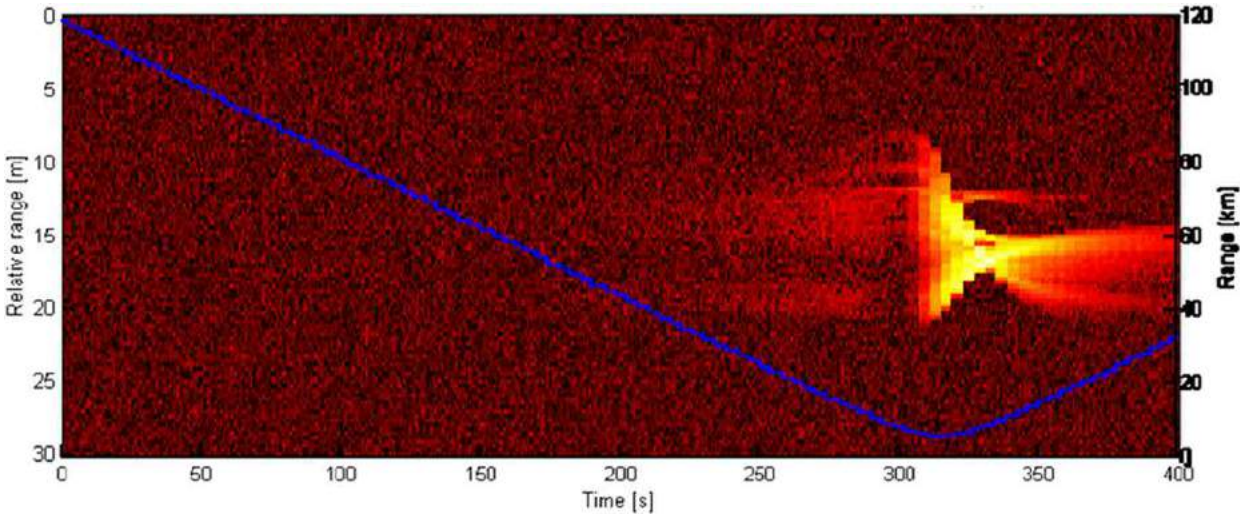


Fig. 34 Simulated laser range profiles for an approaching F16 according to the scenario. Every image column represents a single profile on a linear intensity scale. The blue line indicates the distance corresponding to the right axis. Reproduced from van den Heuvel, J.C., Schoemaker, R.M., Schleijpen, R.M.A., 2009a. Identification of air and sea-surface targets with a laser range profiler. Proceedings of SPIE 7323, 73230Y.

scenario is of high interest as it contains a dynamic change of aspect angle as well as range (and thereby a changing SNR of the waveforms).

Finally, there are a lot of other signal processing techniques which are suitable for target recognition using range profiles. These include machine learning and deep learning involving neural networks, support vector machines and others.

Conclusions

Although not by far as investigated as its microwave counterpart, LRP has some clear advantages to act as a complement or substitute to radar or IR imaging for target classification.

Compared to traditional angle–angle optical sensors LRP can offer range information along the target down to the sub-centimeter level, if needed. While traditional optical imaging needs large apertures, and often turbulence compensation, for long-range target classification, LRP is relatively insensitive to deblurring turbulence and scattering. It is also relative insensitive to first-order pointing error as long some of laser lobe covers the target. LRP can be a natural extension of present laser range finders. The maximum range at which laser range finder type of systems can produce target classification can extend tens of kilometers.

A laser profiling system can complement a microwave radar where it offers additional target classification properties, for example, for small targets using high resolution techniques. It may also back up a radar that is being jammed. The laser in a profiling system may serve as an illuminator for an imaging gated sensor and thus provide a robust recognition containing both angle–angle imaging and range profiling. Lasers can also provide accurate tracking using both imaging and quadrant detector receivers.

Tomography has been shown to be useful to image rotating objects. In general, however, the application of tomography to defense applications is probably limited due to a number of reasons. Among these we note alignment of the profiles may be difficult as well as obtaining many profiles over a sufficiently large angular sector to give a good image quality (Henriksson *et al.*, 2012a,b).

See also: Very High Range Resolution Lidars

References

Andressen, C., Anthony, D., DaMommio, T., *et al.*, 2005. Tower test results for an imaging LADAR seeker. Proceedings of SPIE 5791, 70–82.
Andersson, P., 2006. Long range 3D imaging using range gated laser radar images. Optical Engineering 45 (3), 034301–034310.
Armbruster, W., Hammer, M., 2012. Maritime target identification in flash-ladar imagery. Proceedings of SPIE 8391, 83910C.
Chevalier,T., 2011. A ray tracing based model for 3D ladar systems. In: Proceedings of the International Conference on Computer Graphics Theory and Applications, pp. 39–48.
Dierking, M.P., Heitkamp, F., Barnes L., 1998. High temporal resolution laser radar tomography for long range target identification. In: OSA Signal Synthesis & Reconstruction Conference.
Fujii, T., Fukuchi, T., 2005. Laser Remote Sensing. Boca Raton, FL: CRC Press.
Gschwendtner, A.B., Keicher, W.E., 2000. Development of Coherent Laser Radar at Lincoln Laboratory. Lincoln Laboratory Journal 12 (2), 383–396.

Gustavson, R.L., Davis, T.E., 1992. Diode laser radar for low-cost weapon guidance. Proceedings of SPIE 1633, 21–32.
Henriksson, M., Olofsson, T., Grönwall, C., Brännlund, C., Sjöqvist, L., 2012a. Optical reflectance tomography using TCSPC laser radar. Proceedings of SPIE 8542, 85420E.
Henriksson, M., Olofsson, T., Grönwall, C., Brännlund, C., Sjöqvist, L., 2012b. Optical reflectance tomography using TCSPC laser radar. Proceedings of SPIE 8542, 85420 E.
Hudson, S., Psaltis, D., 1993. Correlation filters for aircraft identification from radar range profiles. IEEE Transactions on Aerospace and Electronic Systems 29 (3), 741–748.
Hutt, D.L., Thériault, J.-M., Larochelle, V.G., Mathieu, P., Bonnier, D., 1994. Estimating atmospheric extinction for eyesafe laser rangefinders. Optical Engineering 33, 3762–3773.
Jin, X., Levine, R., 2009. Bidirectional reflectance distribution function effects in ladar-based reflection tomography. Applied Optics 48, 4191–4200.
Jonsson, P., Hedborg, J., Henriksson, M., Sjöqvist, L., 2015. Proceedings of SPIE 9649, 964905.
Knight, F.K., Click, D.I., Ryan-Howard, D.P., *et al.*, 1989a. Laser radar reflective tomography utilizing a streak camera for precise range resolution. Applied Optics 28, 2196–2198.
Knight, F.K., Kulkarni, S.R., Marino, R.M., Parker, J.K., 1989b. Tomographic techniques applied to laser radar reflective measurements. Lincoln Laboratory Journal 2 (2), 143–158.
Kunz, G.J., Bekman, H.P.T., Benoist, K.W., *et al.*, 2005. Detection of small targets in a marine environment using laser radar. Proceedings of SPIE 5885, 5885F1.
Ladar Seeker/ATA Algoritm Captive Flight Test Results, 2017. Available at: http://www.fas.org/man/dod101/sys/smart/docs/locaas_Industry_Day/sld013.htm (accessed 08.02.17).
Lasche, J.B., Matson, C.L., Ford, S.D., *et al.*, 2009. Reflective tomography for imaging satellites: Experimental results,. In: Proceedings of SPIE, 3815. pp. 178–188.
Laurenzis, M., Woiselle, A., 2014. Laser gated-viewing advanced range imaging methods using compressed sensing and coding of range-gates. Optical Engineering 53 (5), 053106.
Li, H.J., Yang, S.H., 1993. Using range profiles as feature vectors to identify aerospace objects. IEEE Transactions on Antennas and Propagation 41 (3), 261–268.
Lutzmann, P., Göhler, B., van Putten, F., Hill, C.A., 2011. Laser vibration sensing: overview and applications. Proceedings of SPIE 8186, 818602.
Matson, C.L., Mosley, D.E., 2001. Reflective tomography reconstruction of satellite features – field results. Applied Optics 40 (14), 2290–2296.
Molebny, V., McManamon, P., Steinvall, O., Kobayashi, T., Chen, W., 2016. Laser radar: Historical prospective – from the East to the West. Optical Engineering 56 (3), 031220.
Murray, J., Triscari, J., Fetzer, G., et al., 2010. Tomographic lidar. In: OSA Conference Paper. Applications of Lasers for Sensing and Free Space Communications, San Diego, CA.
Pace, P., Steinvall, O., Chevalier, T., 2008. Automatic target classification using one dimensional laser profiling. Technical Memorandum DRDC, Ottawa, TN 2007-000.
Parker, J.K., Craig, E.B., Klick, D.I., *et al.*, 1988. Reflective tomography: Images from range-resolved laser radar measurements. Applied Optics 27, 2642–2643.
Qu, F., Hu, Y., Wang, D., 2011. Lidar reflective tomography imaging for space object. Proceedings of SPIE 8200,820015.
Schoemaker, R., Benoist, K., 2011. Characterisation of small targets in a maritime environment by means of laser range profiling. Proceedings of SPIE 8037, 803705.
Shan, J., Toth, C.K., 2009. Topograhic Laser Ranging and Scanning – Principles and Processing. Boca Raton, FL: CRC Press.
Shaw, A.K., Paul, A.S., Williams, R., 2013. Eigen-template-based HRR-ATR with multi-look and time-recursion. IEEE Transactions on Aerospace and Electronic Systems 49 (4), 2369–2385.
SICK-Sensor Intelligence, 2017. Designing transportation routes to be more efficient,safer and more environmentally friendly. Available at: https://www.sick.com/de/en/industries/traffic/c/g285087 (accessed 08.02.17).
Signal, 2006. Available at: http://www.afcea.org/signal/articles/anmviewer.asp?a=1099&print=yes (accessed 08.02.17).
Sjöqvist, L., Allard, L., Henriksson, M., Jonsson, P., Pettersson, M., 2013. Target discrimination strategies in optics detection. Proceedings of SPIE 8898, 88980K.
Sjöqvist, L., Allard, L., Pettersson, M., *et al.*, 2016. Optics detection and laser countermeasures on a combat vehicle. Proceedings of SPIE 9989, 998908.
Sjöqvist, L., Henriksson, M., Jonsson, P., Steinvall, O., 2014. Time-correlated single-photon counting range profiling and reflectance tomographic imaging. Advanced Optical Technologies 3, 187–197.
Steinvall, O., 2000. Effects of target shape and reflection on laser radar cross sections. Applied Optics 39, 4381–4391.
Steinvall, O., 2009. Laser system range calculations and the Lambert W function. Applied Optics 48, B1–B7.
Steinvall, O., Berglund, D., Allard, L., *et al.*, 2015a. Passive and active EO sensing of small surface vessels. Proceedings of SPIE. 96490I.
Steinvall, O., Chevalier, T., Grönwall, C., 2014. Simulation and modeling of laser range profiling and imaging of small surface vessels. Optical Engineering 53 (2), 029801.
Steinvall, O., Elmqvist, M., Chevalier, T., Brännlund, C., 2012a. Measurement and modeling of laser range profiling of small maritime targets. Proceedings of SPIE 8542, 85420I.
Steinvall, O., Elmqvist, M., Karlsson, K., Larsson, H., Axelsson, M., 2010. Laser imaging of small surface vessels and people at sea. Proceedings of SPIE 7684, 768417.
Steinvall, O., Persson, R., Berglund, F., Gustafsson, F., Öhgren, J., 2015b. Using an eye-safe laser rangefinder to assist active and passive electrooptical sensor performance prediction in low visibility conditions. Optical Engineering 54 (7), 074103.
Steinvall, O., Sjöqvist, L., Henriksson, M., 2012b. Photon counting ladar work at FOI, Sweden. Proceedings of SPIE 8375, 83750C.
Steinvall, O., Tulldahl, M., 2016. Laser range profiling for small target recognition. Optical Engineering 56 (3), 031206.
van den Heuvel, J.C., Pace, P., Steinvall, O., 2008. Laser opportunities in new naval missions. Naval Forces 29 (5), 46–50.
van den Heuvel, J.C., Schoemaker, R.M., Schleijpen, R.M.A., 2009a. Identification of air and sea-surface targets with a laser range profiler. Proceedings of SPIE 7323, 73230Y.
van den Heuvel, V., van Putten, F.J.M., Cohen, F.H., Kemp, R.A.W., Franssen, G.C., 2009b. Results from the Search-Lidar Demonstrator Project for detection of small sea-surface targets. Proceedings of SPIE 7482, 748205.
Wu, P., Ming-Jun, W., Ke. X., Yan-jun, G., Yang, T., 2014. Laser radar range profile, Doppler spectra and range resolved Doppler imaging technologies for the target recognition. Proceedings of SPIE 9299, 929909.
Xinwei, W., Liang, S., Pingshun, L., *et al.*, 2016. Range-intensity coding under triangular and trapezoidal correlation algorithms for 3D super-resolution range-gated imaging. Proceedings of SPIE 9988, 99880U.
Zwart, J.P., 2004. Aircraft recognition from features extracted from measured and simulated radar range profiles. PhD Thesis, Universiteit van Amsterdam.

Micro-Lidars for Short Range Detection and Measurement

Vasyl V Molebny, Academy of Technological Sciences of Ukraine, Kiev, Ukraine

Introduction

The term micro-lidar has its origin from the Greek word *μικρό* (small) and the abbreviation of "light detection and ranging" meaning "detection and ranging with light." Often, a similar abbreviation "ladar" is in use where the word "light" is substituted by "laser" (light amplification by stimulated emission of radiation). Another form is "laser micro-radar" ("radar" here is for "detection and ranging with radio waves"). We shall define the laser radar, or lidar, with a prefix "micro" when the distance between the instrument and the object is of the order of a meter or less. Of course, the subdivisions can be involved, creating a series of prefixes from "mini" to "femto" and smaller. The physical principles of detection and ranging can be the same not depending on distance (range) to the object, but can be modified and fitted only for small distances.

Scope of the Chapter

Actually, the "detection and ranging" in the term "lidar" should be understood wider than only "detection" and "measuring the range." It means acquiring the information about the objects or media due to receiving the optical radiation reflected or scattered in backward direction. Usually, the source of radiation is a laser generating visible or invisible light. There is no severe restriction on the direction of the radiation scattering strongly backward.

The most often used principle of ranging with lidars is based on time-of-flight measurement. This principle is well known when determining the distance from a lightning by taking the count of the time-of-flight of the sound from the moment of light outburst. The idea to use short pulses of light to measure distance was brought out in 1930s by Lebedev (2014). His prototype measured the range up to 3.5 km with the accuracy about 2 m. The principle is applicable for a wide range of distances, from thousands of kilometers to micrometers – for measuring the distance to the Moon, tracking the player position by a play station, or checking the topography of the surface of microchips. It can be a single-dimension measurement (the distance between the lidar and the object), two-dimensional (2D) measurement (e.g., the profile of the surface), and three-dimensional (3D) measurement (e.g., a tomogram of the object).

The light propagates about 0.3 m for 1 ns. If to use bursts of light for measuring a small distance, their duration should be at least an order shorter. There would not be a problem to generate short laser pulses, but when designing a lidar for short distances, the problem is shifted to the detectors and processing. That is why, the time-of-flight technique at small distances uses a phase shift between the sent and the received signals. The micro-lidar can be designed to measure the phase difference between the light carrier frequencies, or between the modulation frequencies.

Continuous-wave (CW) light radiation is used for surface profile measurement, length measurement of small objects, like intraocular distances, small thickness layers or films. To get higher sensitivity, synthetic aperture is added by moving the instrument relatively to the investigated object. Specific field structure can be created by an array of light emitting diodes, enabling to acquire 3D image.

Interferometric methods initiated a new technology of the optical low-coherent tomography (OCT) with multiple modifications and wide practical use, the most exciting of which is a 3D imaging of the structure of the eye. Impressive are also results of the OCT in cerebrovascular and cardiovascular studies and diagnosing. Low-cost nomenclature of OCT using a smartphone is represented by multi-reference technique with its implementation in dermatology.

With micro-lidar, a series of media parameters can be measured. Analysis of Raman spectra allows differentiation between malignant and benign tissues. Analysis of fluorescence delivers information on blood circulation and can help in diagnosing the diseases of the eye. Differences in absorption of light in the anterior chamber of the eye can serve for measurement of the glucose level. Spectral differences in light scattering and absorption for different wavelengths are the basis for the method of measurement of blood oxygenation, including the oxygenation of the vessels of the retina.

In long-range laser radars, it is very important to keep the direction on the target. Similar, but not less sophisticated is the micro-lidar of the CD/DVD pickup head that should keep the head respectively to the fast running track not only along its run, but also at a certain distance in focus of the photodetector. Computer mouse is another example of micro-lidar. Massive production of computers and their periphery made their components so perfect, that they become the basics for other devices, like, for example, the pickup head in the role of the driver of the reference mirror in the smartphone-based OCT instrument.

With the advent of new technologies of vision correction, a problem arose of measurement of the refraction non-homogeneity of the optical system of the eye. Several solutions were found, some of them being typical laser radars. Ray tracing consisted in sounding the eye with narrow laser beam in different points of the eye aperture and measuring the coordinates of the light spot of the laser projection on the eye bottom. The inversion of this method was projecting a single laser beam on the retina, that created there a secondary source of light, and measuring the distortion of the image of this spot of light by the optical system of the eye at the exit from the eye.

 doi:10.1016/B978-0-12-803581-8.09433-9

The potential of the micro-lidar to measure the velocity of small particles in gas or liquid flow was one of the earliest implementations. Laser velocimeter was called sometimes the anemometer. Among its applications was measurement of missile plume flow, turbulence statistics of the submarine wake, velocity of blood flow in retinal vessels, etc. Several methods were developed using noncoherent and coherent radiation. Translational and axial components of the flow were measured. Doppler principle was incorporated into the OCT instruments enabling the measuring and visualization of the 3D distribution of the velocity in the flows.

Distance Measurement Techniques

Pulse Mode

Since the middle of the 19th century, short light pulses assisted in many discoveries in nature studies (Shapiro, 1984). It was natural that the development of picosecond lasers initiated new experiments in designing the micro-lidars. Park *et al.* (1981) demonstrated an optical ranging system which can see through opaque material. It uses picosecond pulses and a very fast gate with a variable delay at the detector. The system could detect a reflection from a mirror inside an opaque mixture with 1.7 mm range resolution and 0.1 mm transverse resolution. The depth of penetration corresponded to 0.67 cm in human tissue.

A possibility of gated viewing at a small-pathlength difference between object beam and reference beam was demonstrated on recording a hologram with the use of laser light of short pulse duration or short coherence length (Abramson, 1978). Only those parts of the object surface were holographically recorded that corresponded to a small-duration gate.

Fujimoto *et al.* (1986) described the application of femtosecond laser pulses to perform optical ranging using nonlinear cross-correlation gating. The laser pulses are focused onto the sample under study, and the remitted reflections and echoes are measured by nonlinear-optical cross correlation with the incident pulse. The experimental setup (**Fig. 1**) uses pulses of 65 fs duration at a repetition rate of 125 MHz, laser wavelength 625 nm, and an average power of up to 20 mW. The laser pulses are focused onto the sample under study, and the remitted reflections and echoes are measured by nonlinear-optical cross correlation with the incident pulse. The reference pulse travels through an optical delay line controlled by a stepping motor. The signal and reference are correlated by focusing with a crossed-beam geometry into a 0.5-mm length of angle-phase-matched potassium dihydrogen phosphate (KDP) second-harmonic crystal. Measurement of the integrated second-harmonic energy as a function of the temporal delay between the signal and the reference yields the cross correlation. Spatial resolutions of less than 15 μm was achieved. An example of signals is shown in **Fig. 2**.

CW Phase Difference Mode

CW modulated signals

In a phase-shift rangefinder, the optical power is modulated with a constant frequency. After reflection from the target, a photodetector collects a part of the laser beam. Measurement of the distance D is deduced from the phase shift $\Delta(\varphi)$ between the

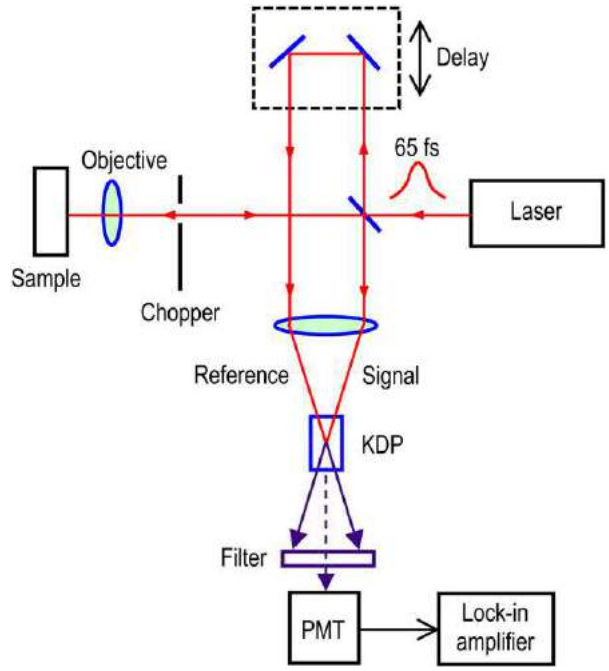


Fig. 1 Schematic of femtosecond optical ranging experiment. BS, beam splitter; PMT, photomultiplier tube; XTAL, nonlinear potassium dihydrogen phosphate (KDP) crystal. Reproduced from Fujimoto, J.G., De Silvestri, S., Ippen, E.P., *et al.*, 1986. Femtosecond optical ranging in biological systems. Optics Letters 11, 150–152.

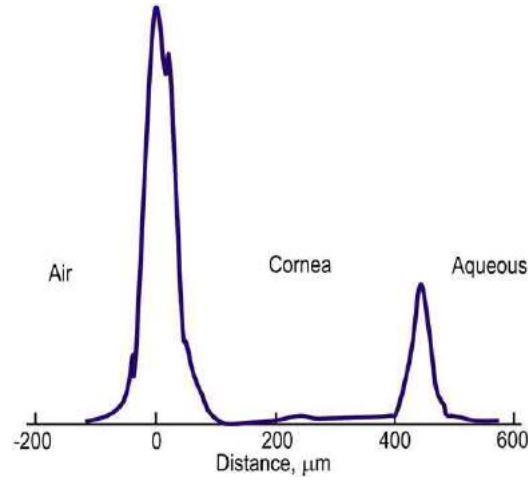


Fig. 2 Optical ranging measurement of the cornea of the rabbit eye performed in vivo. Reproduced from Fujimoto, J.G., De Silvestri, S., Ippen, E. P., *et al.*, 1986. Femtosecond optical ranging in biological systems. Optics Letters 11, 150–152.

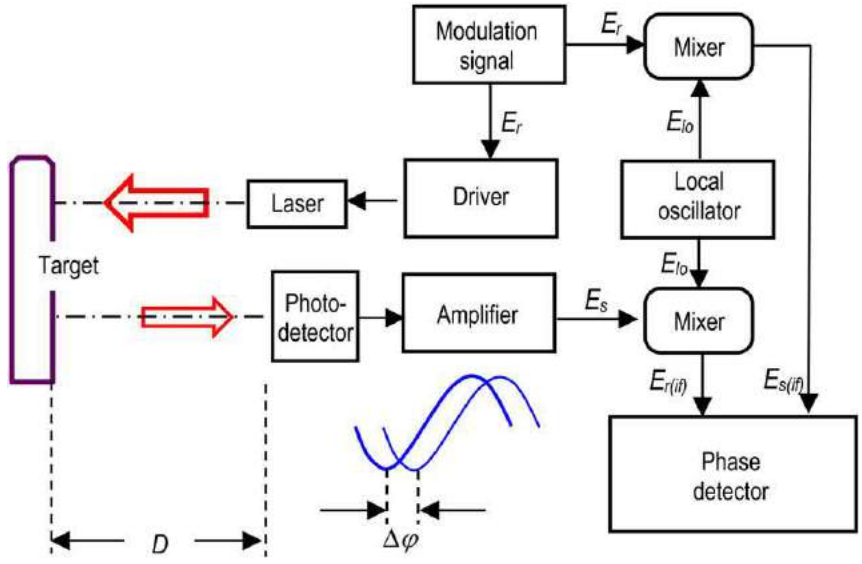


Fig. 3 Block diagram of a phase-shift laser rangefinder using a heterodyne technique. Reproduced from Hu, P., Tan, J., Yang, H., *et al.*, 2011. Phase-shift laser range finder based on high speed and high precision phase-measuring techniques. In: 10th International Symposium on Measurement Technology and Intelligent Instruments. June 29–July 2, 2011, KAIST, Daejeon, Korea.

received and modulated emitted signal: $D=(cF/2)((\varphi)/2\pi)$. The unambiguous distance measurement is limited to $(\varphi)=2\pi$. This principle is often used with conversion of the modulation frequency to a lower intermediate frequency (Amann *et al.*, 2001; Hu *et al.*, 2011).

In the example of **Fig. 3**, the laser beam is modulated by a sine voltage E_r. After the detection and filtering, the voltage E_m is selected. Both voltages E_r and E_m are mixed with the signals E_{lo} from the local oscillator. The phase difference (φ) is measured between the signals E_{ifm} and E_{ifr} from the outputs of the mixers. This phase difference is then recalculated into the distance D. To get the measurement resolution 2 mm, at the measurement range $D=15$ m and with the modulation frequency 10 MHz, the phase-shift should be measured with the resolution 0.05 degree.

Journet and Poujouly (1998), measured the phase shift at an intermediate frequency of 125 kHz. The emitter was a laser diode with 12 mW average output power, the receiver used an avalanche photodiode. The phase shift was measured by direct counting with more than 10 bits resolution, that provided about 0.4 mm for a 60 cm range without phase ambiguity. A computer controls the range-finder through a specially developed interface.

In a later version, Poujouly and Journet (2000) used a digital phase-locked loop to reduce the phase noise (Staszewski and Balsara, 2005). They also described the technique based on under-sampling, applied to digital synchronous detection. Its main advantage is a global simplification of the electronic system, leading to a quite simple development of a twofold

modulation system. This new technique is also very interesting to move toward a kind of smart rangefinder able to adapt different parameters to the different steps of the measurement.

Journet and Lourme (2000) proposed to measure the correlation between the transmitted and the received modulation signals this type of modification is not restricted by the sine modulation, because the distance is determined by the delay of the reference signal, at which the correlation has maximum.

Sugimoto *et al.* (2013) proposed to use two sine modulation signals of different frequencies forming a sinc-similar beating signal (**Fig. 4**), specific points of which are used as references to measure the delay of the received signal relatively to the transmitted signal.

In a wide use in the long-range laser radars (Agishev, 2009), the continuous-wave frequency modulation (CWFM) with a linearly varying frequency is used (**Fig. 5**), often called chirp modulation due to its similarity to bird's chirping. For this technique,

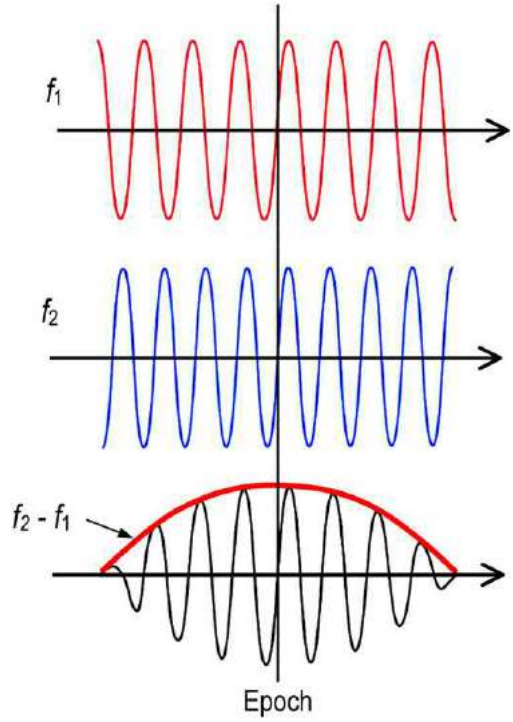


Fig. 4 Two modulation frequencies and their sum (an epoch is a specific point of a central zero crossing in the beating signal). Reproduced from Sugimoto, M., Nakamura, S., Inoue, Y., *et al.*, 2013. LT-PAM: A ranging method using dual frequency optical signals. International Journal on Smart Sensing and Intelligent Systems 6 (3), 791–809.

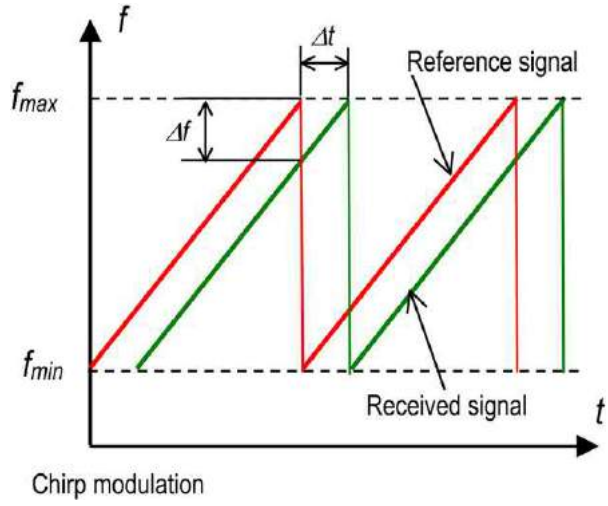


Fig. 5 Time-frequency relationship in a chirp signal. Reproduced from Agishev, R.R., 2009. Lidar Monitoring of the Atmosphere. Moscow: Fizmatgiz (in Russian).

the frequency difference between the transmitted and received signals is proportional to the time delay and correspondingly, to the measured distance.

Frequency difference techniques

Axial eye length measurement

Sekine *et al.* (1993) described a noncontact technique for axial eye length measurement. According to this method, the wavelength shift of a single-mode laser-diode beam that is irradiated onto the eyeball causes a phase shift in the interference fringes of reflections from the retina and the cornea. Then the optical distance between the cornea and the retina is obtained from the phase-shift measurement.

The schematic arrangement of the optical system is shown in **Fig. 6** with the following notations: BS – beam splitter; C – cornea; CL – collimator lens; OA – optical path compensator; OI – optical isolator; R – retina; SF – spatial filter. The light source is a single-mode laser diode (LD). Illumination light coming from the LD is first split to a measuring interferometer and a reference light path. Light that enters the interferometer is again split by a beam splitter and converged separately at the cornea and retina. Reflections from the cornea and retina return to each optical system and interfere with each other. They are then received by a photodetector PD_1. On the other hand, light that enters the reference light path is reflected by mirrors M_l and M_2, and these reflections cause interference before they are received by photodetector PD_2.

Smoothly varying the light frequency in a given range, the number of phase zeros is calculated in the measuring and in the reference arms (N_m and N_r). It is shown that the eye length L_{eye} can be calculated as $L_{eye}=L_{ref}\,(N_m/N_r)$, where L_{eye} is the eye axial optical length, $L_{ref}=L_1-L_2$. $L_{opt}=L_{opt1}-L_{opt2}$ is adjusted to be zero.

Compared to the technique that uses partially coherent light, this technique is inferior in terms of measurement accuracy but superior in its wide, measurable range of 16–32 mm. The results of measurements showed that double standard deviation of measurement was $2\sigma=\pm0.11$ mm.

Wavelength tuning interferometry

The idea of measuring the distance using the interferometer with changing frequency was implemented by Lexer *et al.* (1997) as the laser wavelength-tuning interferometry of intraocular distances (**Fig. 7**). The notations are as follows: OS – object signal; RS – reference signal; SP – scattering potential; MI – reference Michelson interferometer; FT – Fourier transform; NC – numerical correction; PD_1, PD_2 – photodetectors; TL – tunable laser. Shifting the wavelength of a tunable laser diode causes intensity oscillations in the interference pattern of light beams remitted from the intraocular structure. A Fourier transform of the photodetector signal yields the distribution of the scattering level along the light beam illuminating the eye (**Fig. 8**). The authors demonstrated simultaneous measurement of the anterior segment length, the vitreous chamber depth, and the axial eye length in human eyes in vivo with data-acquisition times in the millisecond range.

Dual-beam double frequency configuration

See *et al.* (1985) applied the differential phase-contrast scanning microscope for the surface studies. Molebny *et al.* (1996a,b) described similar configuration to measure the ablation profiles after the Lasik surgeries for the correction of the human vision (**Fig. 9**). The laser beam having its carrier frequency f_0 is split by the acousto-optical modulator in two components (f_0+f_1) and (f_0+f_2). The frequencies f_1 and f_2 are generated by the driving generator. After being reflected from the specimen, these components are fed to the photodetector whose output is connected to the phase detector. The phase difference $\Delta(\varphi)$ at the frequency $\Delta f=(f_2-f_1)$ is the measure of the height difference Δh of the surface along the line connecting the beam projections on the surface. To reconstruct the profile, the surface is scanned along the line in the direction of the line connecting

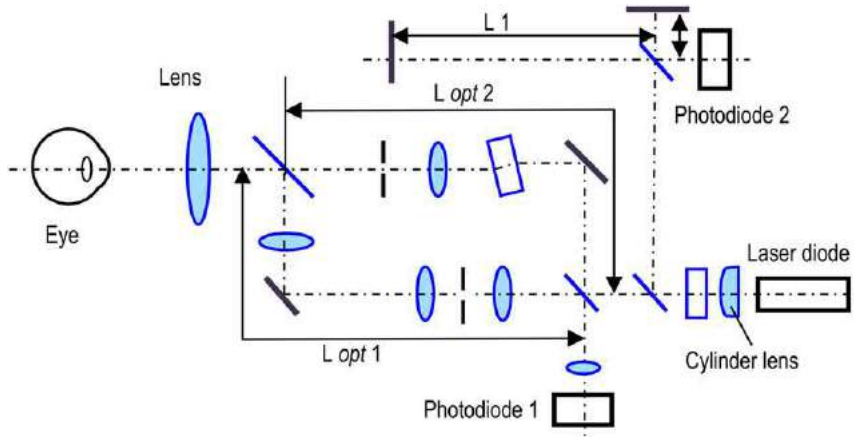


Fig. 6 Scheme of the interferometric system for axial eye-length measurement. Reproduced from Sekine, A., Minegishi, I., Koizumi, H., 1993. Axial eye-length measurement by wavelength-shift interferometry. Journal of the Optical Society of America A 10, 1651–1655.

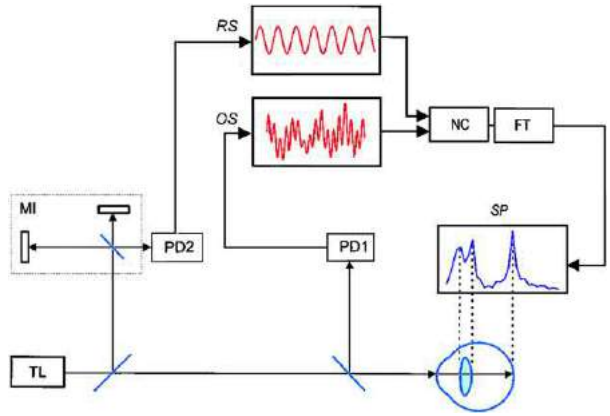


Fig. 7 Scheme of the wavelength tuning interferometer. Reproduced from Lexer, F., Hitzenberger, C.K., Fercher, A.F., Kulhavy, M., 1997. Wavelength-tuning interferometry of intraocular distances. Applied Optics 36, 6548–6553.

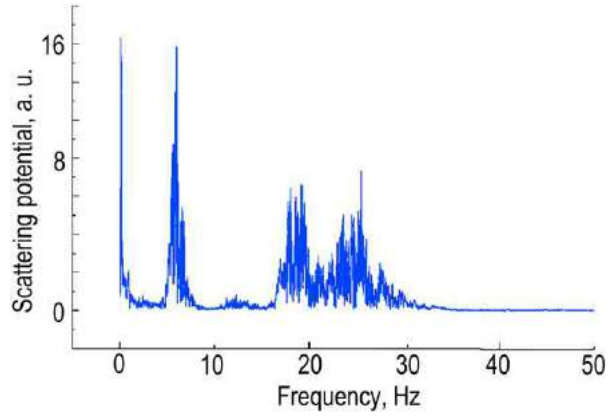


Fig. 8 Interferometry scan of an eye model. Reproduced from Lexer, F., Hitzenberger, C.K., Fercher, A.F., Kulhavy, M., 1997. Wavelength-tuning interferometry of intraocular distances. Applied Optics 36, 6548–6553.

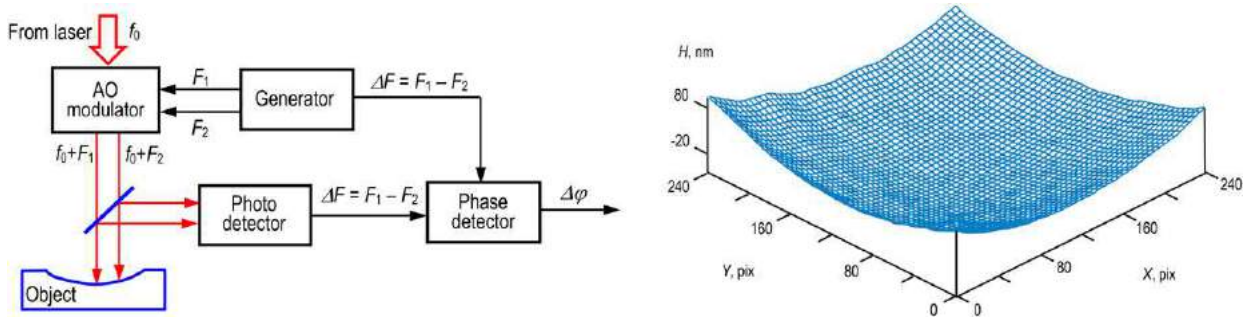


Fig. 9 Dual-beam double frequency configuration for measurement of the ablation profiles after Lasik surgery of vision correction. To the right: an example of the reconstructed profile using a triple-beam three-frequency design. Reproduced from Molebny, V.V., Pallikaris, I.G., Naoumidis, L. P., *et al.*, 1996. Dual-beam dual-frequency scanning laser radar for investigation of ablation profiles. Proceedings of SPIE 2748, 68–75.

the beam projections (e.g., along the x direction). The next line of scanning is one-step shifted in y-direction. With laser wavelength 630 nm, $f_0 = 80$ MHz and $\Delta f = 1$ MHz, the sensitivity to profile measurement was achieved better than 10 nm. To make the data for the next line of scan more defined, a triple beam configuration was used with beams split in x and y directions (Molebny *et al.*, 1998).

Instead of splitting the frequencies with high-cost acousto-optic modulators (AOMs), a gas laser (Bourdet and Orszag, 1979) or modern two-wavelength laser diodes can be used (Ishii and Onodera, 1991). High relative wavelength stability for large variations in optical path length can be achieved with two or more single-mode lasers locked to a Fabry–Pérot etalon

(de Groot and Kishner, 1991; Gerstner and Tschudi, 1994). A simpler configuration for short path differences can be designed using multimode laser diodes (Rovati *et al.*, 1998; de Groot, 1991).

3D laser microvision

The 3D laser microvision developed by Shimotahira *et al.* (2001) is based on the principle of a step frequency radar. Optically, it is an interferometer that derives the distance having the information of the phase delay of the scattered signal (**Fig. 10** with the notations: NBS – non-polarizing beam splitter; AOM – acousto-optic modulator; PZT – piezoelectric transducer). The carrier frequency from the superstructure grating (SSG) laser source is swept stepwise: $f_n = (f_0 + n\Delta f)$. The wavelength of the SSG laser diode is centered at 1.545 μm. In each frequency step, the following procedure is repeated. The output laser light is split into the probe and reference beams by means of an AOM. The non-refracted beam (the zero-order beam) from the AOM is used as a probe beam, and the refracted beam (the first-order beam) from the AOM whose carrier frequency is shifted from that of the probe beam by the frequency of excitation of the AOM is used as a local oscillator beam. The probe beam reflected from the target is mixed with the reference beam from the AOM in a coherent detector. The heterodyne-detected signal, both amplitude and phase, is stored. After all frequency stepping is completed, the stored data are processed and a 3D image of the target is displayed.

The lateral resolution (x–y plane resolution) depends on the focusing parameters of the objective lens, and it is best only at one focused depth. To remove this difficulty, the synthetic aperture approach is used coming from that of the microwave side-looking airborne radar (SAR) which makes high-resolution maps from scanned microwave echoes. A linear flight path of an airplane is used as a linear scan path of the probing antenna. With the 3D laser microvision, the light scattered from the target is measured while the probe is moving away from the target along a path perpendicular to the surface of the target. Thus the scanned probe collects information about the scatterer and significantly improves the lateral resolution as well as the depth resolution of the 3D laser microvision.

In this instrument, the laser wavelengths cover a range as wide as 30 nm centered at 1.545 μm with an output power of 1 mW. A specially developed digital phase meter can measure the phase with an accuracy of 0.03 degree as the time delay between the rising edge of both, returned and reference signals. To obtain 3D information, the probe beam has to be scanned over a 2D surface, 128 × 128 pix large, and at each location of it, the depth information is probed to construct a 3D image.

Fig. 11 demonstrates the effectiveness of the synthetic aperture method using a target of a micro-strip line array on a Gallium arsenide (GaAs) wafer (a – without the synthetic aperture scanning, b – with the synthetic aperture scanning). The micro-strip lines are shown at the bottom of the figure. The width of each micro-strip line is 15 μm, the thickness is less than 1 μm, and the spacing between the adjacent edges of the micro-strip line is 8 μm. The vertical synthetic aperture was made in the stepwise z direction. Curve b in **Fig. 11** shows the result of the synthetic aperture scanning. The image clearly separates the edge spacing of the micro-strip lines. **Fig. 12** shows a microscope photograph of a tunneling photodiode that is deposited on the GaAs substrate. The overall deposit area is 500 μm × 500 μm.

Interferometric Methods

Optical Interferometric Micro-Lidar

Time-of-flight measurement can be combined with interferometry using a Michelson interferometer, where one of the mirrors is replaced by a sample under test (Dresel *et al.*, 1992; **Fig. 13** left). A light source with short coherence length is used. The reference

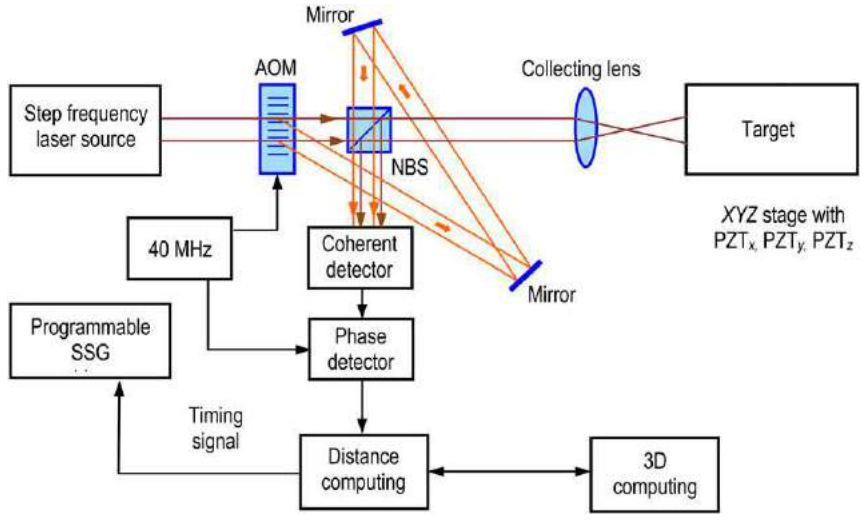


Fig. 10 Block diagram of the 3D laser microvision. Reproduced from Shimotahira, H., Iizuka, K., Chu. S.C., *et al.*, 2001. Three-dimensional laser microvision. Applied Optics 40, 1784–1794.

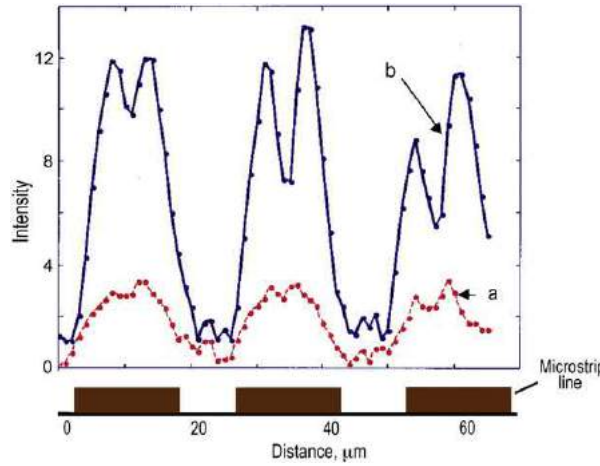


Fig. 11 The target of the three-dimensional (3D) laser microvision – gallium arsenide (Gaas) substrate with thin metal electrodes. Reproduced from Shimotahira, H., Iizuka, K., Chu. S.C., *et al.,* 2001. Three-dimensional laser microvision. Applied Optics 40, 1784–1794.

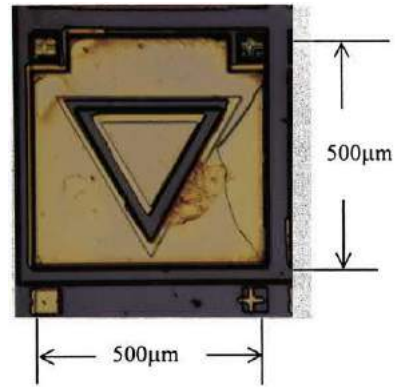


Fig. 12 Measured results of a micro-strip line array on a gallium arsenide (GaAs) wafer. Reproduced from Shimotahira, H., Iizuka, K., Chu. S.C., *et al.,* 2001. Three-dimensional laser microvision. Applied Optics 40, 1784–1794.

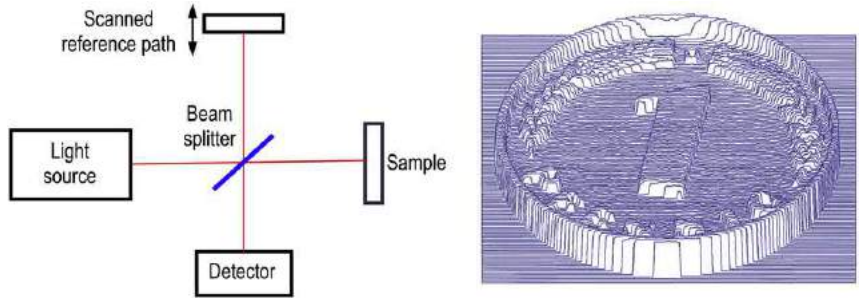


Fig. 13 Basic setup of an optical microradar (left); three-dimensional (3D) reconstruction of a coin image (right). Reproduced from Dresel, T., Häusler, G., Venzke, H., 1992. Three-dimensional sensing of rough surfaces by coherence radar. Applied Optics 31, 919–925.

mirror plane and the sample plane are imaged onto a detector. The image of the sample is superimposed with a reference wave. Interference takes place when the light paths from the reference and the sample are equal and only within those speckles that correspond to the surface elements close to the sample plane. These regions are detected and stored, while the reference mirror is scanned along the z axis.

The profile of the sample can be reconstructed by scanning in (x, y) directions. An example of a reconstructed profile is presented in **Fig. 13** right.

The first biological application of low-coherence interferometry for the measurement of axial eye length was reported by Fercher *et al.* (1988). Different versions of low-coherence interferometry were developed for noninvasive measurement in

biological tissues. Taking an eye as an object, there will be several positions of the reference mirror along the z axis, where the interference gives maximal signal (Fig. 14 left). They correspond to the cornea, lens, and retina. Scanning in (x, y) directions yields the profiles of these surfaces. Fig. 14, right, shows an example of the structure of the human retina reconstructed from the data of measured intensity of light scattered from the eye bottom, the beam being shifted laterally scan-by-scan.

Optical Low-Coherence Tomography

Time-domain optical low-coherence tomography

This interferometric lidar technique was soon baptized by Huang *et al.* (1991) as the optical coherence tomography (OCT). To be more accurate, it should be named the optical low-coherence tomography. OCT imaging system based on a fiber-optic Michelson interferometer is shown in Fig. 15. The advantage of the fiber optics technology is that the setup is robust and easier in alignment. In the schematic, a Michelson interferometer contains a circulator for dual balanced detection. Dual balanced detection adds the signal from the interference of the sample and reference arms and subtracts excess noise from the light source. The sample/probe arm may be interfaced to a variety of imaging devices.

Early OCT instruments used a low-coherence light source and interferometer with a scanning reference delay arm. This method is known as a traditional or time-domain OCT (TD-OCT). OCT found its first application in ophthalmology for diagnosing ocular diseases such as glaucoma, age-related macular degeneration, and diabetic retinopathy.

Spectral-domain optical low-coherence tomography

Later on, the spectral/Fourier domain OCT (spectral-domain optical low-coherence tomography (SD-OCT)) and the swept source/Fourier domain OCT (SS-OCT) were developed, also termed optical frequency domain imaging (OFDI). The SD-OCT has a powerful sensitivity advantage over the time domain detection, since it measures all of the echoes of light simultaneously. This discovery drove a boom in OCT research and development. The sensitivity is enhanced by 50–100 times, enabling a corresponding increase in imaging speeds. The second type of Fourier domain detection, SS-OCT, uses an interferometer with a narrow-bandwidth, frequency swept light source and detectors which measure the interference output as a function of time. It has the advantage that it does not require a spectrometer and line scan camera. Therefore, it can operate at longer wavelengths where

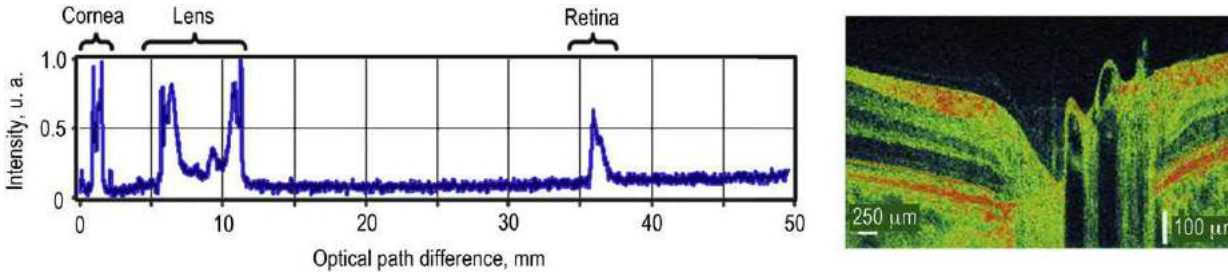


Fig. 14 Intensity distribution along the path of the beam in the eye (left), and the color-coded reconstruction of the intensity of the light scattered from the retina when the beam is scanned laterally (right). Reproduced from Fercher, A.F., Mengedoht, K., Werner, W., 1988. Eye-length measurement by interferometry with partially coherent light. Optics Letters 13, 186–188.

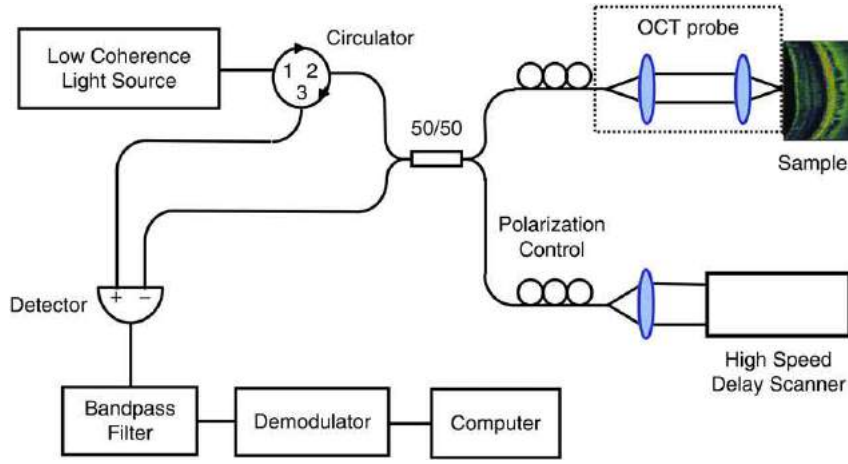


Fig. 15 Schematic of the time-domain optical coherence tomography (OCT). Reproduced from Drexler, W., Fujimoto, J.G., (Eds.), 2015. Optical Coherence Tomography, second ed. Switzerland: Springer.

camera technology is less developed and it can achieve imaging speeds which are much faster than spectral/Fourier domain OCT which is limited by the camera speed. The primary challenge in swept source/Fourier domain OCT is that it requires a high-speed, swept narrow line width light source.

Schematic of a typical spectral domain OCT instrument is presented in Fig. 16. The reference arm has a fixed delay and is not scanned. Interference is detected with a spectrometer and a high-speed, line scan camera. A computer reads the spectrum, rescales it from wavelength to frequency, and Fourier transforms to generate axial scans.

Swept-source optical low-coherence tomography

Schematic of a typical swept source OCT instrument is shown in Fig. 17. It contains a frequency swept light source. The reference arm has a fixed delay and is not scanned. The example shows a sample arm with catheter/endoscope interface and the system is assumed to operate at 1.3 μm wavelengths. This geometry enables dual balanced detection, canceling excess noise in the laser. A portion of the frequency swept light is directed into a Mach–Zehnder interferometer which acts as a periodic frequency filter. This interferometer configuration is more efficient than the classic Michelson interferometer because all of the light is detected. It avoids the spectral resolution and pixel limitations which are inherent in spectrometers and line scan cameras.

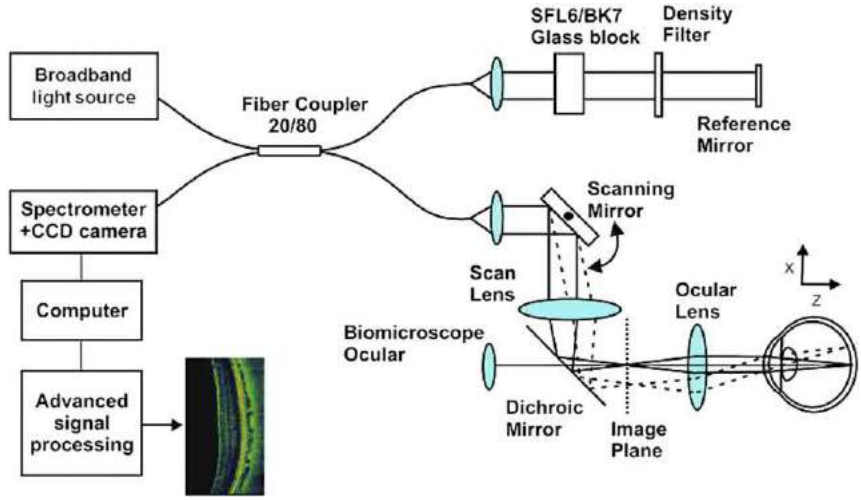


Fig. 16 Schematic of a typical spectral domain optical coherence tomography (OCT). Reproduced from Drexler, W., Fujimoto, J.G., (Eds.), 2015. Optical Coherence Tomography, second ed. Switzerland: Springer.

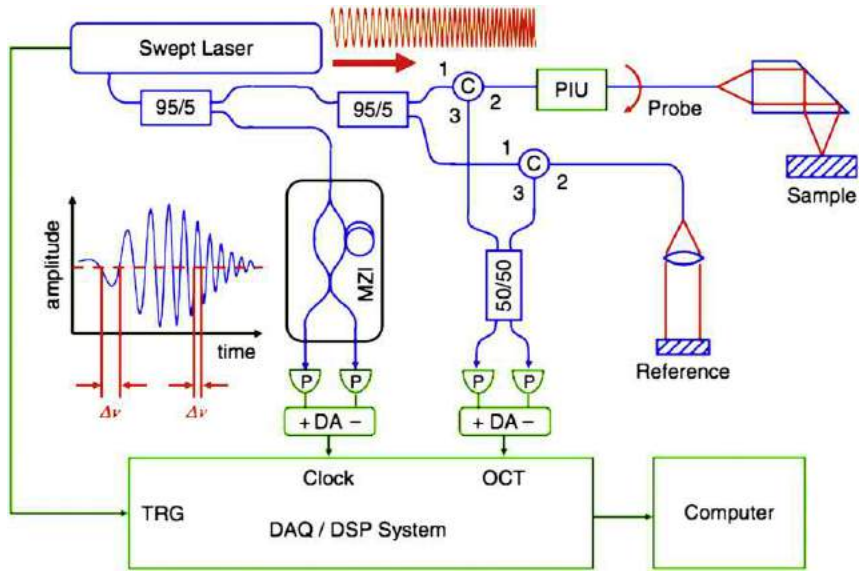


Fig. 17 Schematic of a typical swept source optical coherence tomography (OCT). Reproduced from Drexler, W., Fujimoto, J.G., (Eds.), 2015. Optical Coherence Tomography, second ed. Switzerland: Springer.

Commercial OCT Instruments for Ophthalmology

Fig. 18 shows (from left to right): (1) an early retinal imaging prototype instrument. The OCT system is interfaced with a slit lamp biomicroscope. The OCT beam is scanned using a pair of galvanometer-actuated mirrors. This system was used at the New England Eye Center during the mid-1990s; (2) Zeiss Cirrus spectral OCT – one of the latest, cutting-edge ophthalmic instruments available today; (3) Copernicus REVO spectral OCT.

Color encoding of the OCT data

Display techniques were developed to display OCT data in the form of thickness maps. An early example of a time-domain OCT topographic color-encoded map of retinal thickness is shown in **Fig. 19** left (Hee *et al.*, 1998). An example display of the spectral-domain Zeiss Cirrus instrument is shown in **Fig. 19** right (Mazzarella and Cole, 2015).

Fig. 18 Commercial optical coherence tomography (OCT) instruments for ophthalmology. Reproduced from Drexler, W., Fujimoto, J.G., (Eds.), 2015. Optical Coherence Tomography, second ed. Switzerland: Springer.

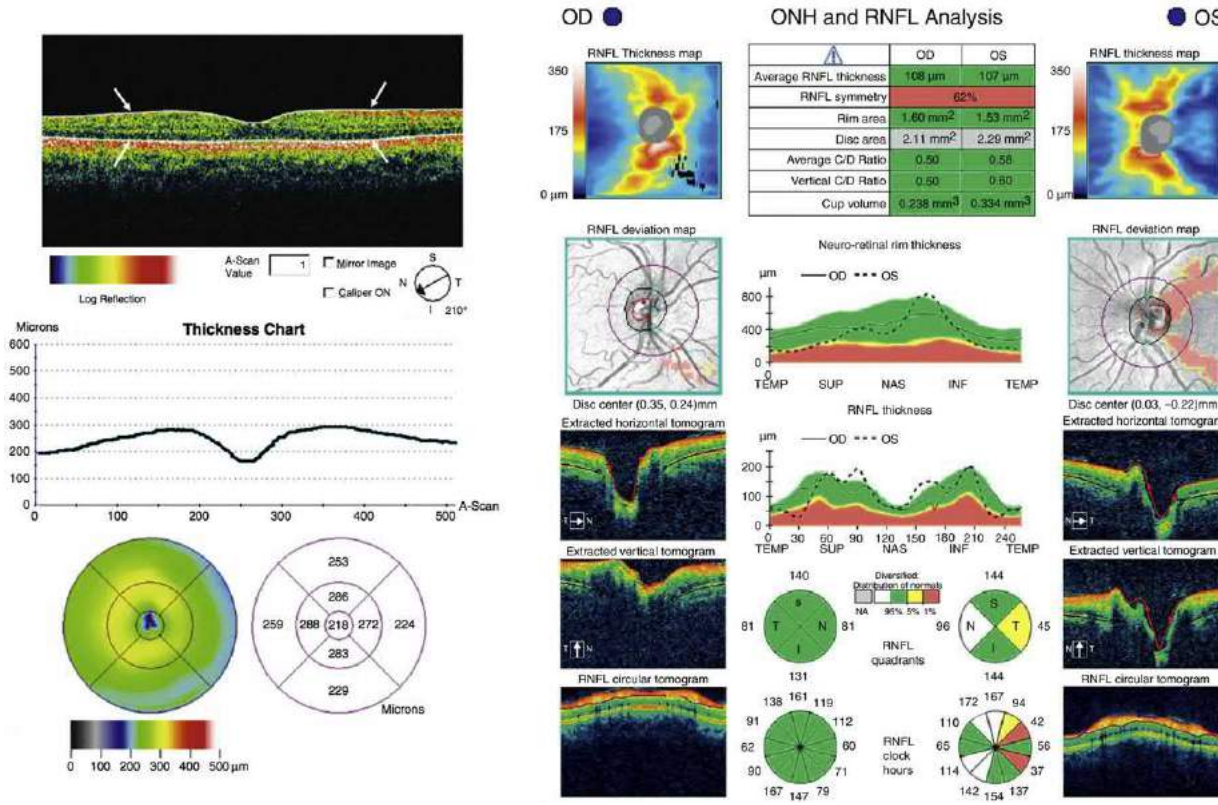


Fig. 19 Display of an early optical coherence tomography (OCT) instrument with color-encoded topographic map of retinal thickness (left). Cirrus OCT display in the glaucoma diagnosing mode (right). *Source*: internet

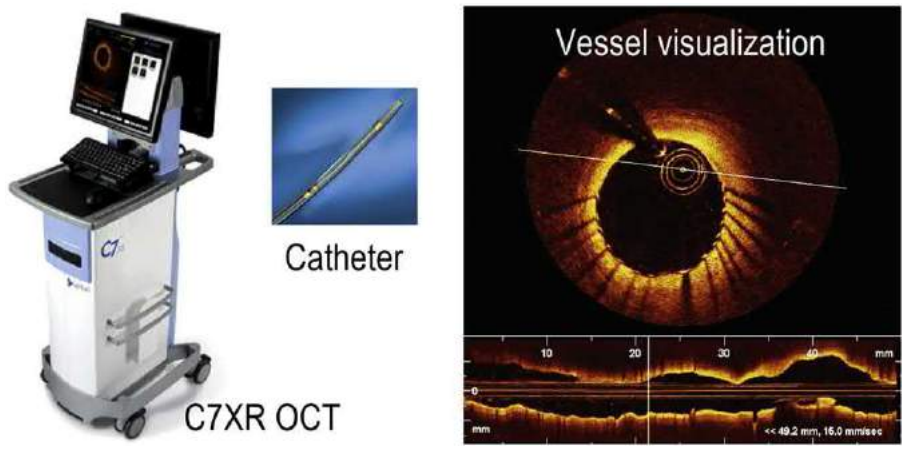


Fig. 20 High-speed intravascular C7XR optical coherence tomography (OCT) system (left); vessel visualization (right). *Source*: internet

OCT in Cerebrovascular and Cardiovascular Research

Fig. 20 left demonstrates the C7XR OCT system (Guiagliumi *et al.*, 2016), that was introduced as the first commercially available FD-OCT system for coronary imaging in 2009. **Fig. 20** right shows an example of intravascular OCT imaging (Adler, 2012). The mechanical actuation necessary to sweep the source wavelength is on the nanometer scale and therefore is accomplished in a compact MEMS-based packaging.

The C7XR system has a spatial resolution of 15 μm and acquires 500 axial lines per frame at a rate of 100 frames/s. At the tip of the catheter, there is a 125-μm-diameter optical fiber assembly that consists of an integral side-looking lens. The optical assembly is encapsulated in a hollow torque wire that translates motion from a drive motor located outside the patient's body. The C7XR system can volumetrically image a 5-cm vessel segment in approximately 3 s.

The latest generation of this type OCT system, the Ilumien Optis, has a higher image acquisition rate (180 frames/s) and can acquire a 75-mm pullback image in only 2.1 s (see Section Relevant Websites). This system has instant 3D reconstruction capability and is used for optimizing stent planning.

Another OCT system developed at the Massachusetts General Hospital (MGH, Boston, MA) has a high-speed wavelength swept laser that uses a polygon-mirror/grating filter to tune over a broad (120-nm-wide) bandwidth centered at 1320 nm. Its axial resolution is 7 μm, and the ranging depth is 4.6 mm in tissue.

The same MGH academic group constructed an OCT system that uses a very broad bandwidth light source and common-path SD-OCT technology, termed microoptical coherence tomography (μOCT), whose resolution is improved by an order (axial resolution $\leq$1 μm and a lateral resolution $\leq$2 μm in tissue). It provides clear pictures of cellular and subcellular features (Liu *et al.*, 2012).

The Terumo (Tokyo, Japan) OFDI system uses a 1.3-μm scanning laser as a light source. The imaging range is 5 mm with an axial resolution of less than 20 μm in water.

Multiple Reference OCT for Dermoscopy

Development of the newest modalities of OCT is oriented not only on super-fast techniques or having super-high resolution, but also on making the instrument as cheap as possible. One of the examples is a multiple reference OCT (MR-OCT) system (Hogan and Wilson, 2009; Dsouza *et al.*, 2014). The idea is to use a simple along-z scanner made, for example, of a voice coil actuator (VCA). Similar actuator can be taken from a typical CD/DVD pickup head (Subhash *et al.*, 2015) whose role in the pickup head was to keep the CD/DVD track in focus at the photodetector (**Fig. 21** left).

In a prototype, the depth scanning was achieved by using a miniature recirculating optical delay based on a voice coil motor actuator and a partial mirror. The actuator was extracted from a CD optical pickup head and can provide an A-scan rate of 50–600 Hz with an axial displacement of about 60 μm. The prototype MR-OCT sensor setup (Subhash, 2014) was based on a smartphone platform using the Android operating system (**Fig. 21** right). The notations in **Fig. 21** are as follows: PUH – pickup head, SLD – superluminescent diode, L_1–L_3 – lenses 1–3, BS – beam splitter, PM – partial mirror, VCA – voice coil actuator, RM – reference mirror, D – distance between PM and RM, S – scan range, TD – total depth scan range. Recirculation is provided by adding the PM in front of the RM at a distance D, recirculating in this way the scan range S several times and extending the TD.

Devices With Photonic Arrays

Focus Tracking Micro-Lidars

For some purposes, critical is not the absolute value of the distance but the deviation from a given value. Typical example is a CD/DVD pickup head whose application in the MR-OCT system was described in the previous paragraph.

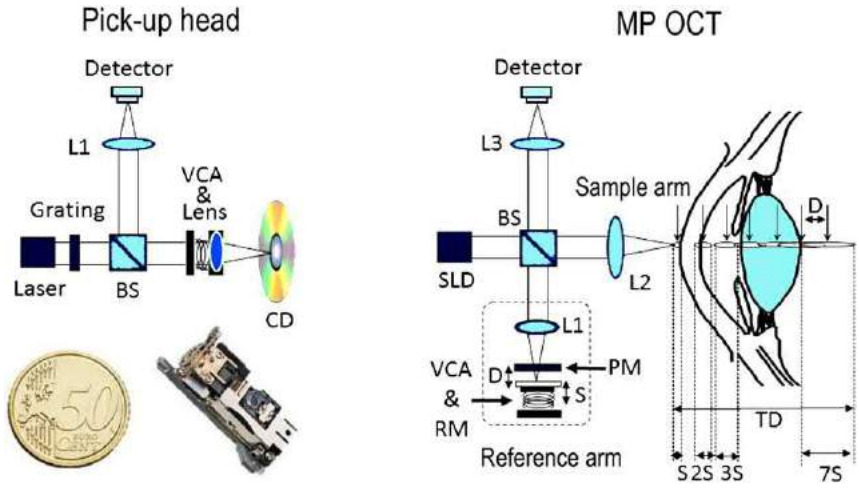


Fig. 21 Pickup head with a voice coil actuator (VCA) and the lens (at the left); experimental setup of a multiple reference-optical coherence tomography (MR-OCT) system using the same VCA with the reference mirror instead of the lens. Reproduced from Subhash, H.M., Hogan, J.N., Leahy, M.J., 2015. Multiple reference optical coherence tomography for smartphone applications. SPIE Newsroom. doi:10.1117/2.1201503.005807.

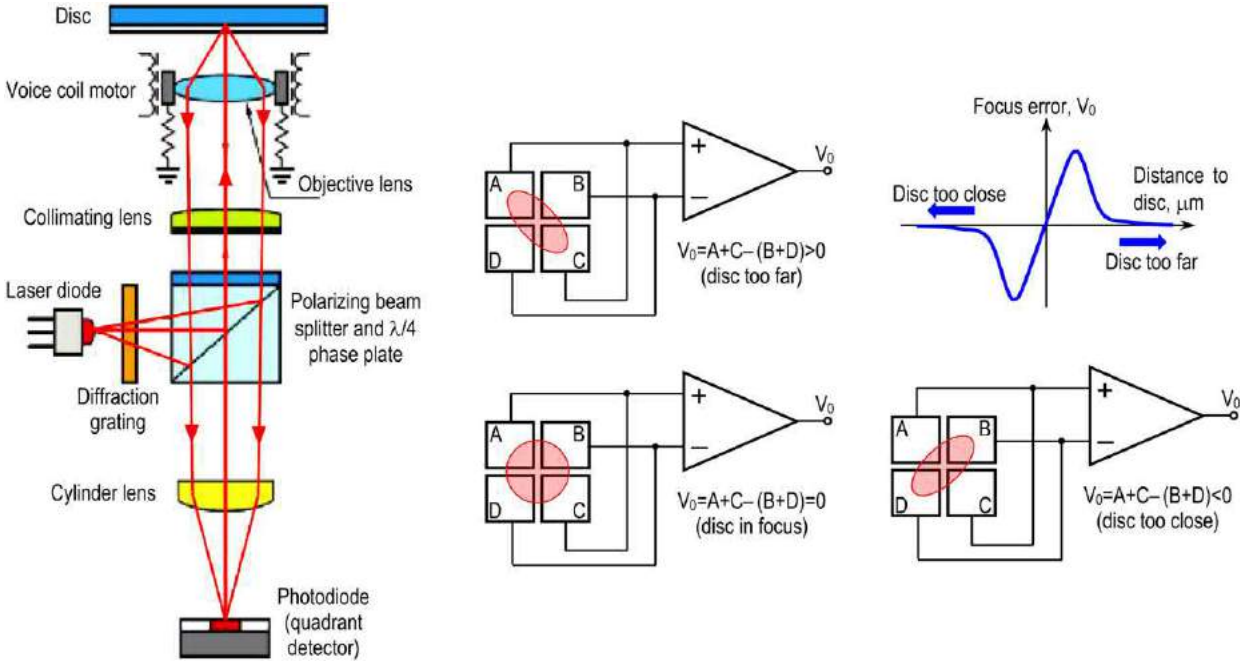


Fig. 22 Pickup head and the principle of focus tracking. Reproduced from Pohlmann, K.C., 2011. Principles of Digital Audio, sixth ed. New York, NY: McGraw-Hill.

A pickup head consists of a transmitting and of a receiving channels (Pohlmann, 2011). The transmitting channel contains a laser diode irradiating the polarized light at an infrared wavelength, a diffraction grating, a polarization beam splitter with a quarter-wave plate, a collimator lens, and an objective lens (**Fig. 22**). Laser radiation reflected from the CD passes through the objective lens, the collimating lens, the polarization beam splitter, and then is focused by a cylinder lens on the four-quadrant photodetector with four photodetectors *A*, *B*, *C*, and *D*. The quarter-wave polarization plate converts the linear polarization of the light into the circular one. On the way back, the circular polarization changes the direction of vector rotation, and after the quarter-wave plate, the linear polarization becomes orthogonal to the initial polarization. This conversion enables transporting the light to and from the disk without losses in the beam splitter.

After the cylinder lens, the beam shape at the detector array is either elliptic or circular. The circular shape occurs only if the reflective surface of the specimen is placed exactly at the focal plane of the optical path. The laser beam in this case has a diameter of less than 1 μm. The elliptic shape occurs when the beam reflection takes place out of focus. The orientation and size of the elliptic beam pattern respectively depend on the location and distance from the focal plane. The detector array has four areas,

denoted as A, B, C, and D. They deliver photocurrents, depending on the illuminance integrated over their area, which are subsequently linearly converted into voltages. From these voltages (denoted also as A, B, C, and D) the focus error V_0 can be derived as $V_0=(A+C)-(B+D)$. When $V_0=0$, the laser beam is focused on the disk, if $V_0>0$, the disk is too far, and when $V_0<0$, the disk is too close. This voltage is applied to the VCA (motor) that adjusts the distance to the position until V_0 equals 0. The mentioned diffraction grating, installed at the exit from the laser, produces a pair of diffracted beams in $+1$ and -1 orders symmetrically situated relatively to the main beam. Projected on the track in the way that the side beams are clinging to the track, one at the left side, another at the right side as shown in **Fig. 23**.

Besides the four areas A, B, C, and D in the form of quadrants, the photodetector array contains also two additional photodiodes E and F. The filtered difference TR of the signals F and E creates the tracking signal keeping the central laser spot on the track.

Range Imaging With Matricial Light-Emitting Diode Array

Reconstruction of the shape of the surface from the interferometric data has a long history, it was summarized by Creath (1988) with many examples. When a fringe pattern is recorded by a detector array, there is an output of voltages representing the average intensity incident upon the detector element over the integration time. As the relative phase between the object and reference beams is shifted, the intensities measured by point detectors will change.

Some techniques change the phase stepwise by a known amount between intensity measurements (Surrel, 1993), whereas others integrate the intensity while the phase is being shifted (Sasaki and Okazaki, 1986). The first is usually referred to as a phase-stepping technique, and the second as an integrating-bucket technique.

An example implementing the four-bucket algorithm was described by Wang *et al.* (2010). The instrument measures 3D surface profiles and is based on an light-emitting diode (LED) array phase-shift rangefinder (**Fig. 24**). It employs the fast electronic scanning of a 2D array of LED light sources instead of mechanical scanning, which means no moving or rotating parts are needed in the system. By using high-sensitivity photodiodes (PD) and high-accuracy analog-to-digital converter (ADC) to sample and demodulate the incoming light signal, a high range resolution can be obtained. **Fig. 25** illustrates four-bucket principle. A sine wave signal is sampled four times (A_0, A_1, A_2, and A_3) within a modulation period, and each sample point is delayed by a quarter

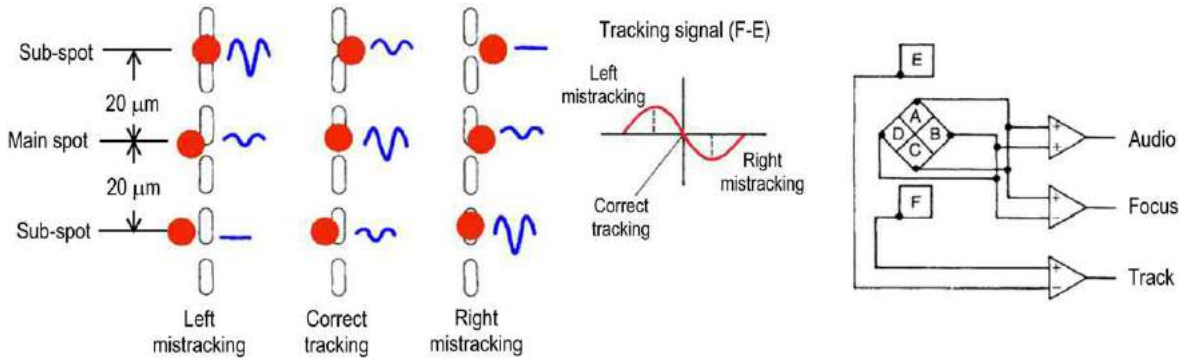


Fig. 23 Principle of keeping the central laser beam on the track. Reproduced from Pohlmann, K.C., 2011. Principles of Digital Audio, sixth ed. New York, NY: McGraw-Hill.

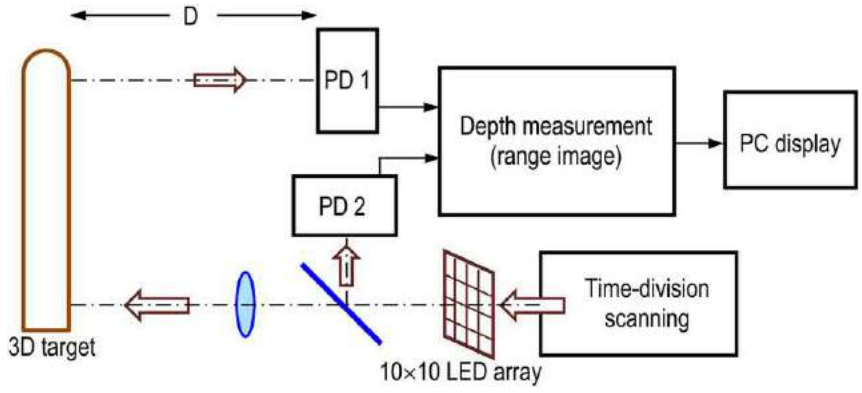


Fig. 24 Block diagram of range imaging system. Reproduced from Wang, H., Jun Xu, J., He, D., *et al.*, 2010. Real-time range imaging system based on a light-emitting diode array phase-shift range finder for fast three-dimensional shape acquisition. Optical Engineering 49 (7), 073201.

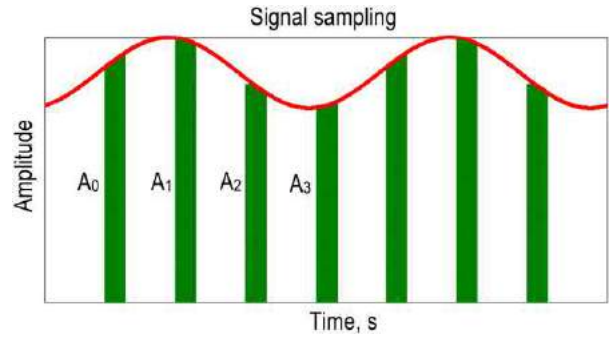


Fig. 25 Principle of phase measurement using four-bucket algorithm. Reproduced from Wang, H., Jun Xu, J., He, D., *et al.*, 2010. Real-time range imaging system based on a light-emitting diode array phase-shift range finder for fast three-dimensional shape acquisition. Optical Engineering 49 (7), 073201.

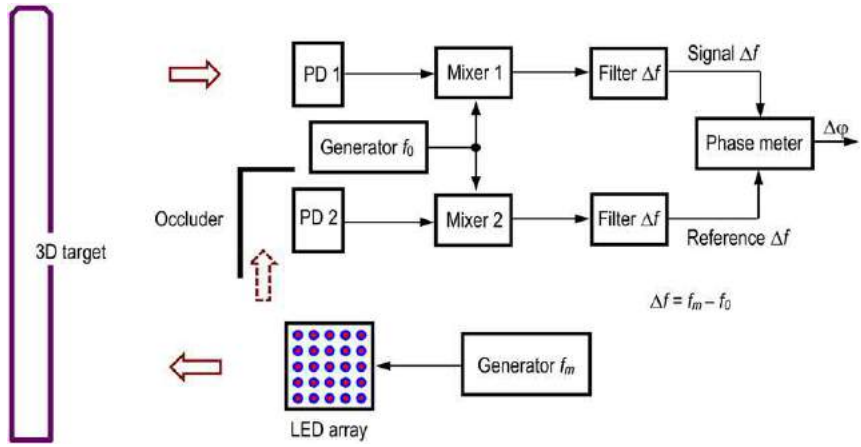


Fig. 26 Block diagram of phase-shift rangefinder using heterodyne technique. Reproduced from Wang, H., Jun Xu, J., He, D., *et al.*, 2010. Real-time range imaging system based on a light-emitting diode array phase-shift range finder for fast three-dimensional shape acquisition. Optical Engineering 49 (7), 073201.

period of the modulation frequency. From this sampling, based on Nyquist theorem, the phase delay of the signal can be determined and recalculated into the measured distance.

Fig. 26 shows the block diagram of the phase-shift rangefinder using a heterodyne technique. Since the modulated LEDs have temperature drift, two receiving channels are used in the system, of which one serves as a reference channel. The signals from the outputs of two mixers are filtered by band-pass circuits tuned on f_1 with a bandwidth Δf_1. Then, the two signals are sent to the phasemeter, where the phase measurement is implemented.

The instrument is insensitive to ambient light, since the dc-offset of the received signal is removed. It is no problem to use a large pixel array size and high spatial resolution by utilizing a low-cost close packaged LEDs. The system can be very robust and can be used to perform inexpensive and real-time range scanning for many robotic and industrial automation applications, including those in hazardous environmental conditions. By combining with the mature and low-cost 2D charge-coupled device/complementary metal-oxide semiconductor (CCD/CMOS) imaging technique, it also can achieve real-time, high-quality 3D imaging, which is another important application prospect for this range imaging system. The authors found that the range images can be acquired at a rate of 10 frames/s with a depth resolution better than ± 5 mm in the range of 50–1000 mm.

Time-of-Flight Principle With Matricial Detectors

To get the information in the 3D space, not only the matrices of emitters can be used, but also the matrices of photodetectors as well (Mutto *et al.*, 2012). Such micro-lidar (sensor) may be conceptually interpreted as a set of a multitude of single devices. It does not mean that the matrix of emitters should be manufactured in a single chip with the matrix of receivers. It only means that a single emitter may provide an irradiation that is reflected back by the scene and collected by a multitude of receivers close to each other.

Once the receivers are separated from the emitters, the former can be implemented as CCD/CMOS pixels integrated in a matrix. The lock-in pixels matrix is commonly called time-of-flight (ToF) camera sensor, and, for example, in the case of the MESA

Imaging SR4000 it is made by 176×144 lock-in pixels (see Section Relevant Websites). Such matricial ToF sensor IR emitters are common LEDs and they can be positioned in a configuration mimicking a single emitter co-positioned with the center of the receivers matrix, as shown in **Fig. 27**. Indeed the sum of all the IR signals emitted by this configuration can be considered as a spherical wave emitted by a single emitter. **Fig. 28** shows the actual emitters distribution of the MESA Imaging SR4000.

The operation of a ToF camera as imaging system can be summarized as follows. Each ToF camera sensor pixel, at each period of the modulation sinusoid, collects four samples of the IR signal reflected by the scene. This information is enough to reconstruct the depth values for each pixel and to build the map of distances in the scene. Accurate description of this kind of ToF camera is given by Lange (2000) in his PhD work. Hansard *et al.* (2012) and DAGM Workshop Proceedings (Kolb and Koch, 2009) can be recommended for further reading.

Pattern-Projecting Matricial Sensors

In 2012, the Microsoft Kinect version of the play station was released which included the depth mapping with a light-coded range camera capable to estimate the 3D geometry of the acquired scene at 30 fps with VGA (640×480) spatial resolution (**Fig. 29**). From the functional point of view, the Kinect range camera is similar to the ToF camera described previously, since they both estimate the 3D geometry of dynamic scenes, are similar also in that they use matricial detectors, but they differ in the principles: the Kinect technology is based on pattern-projection analysis, several patents were issued to PrimeSense (now Apple) (Freedman *et al.*, 2012; Sali and Avraham, 2014; Cohen *et al.*, 2016). Some commenting authors (Mutto *et al.*, 2012) call this technique "matricial active triangulation." Light patterns can be generated by a light modulator as random distributions of light and dark spots. It can be also a diffraction or speckle pattern. García *et al.* (2008) projected coherent light through ground glass generating

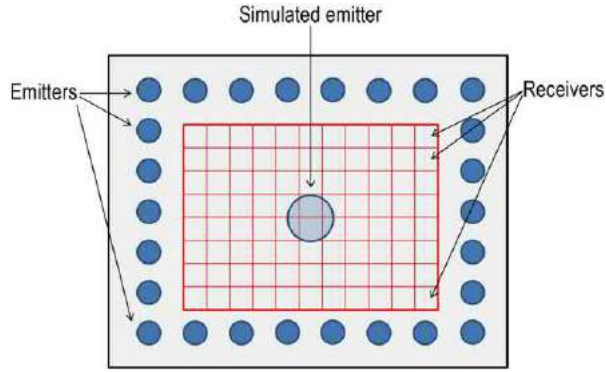


Fig. 27 Scheme of a matricial time-of-flight (ToF) camera sensor. The emitters are distributed around the matrix. Reproduced from Heptagon. SR4000/SR4500 User manual. Available at: http://mesa-imaging.ch.

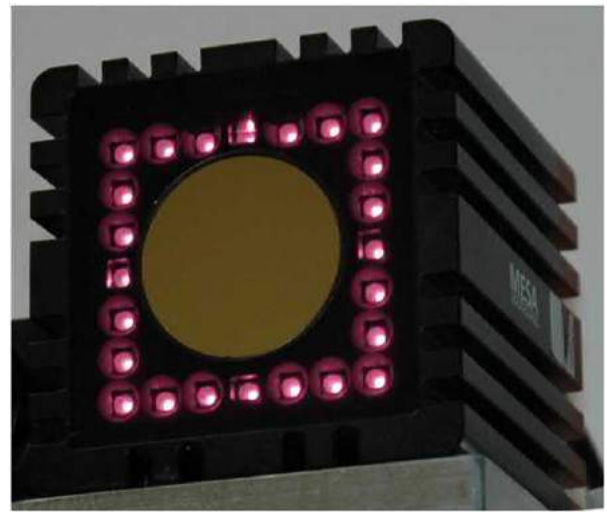


Fig. 28 The emitters of the MESA Imaging SR4000 are the red light emitting diodes (LEDs). Reproduced from Heptagon. SR4000/SR4500 User manual. Available at: http://mesa-imaging.ch.

Fig. 29 Microsoft Kinect play station. *Source*: Internet

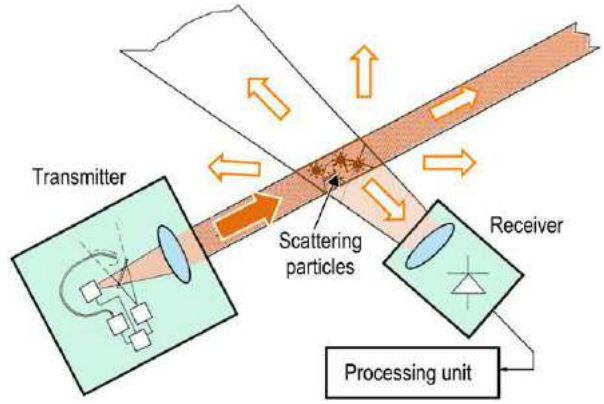


Fig. 30 Typical configuration of the lidar measuring the scatterance of the media. Reproduced from Lonnqvist, J., 1999. Apparatus and Method for Measuring Visibility and Present Weather. US Patent 5,880,836.

random speckle patterns. The spatially random patterns are seen by the sensor. Low correlation between different patterns is used for both 3D mapping of objects and range finding.

Asus X-tion Pro and X-tion Pro Live (see Section Relevant Websites) are other products with a range camera based on the same chip initially manufactured by PrimeSense. All these range cameras, support an IR video-camera and projecting the IR light coded patterns. In this technology, each pixel needs to be associated to a code-word, i.e., a specific local configuration of the projected pattern. A correspondence estimation algorithm analyzes the received code-words in the acquired images in order to compute the conjugate of each pixel of the projected pattern.

Further modifications of the pattern-projection technology are proposed involving combinations of cameras (Sali and Avraham, 2016), or adding a laser illuminator for TOF range measurement (Devaux *et al.*, 2013; Nehmadi and Guterman, 2016).

Micro-Lidars Measuring the Media Parameters

Nephelometers

The typical lidar configuration is shown in **Fig. 30** which is the same for hundred-kilometers space lidars and centimeter- and even micrometer-scale micro-lidars (Lonnqvist, 1999). The angle θ between the direction of the laser beam and the optical axis of the receiver can vary from 0 (or near zero) to 180 degrees. The transmitter and the receiver can use a common optics (monostatic configuration) or, they can use separate optical systems (bistatic configuration). The second one is often preferable to exclude the scatter of the transmitted laser light in the optical system. The volume, from where the information is got, is limited in the first approximation by the section of the laser beam encircled by the receiver's field of view.

Jerlov (1976) considers scattering in the concepts of large-particle optics as the result of three physical phenomena: (1) through the action of the particle, light is deviated from rectilinear propagation (diffraction); (2) light penetrates the particle and emerges with or without one or more internal reflections (refraction); (3) light is reflected externally. The amount of scatter (scatterance β) depends on the angle θ. The function $\beta(\theta)$ is called indicatrix. The indicatrix describes the amount of light that is scattered at every possible combination of angles of incoming and outgoing light (influx and efflux angles). The character of scatter depends on how polydispersed system is. Monodispese sytem has oscillatory β curve. The most striking and distinctive feature of a polydispersed system is the pronounced forward scattering. **Fig. 31** represents the function $\beta(\theta)$ of the oceanic water at different depths. It is clearly seen that at small depths the indicatrix has more oblong shape in the forward direction. Deeper oceanic waters are clearer and their indicatrix is nearer to that of the pure water.

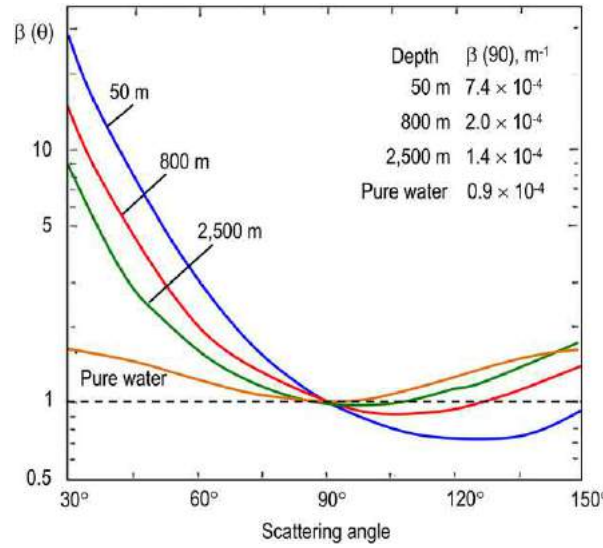


Fig. 31 Indicatrices of the oceanic water at different depths (normalized to β at 90 degree). Reproduced from Jerlov, N.G., 1976. Marine Optics. Amsterdam: Elsevier.

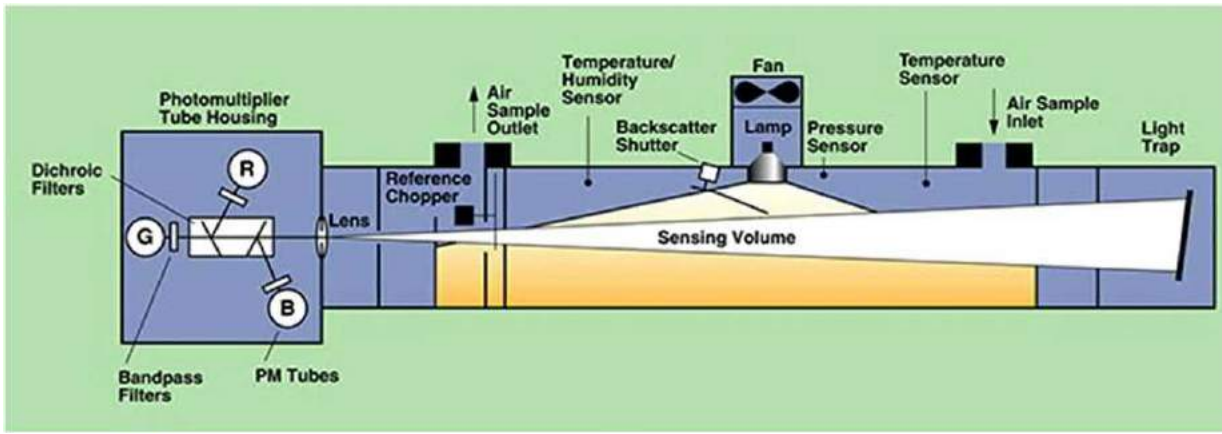


Fig. 32 TSI 3563 nephelometer with additional temperature, humidity, and pressure sensors. Reproduced from Mizrahi, A., Russell, L., Three wavelength integrating nephelometer. Available at: https://www.esrl.noaa.gov/gmd/aero/instrumentation/neph_desc.html.

The instrument for measurement of scatterance (scatterance meter) is called nephelometer. Jerlov (1976) described the nephelometer for measurement of the function $\beta(\theta)$. Angular dependence of scatterance is not always required. In some case, it is enough to have only value of scatter at a certain angle, for example, for turbdity measurement (Mitchell, 2010). For pollution measurement, or for determination of the complex refractive indices of aerosol, use only an integrated value of scatter (Han *et al.*, 2009). At the same time, additional information can be derived when measuring the scatter at different wavelength (Anderson *et al.*, 1996). The three-wavelength model TSI 3563 (**Fig. 32**) splits the scattered light into red (700 nm), green (550 nm), and blue (450 nm) wavelengths. The TSI nephelometer measures back-scattered light at these wavelengths as well. The one wavelength Radiance Research nephelometer measures forward scattering at 550 nm only.

The main body of the TSI 3563 nephelometer is a 10-cm-diameter aluminum tube, 90 cm long (Mizrahi and Russell). Along the axis is an 8-cm-diameter tube set with aperture plates. The plates range from 7 to 170 degrees on the horizontal range of the lamp. The backscatter shutter allows blocking of the angles from 7 to 90 degrees so that only backscattering is measured. The light trap provides a dark reference against which to measure the scattered light.

The receiving optics is located on the other side of the tube from the trap. The light that passes through the lens is separated by dichroic filters into three wavelengths. The first is a color splitter that passes 500–800 nm light and reflects 400–500 nm light. The reflected light passes through a filter centered at 450 nm (blue) into a photomultiplier tube (PMT). The light that passes through the first splitter goes to a second splitter which passes 500–600 nm light and reflects 600–800 nm light. The reflected light passes through a filter centered at 700 nm (red) into a second PMT. The light that passes through both splitters passes through a filter centered at 550 nm (green) into a third PMT. For further reading we recommend the review of integrating nephelometers of Heintzenberg and Charlson (1996).

Smoke Detectors

Lidar principles are widely used in office and home appliances. A smoke detector is an example. Dohl (2015) patented the smoke detector with a plurality of light emitting devices of different wavelengths and with another plurality of scattered light receiving sections. **Fig. 33** demonstrates the design of the smoke detector with one such section in which the optical axes of light emitting diode and of photodetector meet in the scattering volume outside the detector. There can be several LEDs in one transmitting position (**Fig. 34**), for example, blue and infrared. Such coupling of the light wavelengths enables the differentiation of the type of the smoke: if the signal is higher in the longer wavelength channel, the device decides that the larger particles are contained in the smoke, and vice versa: if the signal is higher in the shorter wavelength channel, the instrument decides on the prevailing number of smaller particles. Positioning the receivers at different angles relatively to the optical axis of the light emitting channel can give the information on the complexity of the scattering particles: in the suggestion that the smoke particles have more complicated shape than the particles of the vapor, the device can suggest whether the scattering particles are of the smoke origin or, they are the water vaporization particles.

Flow Cytometry

When analyzing the smoke detector, we noted that the complexity of particle shape results in higher scatter in the lateral direction relatively to the laser beam. This feature of particle scatter is used in flow cytometry that is performed on particles in suspension (e.g., cells or nuclei). The cell suspension is injected into a flow chamber of the fluidic system, where cells are forced into single file and directed into the path of a laser beam (Craig, 2014). Laser light is focused on these single particles/cells and can be measured

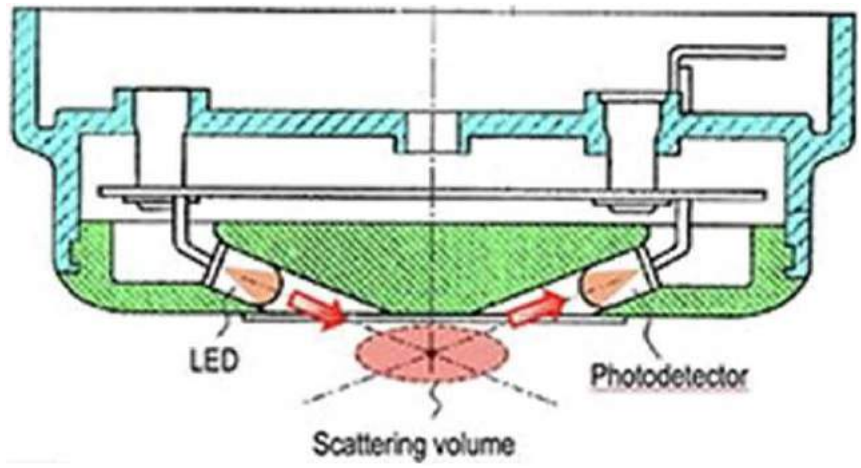


Fig. 33 One channel of the smoke detector. Reproduced from Dohl, M., 2015. Smoke Detector. US Patent 8,941,505. January 27, 2015.

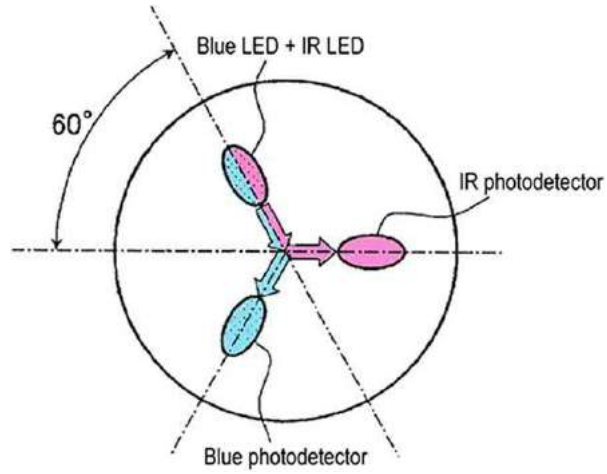


Fig. 34 Two-channel positioning in the multichannel smoke detector. Reproduced from Dohl, M., 2015. Smoke Detector. US Patent 8,941,505. January 27, 2015.

by photodetectors as it is scattered off particles/cells when they pass in front of the beam. Fig. 35 demonstrates the principle of the Sysmex XT-2000i and XT-1800i Automated Hematology Analyzers.

When the laser beam interacts with a single particle/cell, light is scattered but its wavelength is not altered (elastic scattering). The amount of scattered light can be used to identify the particle/cell because it is related to the physical properties of the particle (size, granularity, and nuclear complexity). Light scattered at a 90 degree angle (side light scatter) is related to the internal complexity and granularity of the particle/cell. Neutrophils produce much side scatter because of their numerous cytoplasmic granules. Light that proceeds in a forward direction (forward light scatter) is related to the particle size. Large cells produce more forward scatter than small cells. As illustrated in Fig. 36, differences in light scattering are used to distinguish the particles: the

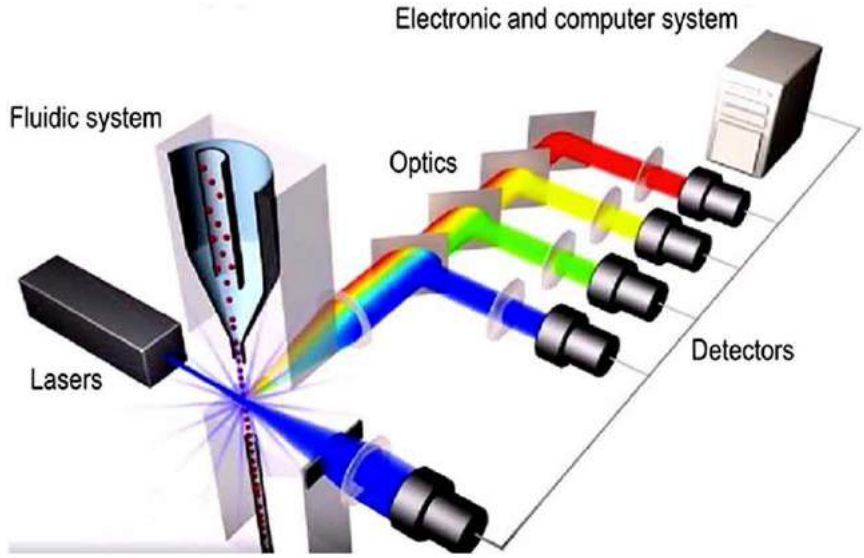


Fig. 35 Configuration of the Sysmex XT-2000i and XT-1800i automated hematology analyzers. Reproduced from Craig, F.E., 2014. Flow cytometry. In: McKenzie, S.B. (Ed.), Clinical Laboratory Hematology, second ed. Essex: Pearson, pp. 956–974.

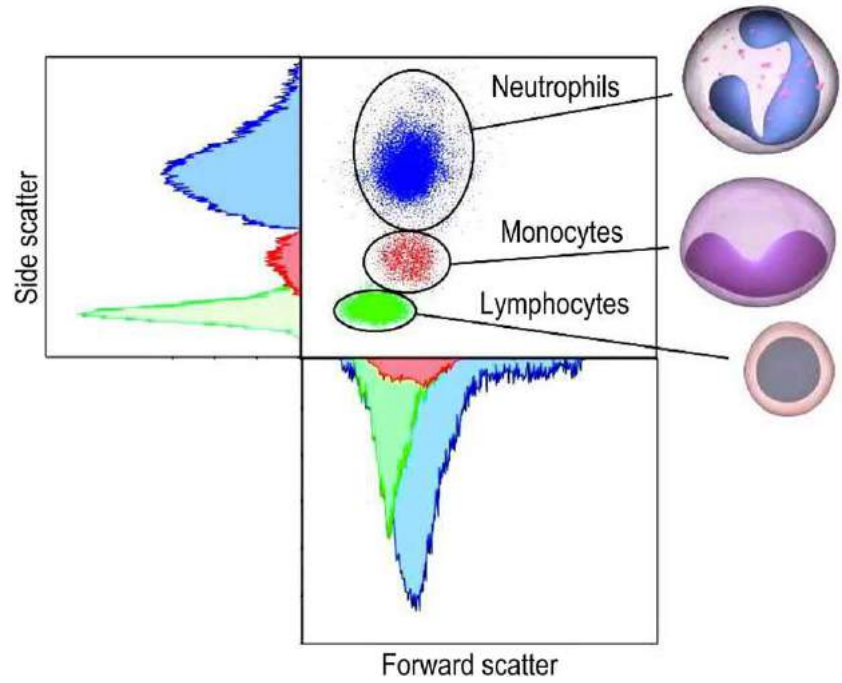


Fig. 36 Distinguishing the particles. Reproduced from Craig, F.E., 2014. Flow cytometry. In: McKenzie, S.B. (Ed.), Clinical Laboratory Hematology, second ed. Essex: Pearson, pp. 956–974.

lymphocytes as the smallest ones are in the low side-signal domain, and the neutrophils, being the biggest ones having the most complicated shape are in the high side-signal domain.

In addition to elastic light scattering, flow cytometers are used to detect bound fluorescent markers (fluorochromes), which are molecules that are excited by light of one wavelength and emit light of a different wavelength (fluorescent light). Fluorochromes can label the antibodies or specific changes in DNA or RNA (**Figs. 37** and **38**). Flow cytometers use light of a single wavelength generated by a laser to excite fluorochromes bound to the particle of interest. Light emitted from the fluorochrome is separated from the incident laser light using a combination of filters and mirrors. A PMT then detects and quantifies the emitted light.

Clinical flow cytometers usually contain an argon laser that generates light at 488 nm. This single wavelength is often used to excite three different fluorochromes, each emitting light at different wavelengths. Using three different fluorochromes allows detection of three different antigens on the cell. Excitation of additional fluorochromes to detect further antigens usually requires a second laser light source (e.g., helium neon, emission 633 nm). Two laser light sources allow to identify up to six antigens on each cell analyzed.

Raman Micro-Lidars

In an inelastic scattering of a photon, some of its energy is lost or increased. Correspondingly, the frequency of the photon is shifted to red or blue. A red shift means that part of the photon energy is transferred to the interacting matter (Stokes Raman scattering). In the blue shift, internal energy of the matter is transferred to the photon (anti-Stokes Raman scattering). Detecting the Stokes or anti-Stokes lines in the scattered spectrum is related to the spectroscopy. Still, in practice, when applied to the techniques of studying the properties of atmosphere, the terms "lidar," "Raman lidar" are used. When coming to the microworld, we may use the term "Raman micro-lidar." Detecting the inelastic scattering and its usage for measuring the parameters of the medium has a wide field of applications. We recommend for further reading the book edited by Baudelet (2013).

The role of the Raman micro-lidar is to make an "optical biopsy," i.e., to diagnose disease such as cancer and atherosclerosis without removing tissue from the body in the way by revealing the differences between cancerous and normal tissues of various organs due to morphological and molecular changes in the tissue. We shall refer here to key optical methods of Raman inelastic

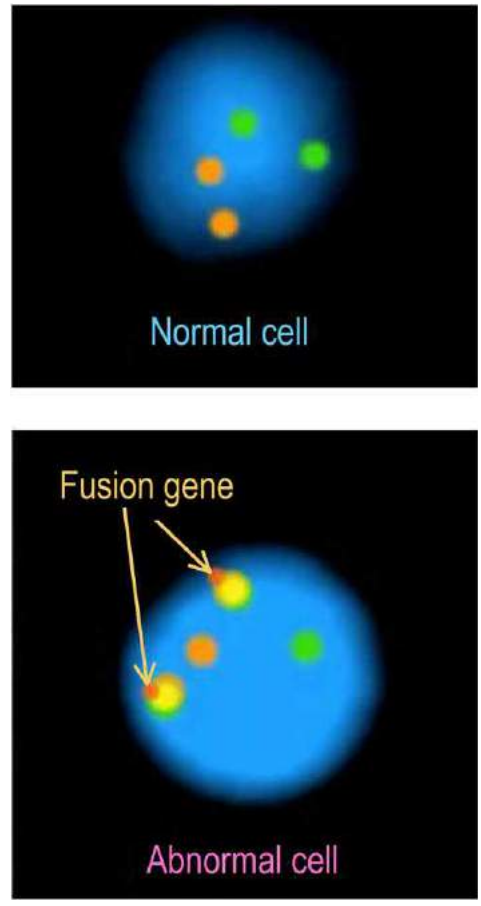


Fig. 37 Normal cell and the abnormal cell with fused genes. Reproduced from Craig, F.E., 2014. Flow cytometry. In: McKenzie, S.B. (Ed.), Clinical Laboratory Hematology, second ed. Essex: Pearson, pp. 956–974.

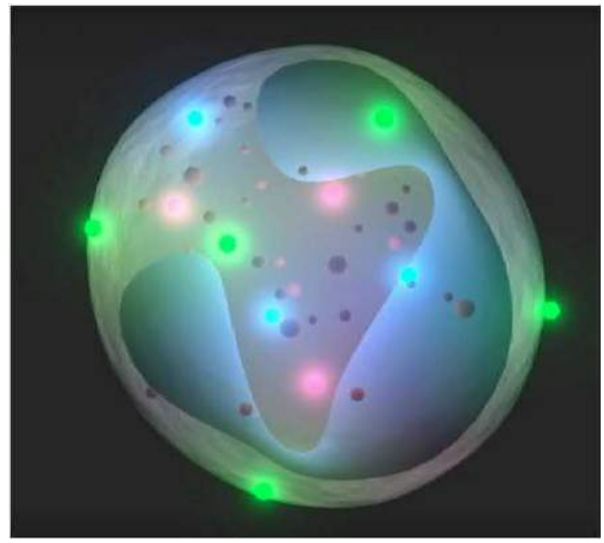


Fig. 38 Neutrophil with fluorescent markers. Reproduced from Craig, F.E., 2014. Flow cytometry. In: McKenzie, S.B. (Ed.), Clinical Laboratory Hematology, second ed. Essex: Pearson, pp. 956–974.

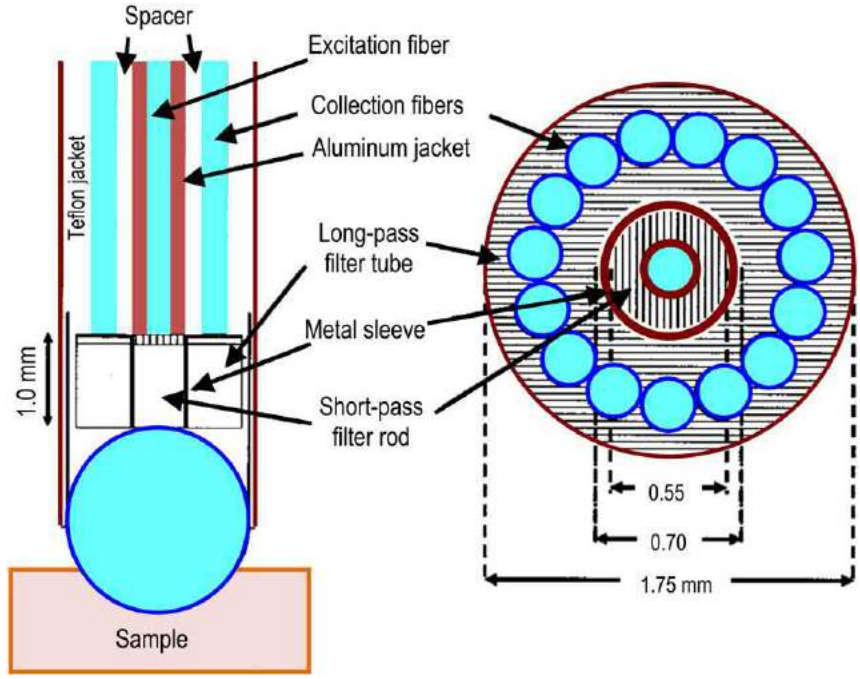


Fig. 39 Schematic of the Raman probe tip. Ball lens B is in contact with the filter module that couples to the fiber bundle. Reproduced from Motz, J.T., Hunter, M., Galindo, L.H., *et al.,* 2004. Optical fiber probe for biomedical Raman spectroscopy. Applied Optics 43, 542–554.

backscattering reflectance and time-resolved spectroscopy. Various human tissue types (prostate, breast, lung, colon, and gastrointestinal) have been studied using optical biopsy (Jelínková, 2013).

The first work on using the Raman scattering to detect cancer was published by Alfano *et al.* (1991). Alfano and Pu (2013) summarized the achievements in the mentioned book of Jelínková. The peak height ratios for Raman data analysis were able to differentiate normal tissue and malignant tumor in breast, gynecological, and cervical tissue. These basis spectra represent the epithelial cell cytoplasm, cell nucleus, fat, β-carotene, collagen, calcium hydroxyapatite, calcium oxalate dihydrate, cholesterol-like lipid deposits, and water.

Motz *et al.* (2004) and Haka *et al.* (2006), both from MIT, developed an optical fiber Raman probe and collected Raman spectra of breast tissue. Their portable Raman system with optical fiber probe is schematically shown in **Figs. 39** and **40**. An 830-nm-diode laser is focused into the excitation fiber of the Raman probe through a bandpass filter. Consisting of a single central excitation fiber

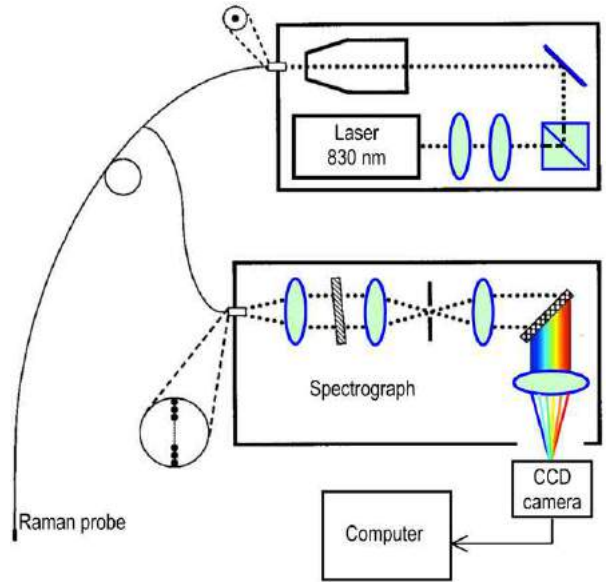


Fig. 40 Schematic of the Raman spectroscopy system used for experimental testing of the Raman probe tip. Reproduced from Haka, A.S., Volynskaya, Z., Gardecki, J.A., *et al.*, 2006. In vivo margin assessment during partial mastectomy breast surgery using Raman spectroscopy. Cancer Research 66 (6), 3317–3322.

surrounded by 15 collection fibers of 200 μm core diameter, the diameter of the probe is less than 2 mm. The linear array of collection fibers at the proximal end is coupled to a spectrograph.

Light delivery and collection can be accomplished by using optical fibers that can be incorporated into a biopsy needle (Haka *et al.*, 2005). As opposed to biopsy, a spectroscopic needle measurement has the advantage of providing immediate diagnosis. With the development of minimally invasive breast cancer therapies, such as radiofrequency ablation (Fornage *et al.*, 2004), there is the potential for diagnosis and treatment to be performed in a single procedure.

Working on even less invasive instrument, Komachi *et al.* (2005, 2009) designed a micro-Raman probe with a 600 μm diameter, that can be inserted into coronary arteries. Custom-made filters attached to the front end of a probe eliminate the background Raman signals of the optical fiber itself. Measurement of the Raman spectra of an atherosclerotic lesion of a rabbit artery in vitro demonstrated its excellent performance.

Further studies resulted in a Raman spectroscopy technique to simultaneously identify micro-calcification status and to diagnose the underlying breast lesion, in real time, during stereotactic core needle biopsy procedures (Barman *et al.*, 2013). An approach was introduced based on diffuse reflectance spectroscopy for detection of micro-calcifications that focuses on variations in optical absorption stemming from the calcified clusters and the associated cross-linking molecules. A new powerful decision algorithm was derived from correlating the diffuse reflectance spectra with the corresponding radiographic and histologic assessment.

Time-Resolved Fluorescence Spectroscopy

Fluorescence spectroscopy is one of the optical techniques used to investigate biomolecules whose photoexcited emission is characterized by spectral distribution, photon yield, lifetime of the excited state, as well as polarization. The first measurements were all performed by methods that involved the human eye. Significant advances in instrumentation, particularly in regard to time-resolved measurements, were achieved with the advent of lasers. Time-resolved fluorescence spectroscopy derived short-pulse lidar principles and can provide information on not only the location and intensity of the key biomolecules but also their local biophysical microenvironments (Alfano and Pu, 2013). It also provides temporal information on the underlying dynamics of molecular processes concerning the chromophores and the environment responsible for the fluorescence in tissues. Picosecond techniques enable the study of dynamics of relaxation in macromolecules.

The experimental arrangement for the time-resolved fluorescence measurements of the human breast tissues is schematically shown in **Fig. 41**. Ultrafast laser pulses of 100 fs duration, 0.1 nJ per pulse, at 620 ± 7 nm from a colliding pulse mode-locked dye laser system at a repetition rate of 82 MHz is used to pump the samples. These laser pulses are amplified and their frequency is doubled in KDP crystal resulting in 310 nm wavelength. The resulting fluorescence emission is collected and directed onto the slit of a synchro-scan streak camera with a temporal resolution of 16 ps. A narrow band filter of 340 ± 5 nm is used to collect the tissue fluorescence and a 310 nm notch filter – to cut off the excitation wavelength. The recorded temporal profiles are analyzed to obtain temporal and polarization information.

Typical time-resolved fluorescence profiles excited by 310 nm from malignant tumor and from normal tissue at 340 nm emission band are shown in the inset of **Fig. 41**. There is a marked difference between the curve profiles of malignant and

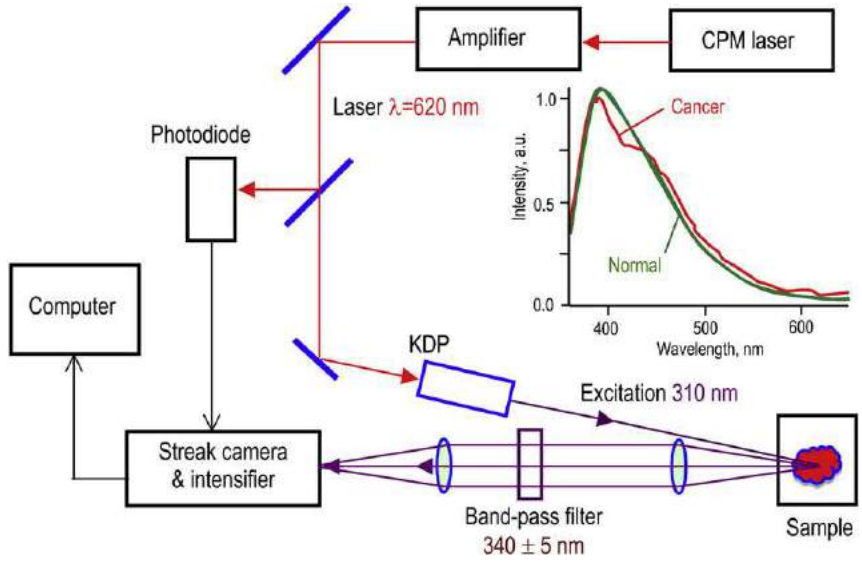


Fig. 41 Time-resolved fluorescence experimental setup and fluorescence profiles with 310 nm excitation and 340 nm emission: malignant tumor (red), normal breast tissue (blue). Reproduced from Alfano, R., Pu, Y., 2013. Optical biopsy for cancer detection. In: Jelínková, H. (Ed.), Lasers for Medical Applications, Cambridge: Woodhead, pp. 325–367.

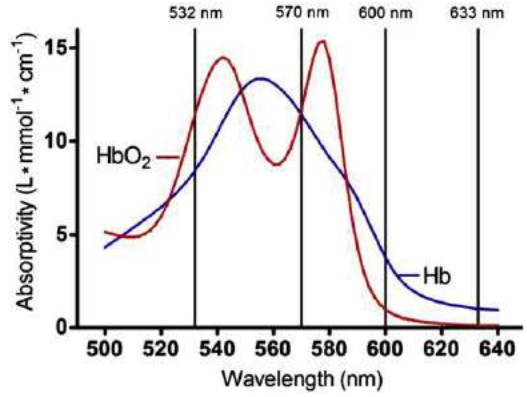


Fig. 42 Absorptivity of oxygenated and deoxygenated hemoglobin. Reproduced from Delori, F.C., 1988. Noninvasive techniques for oximetry of blood in retinal vessels. Applied Optics 27, 1113–1125.

nonmalignant sample. All temporal curves are analyzed to compare the slow and fast components of fluorescence decay. Also, the amplitude ratios of fast to slow components are analyzed to find the differences.

Oxymetry Lidars

The degree of blood oxygen saturation is one of the diagnostic parameters. Different absorptivity of light propagating in the blood with oxygenated and deoxygenated hemoglobin depending on the light wavelength (**Fig. 42**) is the basic for pulse oximetry. The technique of pulse oximetry using the cost-effective and patient-friendly sources of light is described in the book of Moyle (2002).

The oxygen supply of the retina is no less important for the human vision. The supply is provided by both the choroidal and retinal circulation. The choroid serves as the oxygen source for the photoreceptors in the avascular outer retina, whereas the retinal circulation plays a critical role, mediated by its autoregulatory capacity, in maintaining the oxygen supply to the neural elements and nerve fibers in the inner retina. Because of the high oxygen needs of the retina (cerebral tissue), any alteration in circulation such as seen in diabetic retinopathy or vascular occlusive diseases results in functional impairment and extensive retinal tissue damage.

Measuring oxygenation of blood could help to assess the circulatory and respiratory condition of patients. The retina is highly metabolically active tissue and information on retinal and choroidal oxygenation would give vital clinical information on the metabolic state of the retina, which may improve the understanding of retinal metabolism in health and disease. Malfunction of the retinal vasculature can result in serious eye diseases, which can affect vision considerably, for example, central retinal arterial occlusion or central retinal vein occlusion.

But how to get information on oxygenation of blood vessels from the bottom of the eye? The principle of pulse oximetry cannot be applied directly since it operates in the transmissive mode. Hickham *et al.* (1963) proposed to measure blood oxygenation in retinal vessels crossing the optic disk by using a two-wavelength (510 nm and 640 nm) photographic technique. This was the first time the oxygenation of the human retina was measured noninvasively, and following these studies many techniques have been developed. Laing *et al.* (1975) modified the instrument having used two other wavelengths (470 nm and 515 nm).

These methods involved laborious microdensitometric analysis of photographic negatives obtained under controlled conditions. To take into account the light scattered by the red blood cells (RBCs) in whole blood, the introduction of a calibration procedure was needed. To overcome the problem, Pittman and Duling (1975) introduced three closely spaced wavelengths. The contribution of scattering to the optical density at each wavelength was determined from optical density values at two isosbestic wavelengths (546 and 520 nm) and the optical density at the third wavelength (555 nm).

Delori *et al.* (1983) implemented this idea using a fundus camera adding there a photoelectric system with control and processing electronics (**Fig. 43**). Actually, it was the transformation of the fundus camera into a micro-lidar (Delori, 1988).

Most of the knowledge on retinal oxygenation comes from animal studies. Hardarson (2012) describes in his PhD thesis the retinal oximeter Oxymap designed at the University of Iceland which is based on a fundus camera to which an image splitter is attached to simultaneously capture four images of the same area of the fundus (**Fig. 44**). The images are captured at 542 nm, 558 nm, 586 nm, and 605 nm. All analyses were performed with the 586 nm (isosbestic) and the 605 nm (non-isosbestic) images. The 542 nm and 558 nm images were not used. Discarding of 542 nm and 558 nm was based on earlier testing, which revealed that optical density ratios, using these wavelengths gave less reliable results.

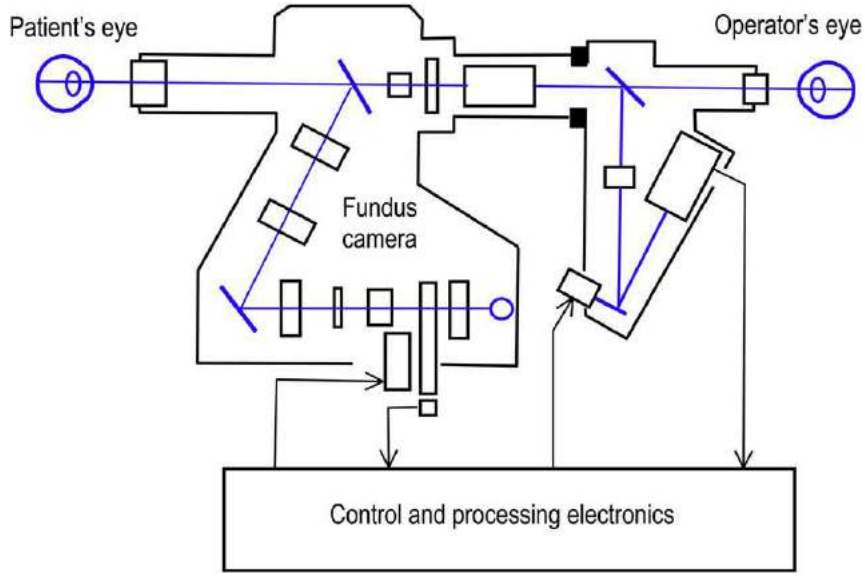


Fig. 43 Fundus camera with incorporated photoelectronic system. Reproduced from Delori, F.C., 1988. Noninvasive techniques for oximetry of blood in retinal vessels. Applied Optics 27, 1113–1125.

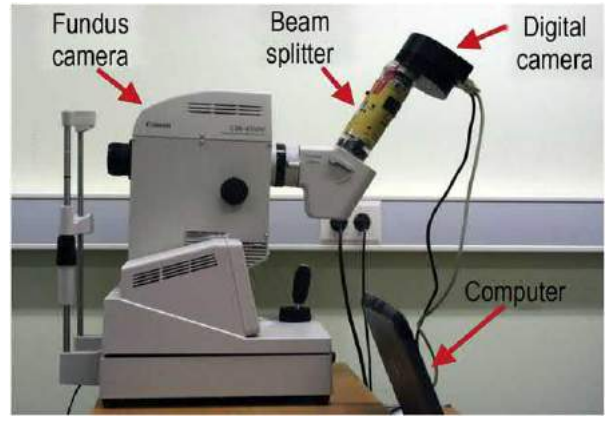


Fig. 44 Retinal oximeter on the platform of Canon CR 6-45NM. Reproduced from Hardarson, S.H., 2012. Retinal oximetry. PhD Thesis, University of Iceland, Reykjavik.

Kristjansdottir (2014) described the Oxymap T1 instrument developed at the University of Iceland (Fig. 45 left) based on the Topcon TRC-50DX mydriatic retinal camera. Two highly sensitive digital cameras, image splitter, optical adapter, and light filters are added. A fundus image captured by camera is split into two by the image splitter. One camera captures the image at 570 nm while the other captures the same area of the fundus simultaneously, at 600 nm (Fig. 46). This is done by inserting two narrow 5 nm band pass filters into the light path to each camera of the oximeter. Another light filter (80 nm band pass filter, 585 nm

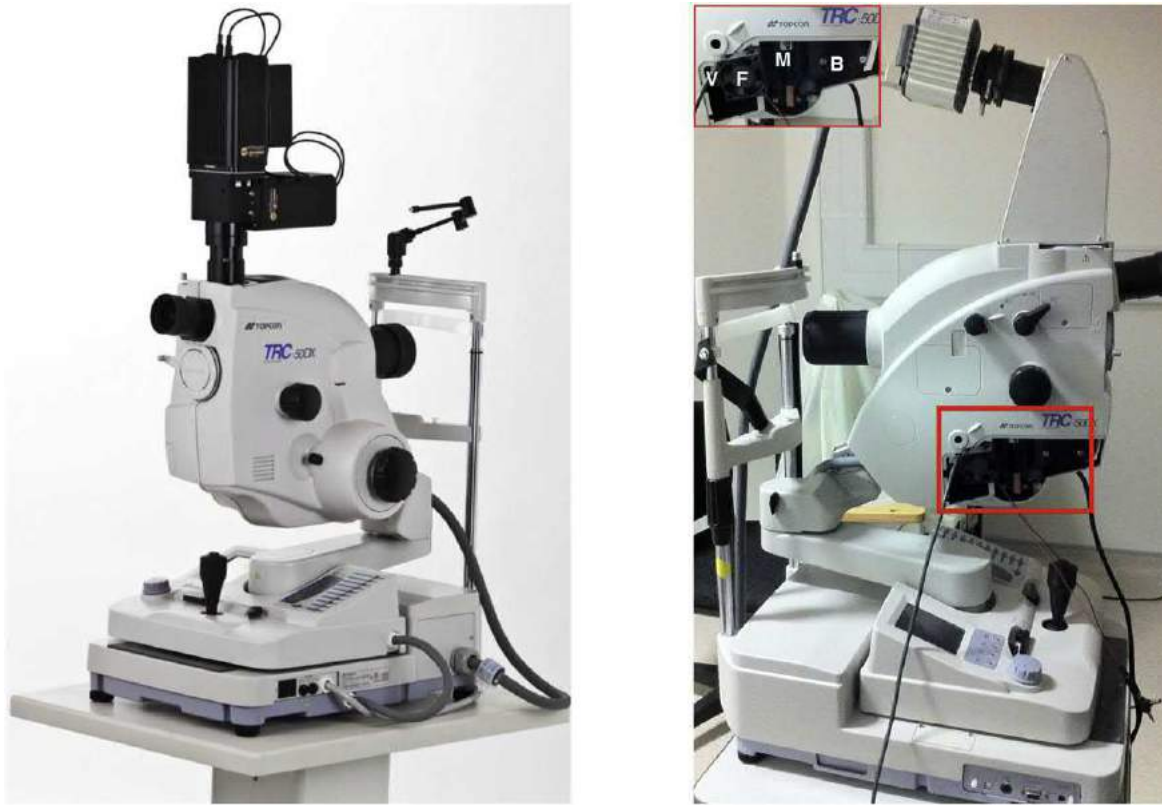


Fig. 45 Two variants of redesigning the Topcon TRC-50DX mydriatic retinal camera: left – (Kristjansdottir, 2014); right – white light was emitted by the standard 100W bulb (B), which was filtered to select a single wavelength of light by an optical filter. Reproduced from Holm, S.P., 2014. Optical imaging of retinal blood flow: studies in automatic vessel extraction, alignment, and driven changes in vessel oxymetry. PhD Thesis, University of Manchester.

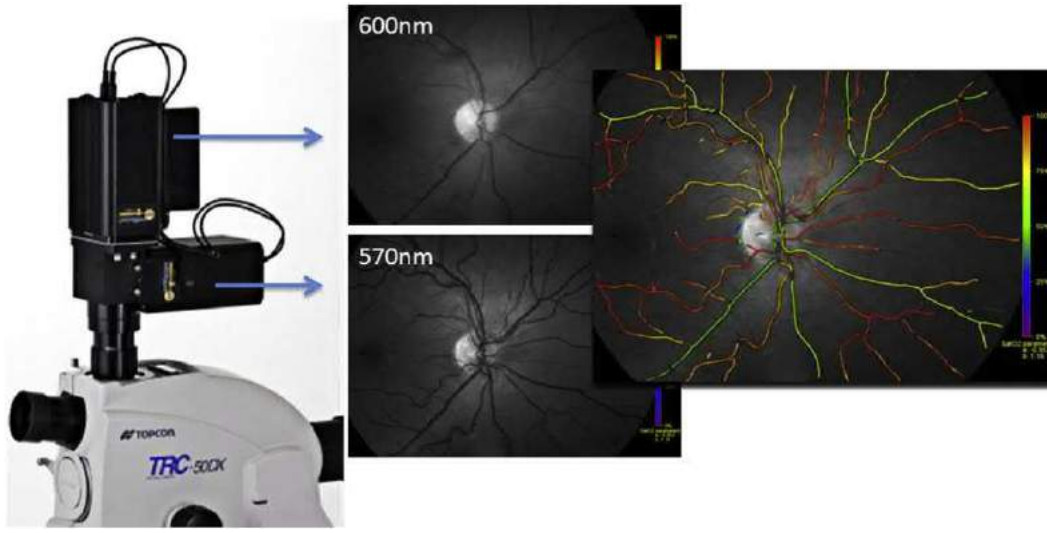


Fig. 46 Reconstruction of the retinal oxygenation map with the Oxymap T1 instrument. Reproduced from Kristjansdottir, J.V., 2014. Choroidal and retinal oximetry. MSc Thesis, University of Iceland Reykjavik.

center wavelength) is also inserted into the fundus camera itself to exclude unnecessary light exposure to subjects' eyes, only allowing light between 545 and 625 nm to exit the camera lens.

A specialized software, Oxymap Analyzer, was developed for calculation of blood vessel oxygen saturation. It analyses two spectral images (570 and 600 nm, the isosbestic is 570 nm wavelength), automatically detects blood vessels, selects measurement points and measures the brightness on vessels at each wavelength. From their ratio, the optical density is calculated. Examples of oxygenation maps of healthy eye and the eye with central retinal vein occlusion are given in **Fig. 47**. A clear difference is seen in venular oxygen saturation (red color indicates 100% and violet indicates 0% oxygen saturation).

Another approach was introduced by Holm (2014). He also used the Topcon TRC-50DX mydriatic retinal camera (**Fig. 45** right) but modified the processing just subtracting one image from another. He developed the intensity-based algorithm using the approach of Saleh and Eswaran (2012) aimed at reducing background variations within the images to improve the contrast between the retinal vasculature and the background tissue, and to reduce the amount of segmentation noise. In the intensity-based algorithm only the green channel is used since it has the maximal contrast of the vasculature. The weighted sum of all three color-channels is used for Gabor filtering. A review of recent achievements in the field of oxymetry is given in the publication of Li *et al.* (2015).

Glucose Lidars

A wide range of optical technologies have been designed in attempts to develop robust noninvasive methods for glucose sensing in biological fluids, especially in human blood (Tuchin, 2009). The methods include infrared absorption; near-infrared scattering; Raman, fluorescent, and thermal gradient spectroscopies; as well as polarimetric, polarization heterodyning, photonic crystal, optoacoustic, optothermal, and OCT techniques. None of them is in a wide clinical use. We shall restrict our interest with two of them that could be easier qualified as lidar techniques.

Differential absorption glucose content measurement

Zhou *et al.* (2011) presented a method of glucose concentration detection in the anterior chamber with a differential absorption optical low-coherence interferometry technique. The anterior chamber is located between the cornea and the lens, with the thickness of 3.13 ± 0.50 mm. It is filled with transparent liquid-aqueous humor, with a total volume about 250 μL. In the experiments, the cornea and aqueous humor can be treated as nearly non-scattering substance. The difference in the absorption coefficient is much larger than the difference in the scattering coefficient, thus the influence of scattering can be neglected. The light directed on the iris being back-reflected passes the anterior chamber twice (**Fig. 48**).

Two light sources, one centered within (1625 nm) and the other centered outside (1310 nm) the glucose absorption band, are used for differential absorption measurement. Schematic diagram of the differential absorption system is shown in **Fig. 49**. The notations are as follows: FC – fiber coupler; OF – optical filter; WDM_1 – wavelength division multiplexer; WDM_2 – wavelength demultiplexer; M – vibrating mirror; L – lens. The light emerging from the eye carries the information of glucose concentration.

In the eye model and pig eye experiments, the authors obtained a resolution glucose level of 26.8 mg/dL and 69.6 mg/dL, respectively. This method has a potential application for noninvasive detection of glucose concentration in aqueous humor, which is related to the glucose concentration in blood.

Polarization-based glucose content measurement

Malik and Cote (2010) described their experiments with the idea to use the polarization parameters of anterior chamber aqueous humor to measure the glucose content. Glucose sensing using optical polarimetry is based on the phenomenon of optical activity,

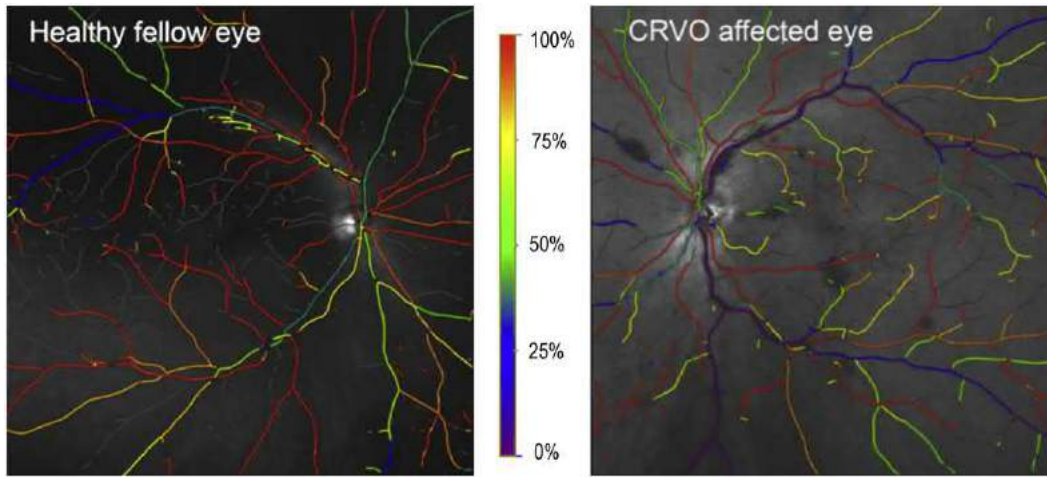


Fig. 47 Oximetry images from patient eyes: healthy eye (left) and the eye with central retinal vein occlusion (right). Reproduced from Kristjansdottir, J.V., 2014. Choroidal and retinal oximetry. MSc Thesis, University of Iceland Reykjavik.

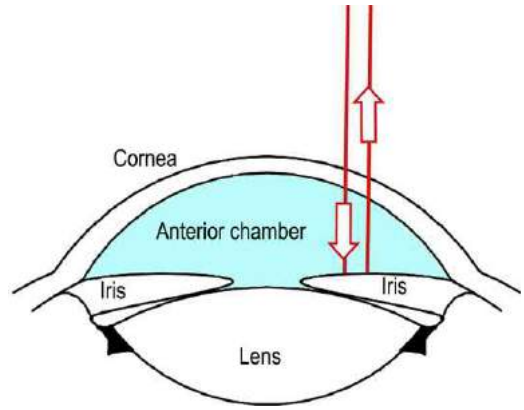


Fig. 48 Optical path through the anterior chamber with back reflection from iris. Reproduced from Zhou, Y., Zeng, N., Ji, Y., *et al.*, 2011. Iris as a reflector for differential absorption low-coherence interferometry to measure glucose level in the anterior chamber. Journal of Biomedical Optics 16 (1), 015004.

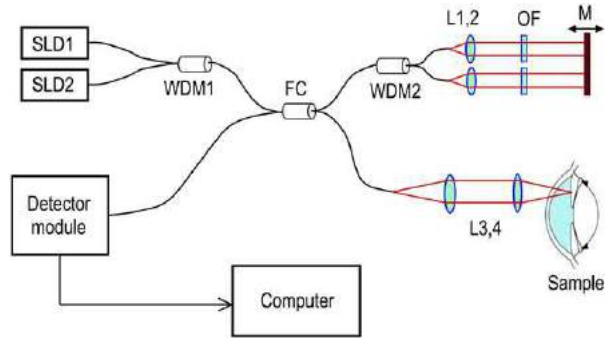


Fig. 49 Schematic diagram of the differential absorption system. Reproduced from Zhou, Y., Zeng, N., Ji, Y., *et al.*, 2011. Iris as a reflector for differential absorption low-coherence interferometry to measure glucose level in the anterior chamber. Journal of Biomedical Optics 16 (1), 015004.

which is the rotation of the orientation of plane polarized light passing through a solution of optically active molecules. The idea of the experiment was to measure the polarization parameters of the light that has passed through the cornea across the anterior chamber of the eye (inset of **Fig. 50**). The light has to be incident at a relatively glancing angle with respect to the posterior corneal surface in order for the beam to exit the anterior chamber through the cornea.

The problem, the authors met, is the birefringence of the cornea. In polarimetric glucose sensing through the anterior chamber of the eye, this corneal birefringence masks the optical rotation due to glucose in the aqueous humor, and therefore acts as a noise source. When the cornea moves due to eye motion, its birefringence changes the polarization of the light and this is a significant source of polarization noise. In addition, the index mismatch between the cornea and air causes the light beam to bend as it propagates and this complicates the ability to couple the light across the anterior chamber of the eye. It was demonstrated that change in the polarization vector due to corneal birefringence is at least an order of magnitude larger than that due to the change in optical path length and glucose concentration.

In his PhD work Malik (2011) made an effort to understand whether corneal birefringence is wavelength independent. This is important to know since the assumption is that by using multiple wavelengths this noise source may be canceled out. He designed a prototype and used two wavelengths: 635 nm and 532 nm. Experimental setup of the dual-wavelength optical polarimeter is presented in **Fig. 50**. The notations are as follows: P – polarizer; FC – Faraday compensator; BS – beam splitter; FM – Faraday modulator; PD – photodetector; LIA – lock-in amplifiers. The conclusion of this study was that using a dual-wavelength polarimetric approach has a potential to minimize the corneal birefringence noise and thus quantify blood glucose concentration.

Micro-Lidars for Velocity Measurement

Laser Velocimetry

Velocity measurement was another option that was tried as one of the first applications for lasers as transmitters in laser radars. The earliest publication was made by Yeh and Cummins (1964) from Columbia University who used a microscale lidar to study flow patterns in fluids by injecting dyes into the flow stream, illuminating the stream with monochromatic laser light, and measuring

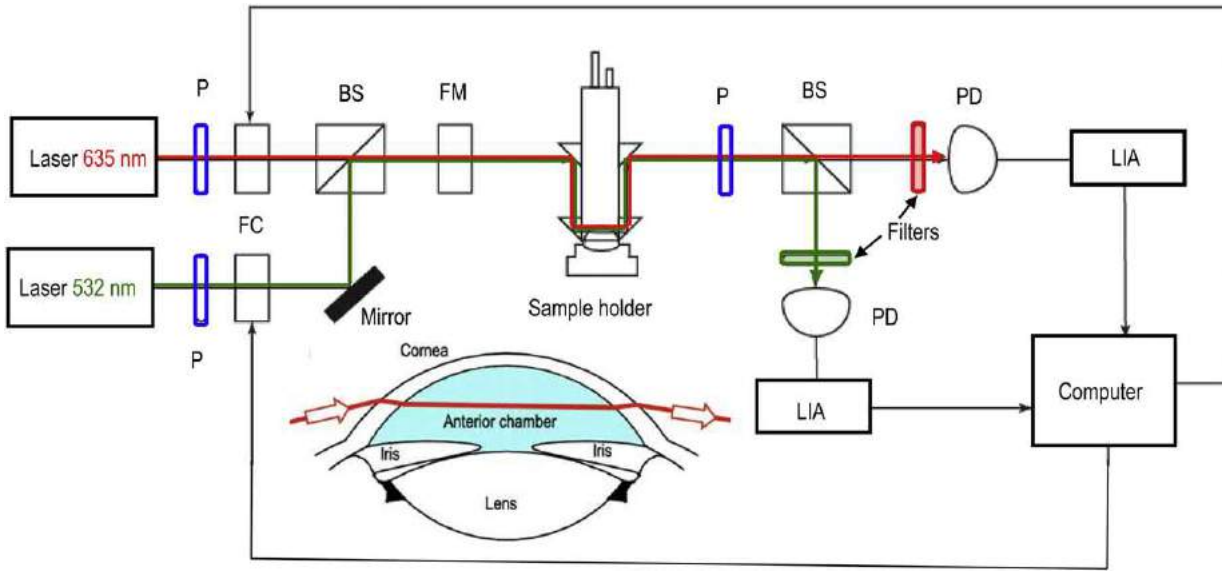


Fig. 50 Experimental setup of the dual-wavelength optical polarimeter. Reproduced from Malik, B.H., 2011. Dual wavelength polarimetry for glucose sensing in the anterior chamber of the eye. PhD Dissertation, Texas A&M University, College Station.

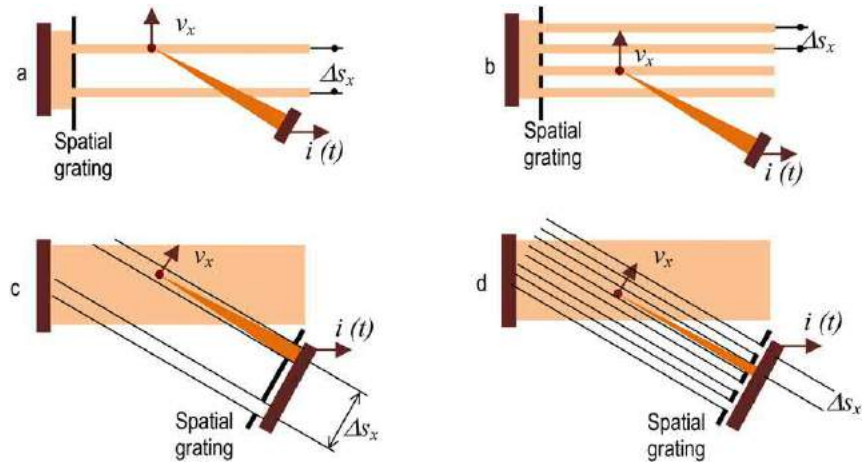


Fig. 51 Flow measurement techniques using an optically fixed measurement volume. ((a),(b)) Show the versions of the illumination by strips of light and ((c),(d)) show formation of detected zones at the receivers. Reproduced from Albrecht, H.-E., Borys, M., Damaschke, N., Tropea, C., 2003. Laser Doppler and Phase Doppler Measurement Techniques. Berlin: Springer.

the Doppler shifts in the Rayleigh scattered light. Twelve years later, the so-called "laser dual beam method" was patented (Schodl, 1976), according to which, two adjacent remote focal points are formed, where the light beams are focused. The particles in the gas stream passing through the focal points reflect light, generating start and stop pulses in the detectors, their time difference being used to calculate the velocity of the stream. Rotating the plane of the pair of the beams, the velocity can be determined in two dimensions. The first of these methods that measured the radial component of the velocity required monochromatic light, whereas the second one measuring the transversal components of the velocity is not limited by the degree of the coherence of light.

Analyzing different approaches to modeling and classification of the methods of velocity measurement, Molebny (1981) paid attention that the model of any of the known methods can be designed to include a field of readout points or readout strips, which would serve as references to measure the velocity. The readout points can be created by focusing the beams of light in adjacent points of space, or creating interference strips in the volume under study. These readout points can be immovable, fixed in space, like in the second of the above mentioned methods, or they can move (run), like in the field of interference of two beams of light with differing wavelengths. The interference can take place in the volume under study, or in the detector plane. Following this classification, we can get the information on the velocity from the field of fixed points, or from the field of moving points.

Fig. 51 left (a) and (c) illustrates the principle of measuring the velocity when the scattering particle crosses two zones. In the case (a) they are the zones (strips) illuminated by light, and the receiver sees both of them. In the case (b) the zones are created by

the fields of view of the receivers. For a single scattering particle, the interval Δt is measured. For multiple particles, the correlation interval Δt is measured.

Fig. 51 right **(c)** and **(d)** illustrates multiple zones created by a spatial grating at the transmitter side **(a)** or at the receiver side **(b)**. In both cases the frequency is measured that is created by bursts of scattered light when the moving particle crosses the illuminated zone.

The book summarizing the first decade of intensive studies of the new instrument – laser Doppler Anemometer (LDA) was published by Durst *et al.* in 1976, its second edition appeared in 1981 (Durst *et al.*, 1981). For further reading we recommend later publications of Albrecht *et al.* (2003), Zhang (2010), and specifically for ocular blood flow – the book of Schmetterer and Kiel (2012).

Blood flow measurement in retinal vessels

Successful application of laser velocimetry for flow measurements was described with application to blood vessels. The feasibility of using laser Doppler velocimetry (LDV) to measure blood flow in individual retinal vessels was first demonstrated by Riva *et al.* (1972), patented later (Riva, 1979). In accordance to the patent, the laser beam was directed into the eye (**Fig. 52**), the light scattered from the retina was received by the detector. After the detector, the signal was amplified, a loudspeaker was used as a measuring device. The type recorder and other devices was also foreseen for analysis. The laser light scattered from the retina contained the components scattered by RBCs flowing in the retinal vessels and the light scattered from the walls of the vessels and other structures that did not move (**Fig. 53**). At the output of the detector, the difference frequency was filtered and amplified. It was the Doppler-shift frequency spectrum that was further analyzed. In the experiments, the blood flow in a retinal artery of an anesthetized rabbit was studied. The maximum Doppler frequency shift arising from the light scattered by the RBCs flowing at the maximum speed was estimated from the spectrum and from the intraocular scattering geometry.

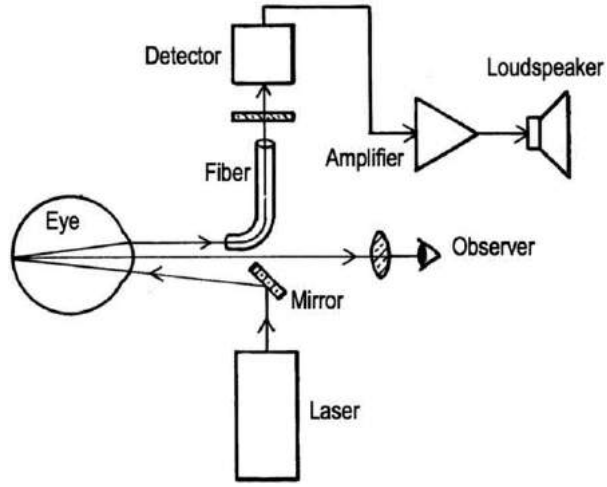


Fig. 52 Schematic of laser Doppler velocimeter patented by Riva in 1979. Reproduced from Riva, C.E., 1979. Blood Flow Measurement. US Patent 4,142,796. March 6, 1979.

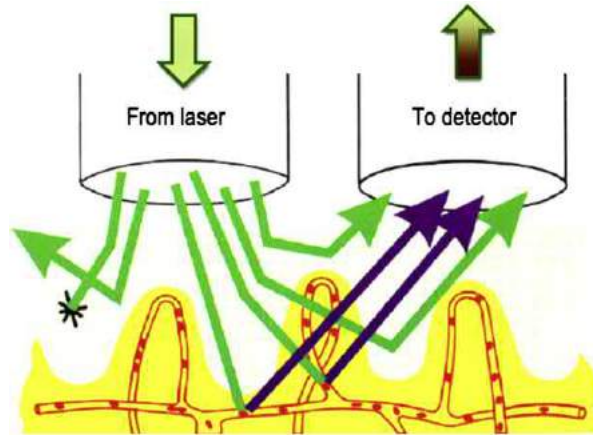


Fig. 53 Light scattered from moving (violet) and nonmoving (green) structures creating the Doppler-shift frequency spectrum. *Source*: Internet

Two-directional velocity measurement

Several years later, a new approach for determining blood velocity was proposed (Riva *et al.*, 1979, 1983). The procedure involves collecting the light scattered by the RBCs in two distinct directions separated by a known angle. Subsequent analysis yields an absolute measure of the velocity that is independent of the exact orientation of the vessel and of the relative angular orientation of the incident and scattered light beams with respect to the flow direction.

Velocity Measurement With Optical Doppler Tomography

Conventional LDV found its application to measure mean blood perfusion in the peripheral microcirculation. However, strong optical scattering in biological tissue limits the spatial resolution of flow measurements by LDV. Chen *et al.* (1997) described the OCT technique that provides not only in vivo blood flow measurement at discrete locations but also the tissue structure surrounding the vessel. Blood flow in a vein is imaged in **Fig. 54**, displaying a color-coded velocity image, and a velocity profile along the vein cross-section. The magnitude of blood flow velocity is maximal at the vessel center and decreases monotonically toward the peripheral wall. In a horizontal cross-section of the velocity image near the vessel center, an excellent fit of the velocity profile to a parabolic function indicates that blood flow in the vein is laminar.

Time dimension in velocity measurement

Ma *et al.* (2010) introduced a novel method to measure absolute blood flow velocity in vivo. It measures the velocities of blood plasma across the heart outflow tract. In experiments involving a chicken, the authors acquired four-dimensional (4D) $[(x, y, z) + t]$ images of absolute velocity distributions of the blood plasma with high spatial and temporal resolution in vivo. They reconstructed 4D microstructural images and obtained the orientation of the heart outflow tract at its maximum expansion. Assuming flow is parallel to the vessel orientation, the obtained centerline indicated the flow direction. Using this method, the flow velocity profiles were compared at various positions along the heart outflow tract of the chicken embryo.

OCT capillary velocimetry

OCT found its applications also in cardiovascular diagnostics and studies of cranial vascularization, etc. To demonstrate applications of OCT capillary velocimetry in cerebrovascular research, a 1310 nm spectral/Fourier domain OCT microscope was used for in vivo imaging of the rat cortex (Srinivasan *et al.*, 2012). The light source consisted of two SLDs yielding a spectral bandwidth of 170 nm. The axial (depth) and transverse resolution was 3.6 μm in tissue, full-width-at-half-maximum, and the imaging speed was 47,000 axial scans per second (17.3 μs exposure time), achieved by an In GaAs line scan camera. Capillary velocimetry and its imaging was performed during a hypercapnic challenge in a rat. An example of such imaging is given in **Fig. 55**.

Speckle contrast tomography

A novel tomographic method based on laser speckle contrast was introduced by Varma *et al.* (2014) that allows the reconstruction of a 3D distribution of blood flow in deep tissues. The experimental setup uses a temperature controlled continuous laser diode (785 nm, 90 mW) to probe the sample. A pair of galvo controlled mirrors are used to scan the laser point source. The light source is focused on the bottom of the sample and the produced speckle patterns are imaged from the top with a monochrome scientific CMOS camera, with exposure time 1 ms. The horizontal field of view is about 4 cm, a pixel diameter is 3×10^{-4} cm. The laser was set in every position during 0.5 s to acquire 35 intensity images. An example of a 3D slice plot of the reconstructed flow velocity for original velocity of 3.18 cm/s is shown in **Fig. 56**.

Triple-beam spectral-domain OCT

The three beam spectral-domain SD-OCT system (**Fig. 57**) consists of three independent subsystems. The SLD sources are operated at a central wavelength of 840 nm with a bandwidth of 50 nm and coupled into three miniature fiber collimators to reconstruct the 3D velocity vector. The three beams are split into reference and sample beam via a beam splitter. The notations in **Fig. 57** are as follows: C – miniature fiber collimator, FC – fiber collimator, L – lens, G – grating, BS – beam splitter, M – mirror, PP – dispersion-compensating prism pairs, NF – neutral density filter, FPT – facet prism telescope, T – telescope, MEMS – two-axis gimbal-less scanning mirror, LS – linear stage, SLD – superluminescent light emitting diode, CAM – line scan camera.

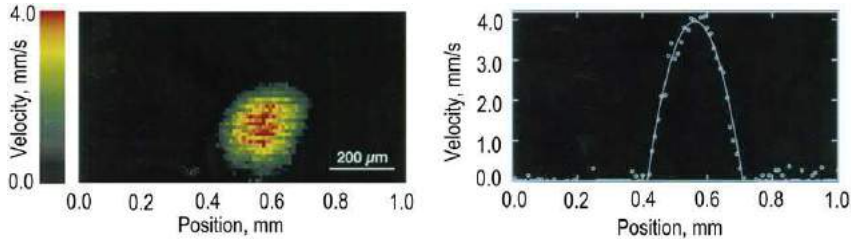


Fig. 54 Optical Doppler tomography images of blood flow in an in vivo biological model. Reproduced from Chen, Z., Milner, T.E., Srinivas, S., *et al.*, 1997. Noninvasive imaging of in vivo blood flow velocity using optical Doppler tomography. Optics Letters 22, 1119–1121.

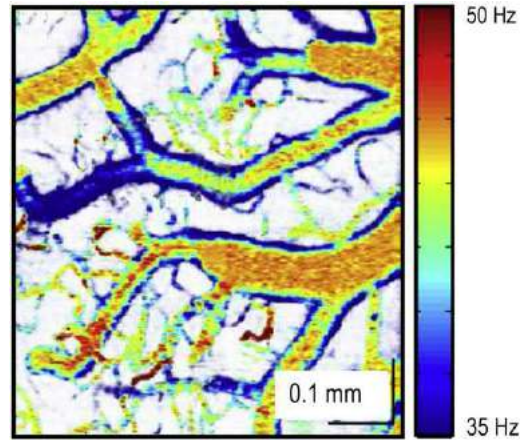


Fig. 55 Cortex velocity map during a hypercapnic challenge in a rat. Reproduced from Srinivasan, V.J., Radhakrishnan, H., Lo, E.H., *et al.,* 2012. OCT methods for capillary velocimetry, Biomedical Optics Express 3, 612–629.

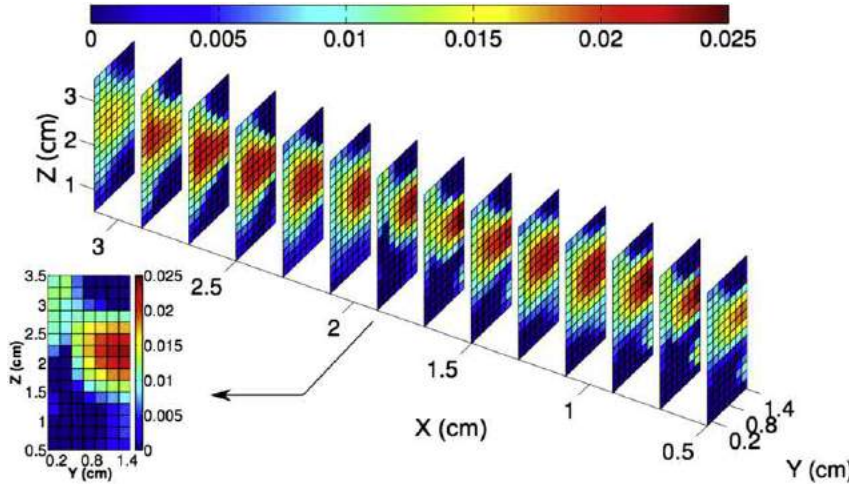


Fig. 56 Reconstructed 3D slice plot of blood flow in deep tissues. Reproduced from Varma, H.M., Valdes, C.P., Kristoffersen A.K., *et al.,* 2014. Speckle contrast optical tomography: A new method for deep tissue three-dimensional tomography of blood flow. Biomedical Optics Express 5, 1275–1289.

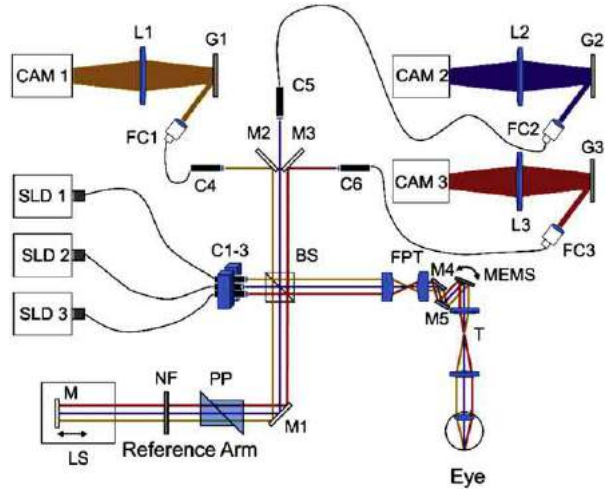


Fig. 57 Three beam spectral-domain optical low-coherence tomography (SD-OCT). Reproduced from Haindl, R., Trasischker, W., Wartak, A., *et al.,* 2015. Total retinal blood flow measurement by three beam Doppler optical coherence tomography. Biomedical Optics Express 7, 287–301.

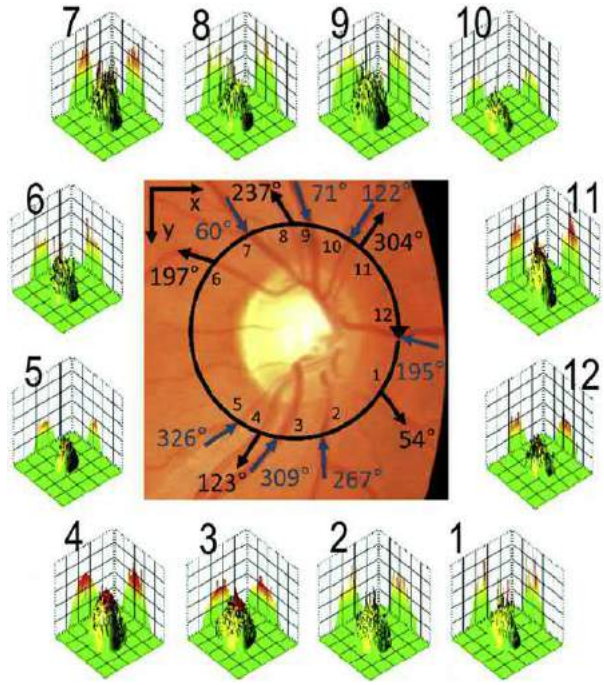


Fig. 58 Fundus photo with reconstructed velocity profiles. Reproduced from Haindl, R., Trasischker, W., Wartak, A., *et al.*, 2015. Total retinal blood flow measurement by three beam Doppler optical coherence tomography. Biomedical Optics Express 7, 287–301.

In the sample arm the two-axis gimbal-less MEMS mirror is used which is advantageous compared to a 2D galvo scanner, because both scanning axes are already in the same imaging plane. The MEMS mirror is driven with voltages ranging from 0 to 136 V. Any scan patterns can be realized (linear, raster, circular, and resonant patterns). Maximum optical scan angles of ± 11.5 degree can be achieved.

Fig. 58 displays the 3D velocity profile evaluation in the numbered vessels in the points on the circle. Arrows with numbers show the calculated vessel orientation clockwise relatively to the x axis. Reconstructed velocity profiles for the corresponding vessels around the optical nerve hypoplasia range from 0 to 40 mm/s.

Acoustical Studies and Vibrometry

Laser Doppler vibrometry

Doppler effect is used also to measure vibrations where they can exist. Laser Doppler vibrometry can provide advanced measurements in multiple physiological systems relevant to laboratory and field assessment of human stress and emotion (Tabatabai *et al.*, 2013). It can provide advanced recordings of myocardial and vascular performance, of respiratory efforts and sounds, and of tremor activity. It is responsive to laboratory stressors; muscle vibratory activity can be sensed from multiple muscles, including facial muscles, and the comparison is favorable with electromyography. Laser velocimetry can reliably assess facial muscle activity, associated with emotion and stress at low levels – below the threshold for visible facial deformations.

Wang *et al.* (2011) demonstrated a high-sensitivity pulsed laser vibrometer used to remotely determine the detailed, time-phased mechanical workings of various parts of the human heart. Results reported were validated by electrocardiography and accelerometer readings.

Successive pulse illumination in acoustical studies

In acoustical studies, Degroot (2009) used a method of successive pulse illumination of the investigated field with injected scattering particles (**Fig. 59**). She made videoregistration synchronously with pulse illumination, and measured the shifts of particles in successive frames in different phases of the acoustic process. An example of a 2D distribution of particle velocities is given in **Fig. 60**.

Vibration measurement of the vocal folds

An original technique was developed and used for clinical studies by George *et al.* (2008). To measure vibration dynamics of the vocal folds, they used a high-speed TV camera registration of a laser line projected onto the object in the triangulation mode and calculated the amplitude distribution for each frame (**Fig. 61**). The endoscopic laser projection system stretches a laser line beam in one direction and projects it as a thin laser sheet at an angle onto the surface of the vocal folds. The laser projection channel consists of a semiconductor diode laser and a cylindrical optical system.

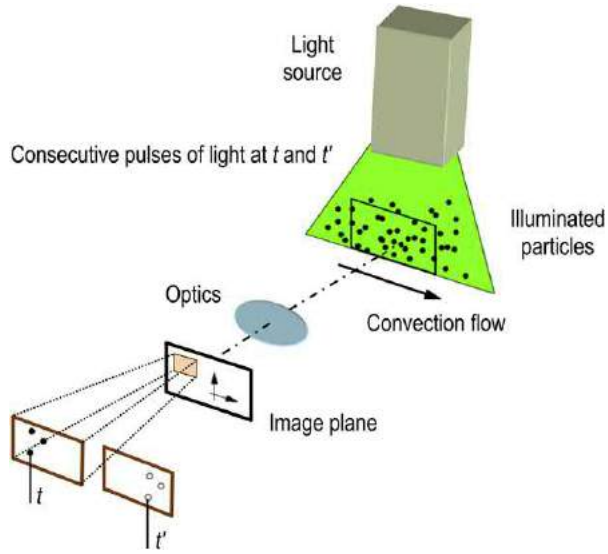


Fig. 59 Method of successive pulse illumination. Reproduced from Degroot, A., 2009. Contribution à l'estimation de la vitesse acoustique par vélocimétrie laser Doppler & application à l'étalonnage de microphones en champ libre. Le Mans: Université du Maine, p. 209.

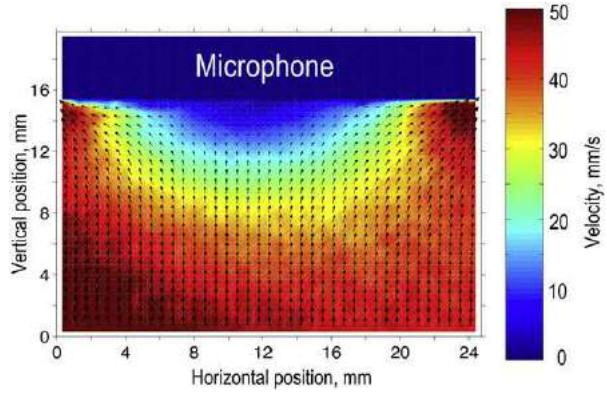


Fig. 60 2D distribution of particle velocities. Reproduced from Degroot, A., 2009. Contribution à l'estimation de la vitesse acoustique par vélocimétrie laser Doppler & application à l'étalonnage de microphones en champ libre. Le Mans: Université du Maine, p. 209.

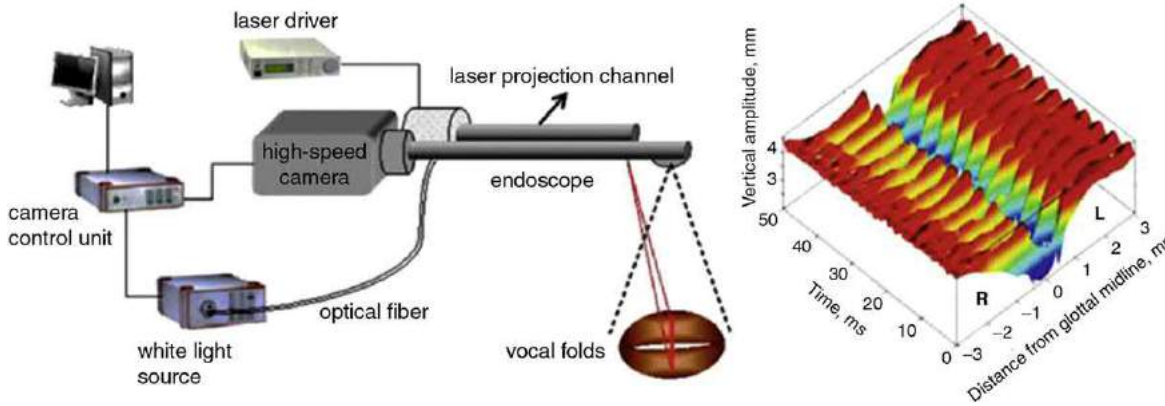


Fig. 61 Schematic view of the laser triangulation laryngoscope and color coded 3D vibration pattern of the vocal folds. Reproduced from George, N.A., de Mul, F.F.M., Qiu, Q., Rakhorst, G.R., Schutte, H.K., 2008. New laryngoscope for quantitative high-speed imaging of human vocal folds vibration in the horizontal and vertical direction. Journal of Biomedical Optics 13, 064024.

The laser projection channel is firmly attached to a 90 degree rigid endoscope, which acts as the receiving channel. The optical axes of the two channels are separated by a distance of 9 mm at the tip of the system. Within a normal working distance of 60–70 mm from the tip of the endoscope, the laser line is 18–20 mm long and 0.4 mm wide. A semiconductor laser emits at 653 nm, delivering an effective laser power density of 1.8 mW/mm^2, keeping below the exposure limit of 2 mW/mm^2. The red laser is used because it gives minimum absorption and satisfactory reflectance by the tissue in the visible spectral region.

A compact high-speed digital color camera is used for recording the images. It can record images continuously for a maximum of 2 s at the rate of 4000 fps with an image resolution of 256×256 pixels. Temporarily stored data in the camera memory is downloaded to the computer for further analysis. Vibration profiles in both horizontal and vertical directions were calibrated and measured with a resolution of ± 50 mm.

Wavefront Sensing Micro-Lidars

In laser weapon, to deliver the maximum density of energy to the target, atmospheric turbulence should be compensated by means of irradiating the target with a wavefront conjugated laser beam (He, 2002). The wavefront distortion is measured by a laser radar that receives the back reflected laser radiation with a matrix of photodetectors and provides the data for conjugation.

A similar need to measure the wavefront aberrations flared up in ophthalmology with the advent of vision correction using the ablation procedure after the successful wavefront corrections made on live human eye by Pallikaris *et al.* (1990), Seiler *et al.* (1990), McDonald *et al.* (1990), and others. Taboada and Archibald (1981) were the first engineers who paid attention to the potential of ultaviolet excimer laser to ablate the corneal epithelium. Trockel *et al.* (1983) used this effect for making the incisions correcting the ametropia.

Liang *et al.* (1994) used the matricial sensor, manufactured in Arizona by Platt, to measure the wave front aberrations in the human eye using the laser light reflected in outward direction from retina. The task to measure the inward, physiologically genuine, wavefront aberrations was discussed between Molebny and Pallikaris in 1995 (Molebny, 2013). The first clinical ray tracing aberrometer was tested on the live human eyes in the hospital of Crete University in 1998.

Hartmann–Shack Ocular Wavefront Sensing

The idea of light projection through "a perforated screen" was described by Hartmann (1904), developed further by Shack and Platt (1971) who replaced the perforated screen with "an array of contiguous lenticular elements." A complete system designed and fabricated for astronomy was delivered to Cloudcroft, NM in the early 1970s but was never installed, and the facility was decommissioned (Platt and Shack, 2001).

Liang *et al.* (1994) with lenslet array from Platt, and Molebny *et al.* (1996a,b) with home-made holographic Fresnel microlenses built their wavefront sensors in which the laser light was projected on the retina, and back reflection captured by video cameras through the multielement optics.

A simplified schematic is shown in **Fig. 62**. A thin laser beam is directed into the eye, and being reflected from the retina, exits the eye. Usually, the microlenses are arranged in the form of matrix. Each microlens of the lenslet array projects its part of the beam cross-section on the CCD (CMOS, or other type) camera. Computer reconstructs the wavefront, and can calculate any of the derivative parameters of the optical system of the eye. The combination of a set (a matrix) of lenses (a lenslet array) and a 2D photosensitive detector are usually called Hartmann–Shack sensor. The image in the plane of the photosensitive detector is called hartmanngram. The synonyms of Hartmann–Shack sensor are Shack–Hartmann or Hartmann–Platt sensor.

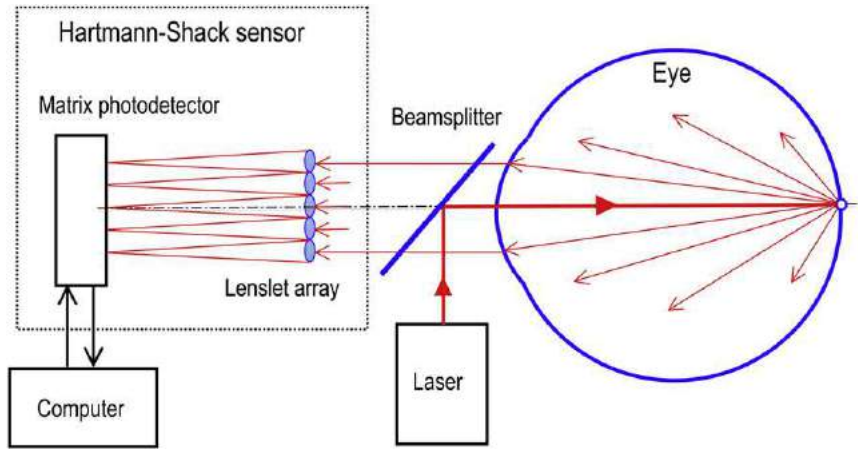


Fig. 62 Principle of measuring the wavefront errors in the eye with Hartmann–Shack sensor. *Source*: Molebny.

Commercial Hartmann–Shack aberrometers

Let us look at some examples of commercial aberrometers based on Hartmann–Shack wavefront sensing. One of the oldest companies on the market was VISX, later integrated with Abbott medical optics (AMO). iDesign is one of the latest instruments (see Section Relevant Websites). The maximal pupil diameter is 7 mm, the range of measured aberrations from -16 D to $+12$ D of sphere and astigmatism up to 8 D, with the dynamic range for higher order aberrations 8 μm root-mean square (RMS). The number of analyzed points is 1250. The previous WaveScan model captured 240 data points, and had the range of higher order aberrations 1.3 μm RMS. Besides wavefront measurement, the instrument combines functions of corneal topography, autorefractometry, pupillometry, and keratometry. The Hartmann–Shack approach is used in this instrument not only for the wavefront sensing but also to collect information on corneal topography.

Topcon having come later to the market, used a lot of experience from the earlier models of other manufacturers, especially for the interface. Like iDesign machine, Topcon's KR-1W has also five functions: wavefront measurement, corneal topography, autorefractometry, pupillometry, and keratometry (see Section Relevant Websites).

iProfiler (**Fig. 63** left) from Carl Zeiss Vision GmbH also keeps the tendency of combining several functions in one instrument: ocular wavefront aberrometer, autorefractometer, corneal topographer, and keratometer (see Section Relevant Websites). Ocular wavefront is reconstructed up to seventh-order Zernike decomposition. The instrument is provided with touch screen. The instrument measures both eyes automatically in approximately 30 s. An example of displayed data with a map of an ocular wavefront errors is shown in **Fig. 63** right. The map is color-coded (the scale is shown to the left of the wavefront map).

Wavefront Reconstruction

Presentation of measured wavefront data and the procedure of the wavefront error reconstruction for ophthalmologic applications were described by Southwell (1980), applied by Lane and Tallon (1992).

In general, aberrometers measure the gradient of the wavefront error function or, in other words, the deflection of rays from an un-aberrated direction. From each measured ray or location in the wavefront, the data contain four numbers. The first two are the horizontal (x) and vertical (y) coordinates of the location given. The second two numbers are the measured horizontal and vertical component values of the gradient. When reconstructing, the value of wavefront error should be calculated in each (x, y) point. 2D distribution (a map) of this error can be found as an approximation using the values of wavefront error in each point, or as an interpolation.

Analytically, the approximated surface of the wavefront error contains a set of 2D polynomials that is similar to a stack of multiple layer transparencies. This way of presentation is called a global one because each member of the polynomial is a component describing the whole map, not a part of it. Interpolation describing local specificities (in each zone of the map) is called a zonal presentation.

The approximation based on Zernike polynomial presentation is not the only possibility. Other types of polynomials can be used: Bhatia–Zernike, Fourier, Fourier-Mellin, Taylor, Bessel, Legendre, Didon, Karhunen–Loeve, etc. In most cases, preference is given to Zernike polynomials, because their components, describing the lower-order aberrations, coincide with the conventional ophthalmologic description. When the wavefront error is reconstructed in Zernike polynomials, it can be transformed into the presentation as a Fourier series, and vice versa (Dai, 2006). The same transformation can be done from Zernike to Taylor polynomials, and vice versa. The simplest way of interpolation is linear, just connecting two points. More sophisticated is connecting by a curve, satisfying certain requirements.

Fig. 63 iProfiler – a version of the aberrometer from Carl Zeiss Vision GmbH (left); a map of ocular wavefront errors (right). *Source*: Molebny.

Ray Tracing Aberrometry

The first ophthalmology-related publications baptizing the technique as ray tracing were made by Molebny *et al.* (1997) and by Navarro and Losada (1997). The ray tracing technique uses measurement of the position of a thin laser beam projected onto the retina. A beam of light is directed into the eye parallel to the visual axis having passed a two-directional (x, y) acousto-optic deflector and a collimating lens (Fig. 64).

The front focal point of the collimating lens is positioned in the center of scanning C of the acousto-optic deflector. Each entrance point, one at a time, provides its own projection on the retina. The position sensing detector measures the transverse displacement δx, δy of the laser spot on retina. An objective lens is used to optically conjugate the retina and the detector plane.

Commercial iTrace visual function analyzer

Fig. 65 shows a version of iTrace visual function analyzer, one of its functions is the wavefront sensing. A set of entrance points overlaid on the image of the eye is encircled by a green outer contour corresponding to the shape of the pupil, which is reconstructed from the eye image. In the process of beam displacement over the eye aperture, data on the transverse aberration for each point of the beam entrance into the eye are collected. They represent a 2D distribution known as a retinal spot diagram (Fig. 65 right bottom). The instrument uses a narrow (0.30 mm) diode laser beam with the wavelength 785 nm being displaced over the entrance pupil of an eye while kept parallel to the visual axis. To measure the positions of laser spots on retina, x and y linear arrays (each of 512 photodetectors) are used.

In the process of measurement, an image of the pupil is captured, its size and position are automatically measured in each TV frame. The permission on firing the probing radiation is given by the software only when the positions of the pupil and visual axis are within a predetermined correspondence. The data, captured with a position-sensing detector, are processed and transferred to

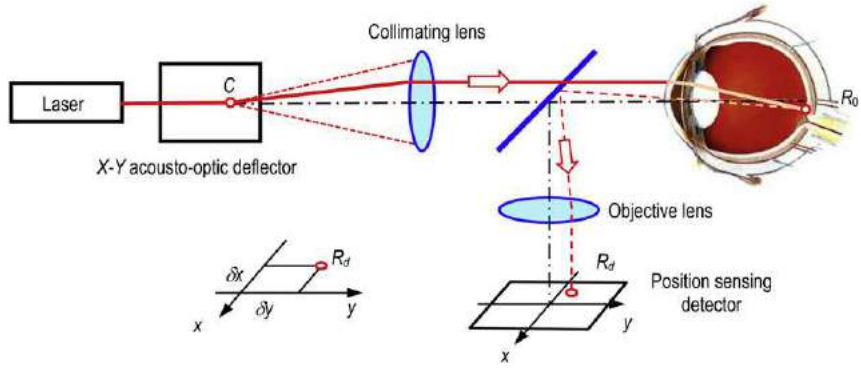


Fig. 64 Ocular ray tracing with a fast *x-y* acousto-optic deflector. *Source*: Molebny.

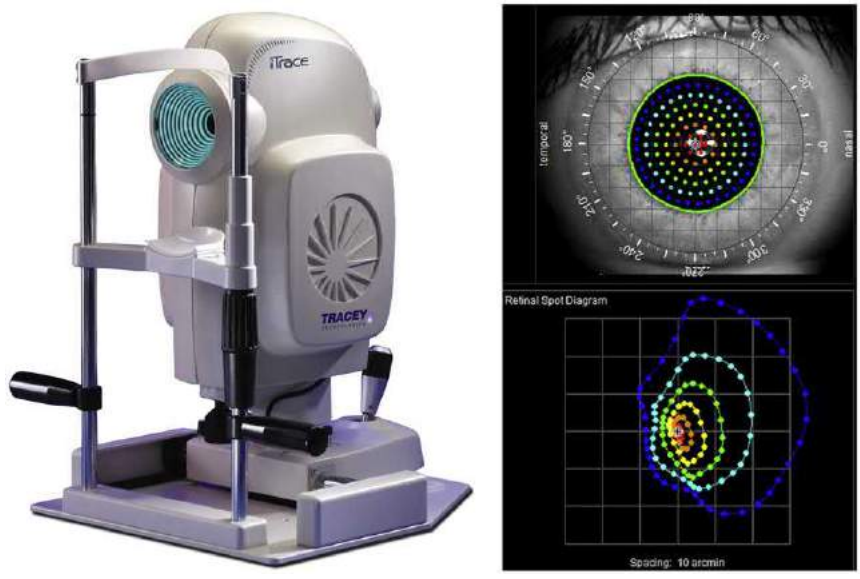


Fig. 65 iTrace visual function analyzer (left); laser beam positions in the entrance aperture of the eye (right top); collocation of laser beam projections on the retina – retinal spot diagram (right bottom). *Source*: Molebny.

the computer. The total time of scanning for the entire aperture of the eye is within 100 ms. The duration depends on the number of test points at the eye entrance pupil (from 64 to 256). Acousto-optic deflectors are very fast devices: transition time for switching a position is about 10 μs, i.e., for 256 points, less than 3 ms is necessary. This is a great advantage of acousto-optics as compared to much slower scanners (no more than 20 points/s), used in the laboratory studies of Navarro and Moreno-Barriuso (1999).

Reconstructed maps and functions

Examples of reconstructed wavefront maps are given in **Fig. 66**, both in 2D and 3D versions. The acquired data are used also to calculate the derivative parameters of the optical system of the eye like point spread function (PSF) and modulation transfer function (MTF). Examples of the PSF and MTF are given in **Fig. 67**. The instrument allows both 2D and 3D display of both of them, as well as including in calculations any combination of Zernike modes. Any of these functions can be calculated for any size of the pupil.

Locating the visual axis

To optimally correct refractive errors, the eye must be properly positioned for measurements before and during surgery. Manzanera *et al.* (2015) showed that the optimal positioning is in favor of the visual axis. According to Reinstein (2009), corneal refractive ablations centered on the visual axis result in high quality of vision. Based on three-beam scanning micro-lidar, a technique was proposed to objectively locate the visual axis of the eye (Molebny, 2017) which is defined as the line connecting the point of fixation and the center of the foveola. The laser beam consists of three components (f_0, $f_0 + F_x$, and $f_0 + F_y$), two of them ($f_0 + F_x$ and $f_0 + F_y$) being shifted along x and y axes relatively to the central one (f_0). The triple-beam is scanned along a closed trajectory (**Fig. 68**), the difference in depth between two pairs of beams Δh_x and Δh_y is determined in the process of scanning. The center of the foveola is defined as the crossing of steepest inclinations (**Fig. 69**).

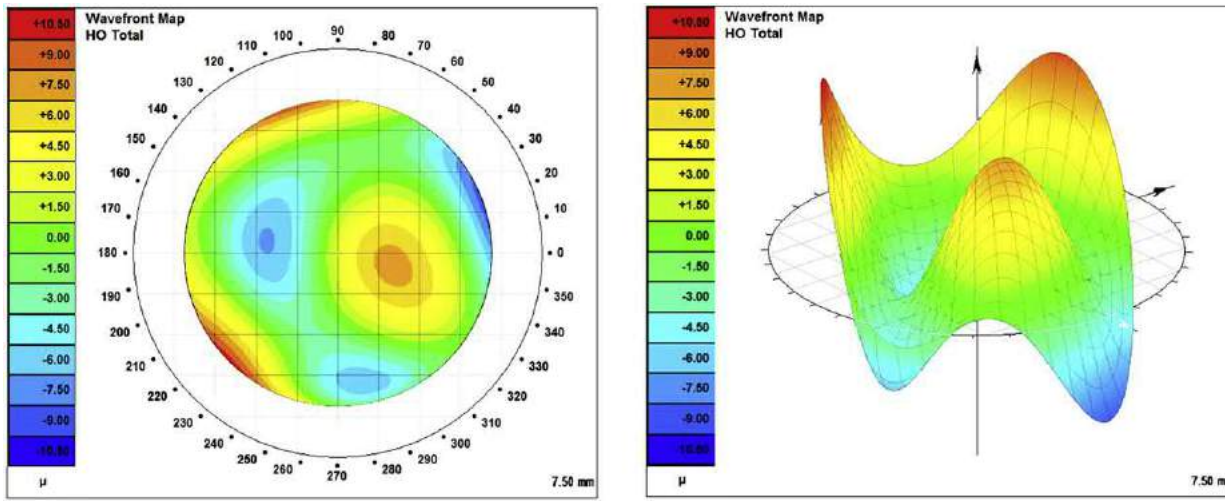


Fig. 66 Wavefront map displayed in Trace in 2D and 3D versions. *Source*: Molebny.

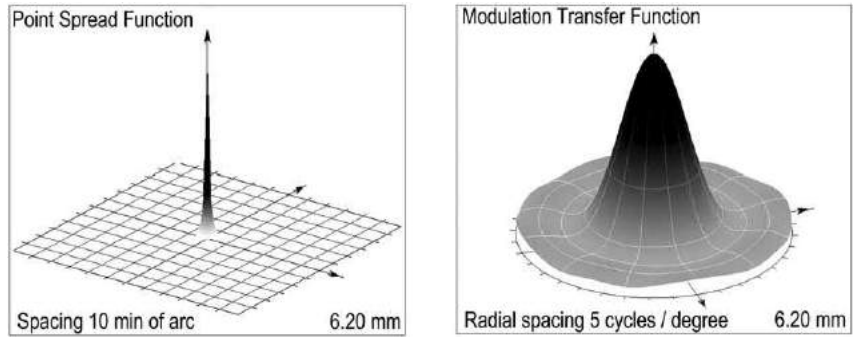


Fig. 67 Reconstructed 3D point spread function (PSF) (left) and 3D modulation transfer function (MTF) (right) of the human eye. *Source*: Molebny.

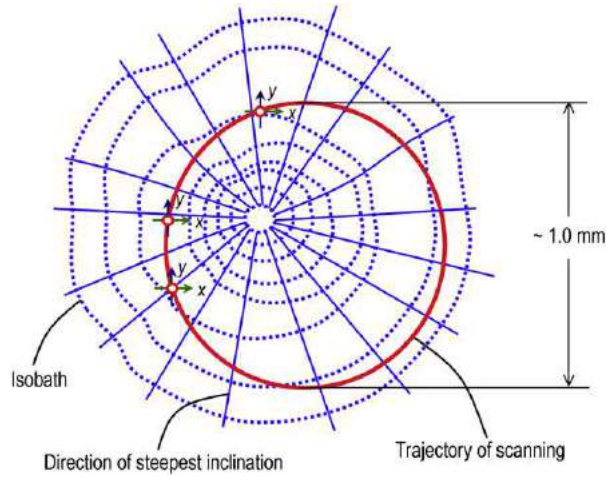


Fig. 68 Trajectory of scanning when determining the center of the foveola. *Source*: Molebny.

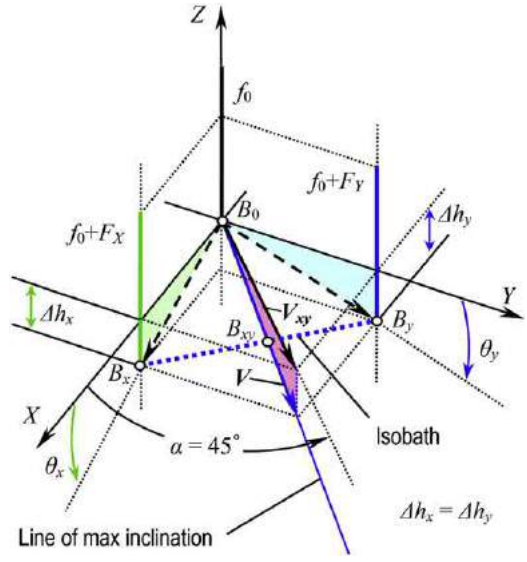


Fig. 69 Triple-beam configuration enabling the measurement of the depth difference. *Source*: Molebny.

Conclusion

When I started working on this chapter, it seemed to be a short walk across a small section of the big and mature field of laser radars, from which this younger sister section borrowed the principal ideas. The reality brought me to a shoreless ocean of new ideas and new opportunities. Technology solutions in this younger sister field appeared even more sophisticated: think about the clever CD/DVD head-up multiplied in millions of computers and other cutting-edge gadgets. Or recollect the flow cytometry taking the responsibility not only to quantitatively evaluate the number of any kind of cells in the human blood, but also to interrogate the signaling from the most dangerous tumor qualifying cells.

The intelligence of military rangefinders look like a children toy in comparison with the interactive play stations analyzing a huge amount of information and evaluating not only the distances to several players but also their positions and movements. Micro-lidars opened new perspectives in studying the human vision and measuring its parameters that enabled achieving the eagle acuity. Micro-lidars make feasible what could not be done with other means, for example, to objectively locate the visual axis of the eye that can't be anatomically even defined.

OCT whose birthday was just several years ago having achieved the highest level of both, information values and financial costs, is attacked today to have replaced the expensive distance-modulation parts with the cheap voice coils from the CD/DVD head-ups.

Sophisticated play stations are only the introduction to what is the dream of the mankind – to robotics, where the micro-lidars will play an important role of the robotic eyes communicating with the robot's brain.

Acknowledgment

The author thanks Dr. Paul McManamon for his idea to inspire the writing of this chapter.

See also: Very High Range Resolution Lidars

References

Abramson, A., 1978. Light-in-flight recording by holography. Optics Letters 3, 121–123.
Adler, D., 2012. Bench-to-bedside success: Intravascular optical coherence tomography. SPIE Newsroom. doi:10.1117/2.1201208.004382.
Agishev, R.R., 2009. Lidar Monitoring of the Atmosphere (in Russian). Moscow: Fizmatgiz.
Albrecht, H.-E., Borys, M., Damaschke, N., Tropea, C., 2003. Laser Doppler and Phase Doppler Measurement Techniques. Berlin: Springer.
Alfano, R., Pu, Y., 2013. Optical biopsy for cancer detection. In: Jelínková, H. (Ed.), Lasers for Medical Applications. Cambridge: Woodhead, pp. 325–367.
Alfano, R.R., Tang, G.C., Pradhan, A., *et al.*, 1991. Light sheds light on cancer. Bulletin of the New York Academy of Medicine 67, 143–150.
Amann, M.C., Bosch, T., Lescure, M., *et al.*, 2001. Laser ranging: A critical review of usual techniques for distance measurement. Optical Engineering 40 (1), 10–19.
Anderson, T.L., Covert, D.S., Marshall, S.F., *et al.*, 1996. Performance characteristics of a high-sensitivity, three-wavelength total scatter/backscatter nephelometer. Journal of Atmospheric and Oceanic Technology 13, 967–986.
Barman, I., Dingari, N.C., Saha, A., *et al.*, 2013. Application of Raman spectroscopy to identify microcalcifications and underlying breast lesions at stereotactic core needle biopsy. Cancer Research 73 (11), 3206–3215.
Baudelet, M. (Ed.), 2013. Laser Spectroscopy for Sensing: Fundamentals, Techniques and Applications. Amsterdam: Elsevier and Woodhead Publishing.
Bourdet, G.L., Orszag, A.G., 1979. Absolute distance measurements by CO_2 laser multiwavelength interferometry. Applied Optics 18, 225–227.
Chen, Z., Milner, T.E., Srinivas, S., *et al.*, 1997. Noninvasive imaging of in vivo blood flow velocity using optical Doppler tomography. Optics Letters 22, 1119–1121.
Cohen, D., Rais, D., Sali, E., *et al.*, 2016. Error Compensation in Three-Dimensional Mapping. US Patent 9,330,324. May 3, 2016.
Craig, F.E., 2014. Flow cytometry. In: McKenzie, S.B. (Ed.), Clinical Laboratory Hematology, second ed. essex: pearson, pp. 956–974.
Creath, K., 1988. Phase-measurements interferometry techniques. In: Wolf, E. (Ed.), Progress in Optics. New York, NY: Elsevier Science, pp. 349–393.
Dai, G., 2006. Zernike aberration coefficients transformed to and from Fourier series coefficients for wavefront representation. Optics Letters 31, 501–503.
Degroot, A., 2009. Contribution à l'estimation de la vitesse acoustique par vélocimétrie laser Doppler & application à l'étalonnage de microphones en champ libre. Le Mans: Université du Maine, p. 209.
de Groot, P., 1991. Interferometric laser profilometer for rough surfaces. Optics Letters 16, 357–359.
de Groot, P., Kishner, S., 1991. Synthetic wavelength stabilization of a two-color laser diode interferometer. Applied Optics 30, 4026–4033.
Delori, F.C., 1988. Noninvasive techniques for oximetry of blood in retinal vessels. Applied Optics 27, 1113–1125.
Delori, F.C., Weiter, J.J., Mainster, M.A., Flook, V.A., 1983. Oxygen saturation measurements in retinal vessels. Investigative Ophthalmology & Visual Science 24 (Suppl.), 13. ARVO.
Devaux, J.C., Abdelkader, H.H., Colle, E., 2013. A multi-sensor calibration toolbox for kinect: application to kinect and laser range finder fusion. In: 16th International Conference Advance Robotics (ICAR 2013), November 2013, Montevideo, Uruguay.
Dohl, M., 2015. Smoke Detector. US Patent 8,941,505. January 27, 2015.
Dresel, T., Häusler, G., Venzke, H., 1992. Three-dimensional sensing of rough surfaces by coherence radar. Applied Optics 31, 919–925.
Dsouza, R., Subhash, H., Neuhaus, K., *et al.*, 2014. Dermascope guided multiple reference optical coherence tomography. Biomedical Optics Express 5, 2870–2882.
Durst, F., Melling, A., Whitelaw, J.H., 1981. Principles and Practice of Laser-Doppler Anemometry, second ed. London: Academic Press.
Dynamic 3D Imaging: Dyn3D Proceedings, DAGM 2009 Workshop. Kolb, A., Koch, R. (Eds.), Berlin: Springer.
Fercher, A.F., Mengedoht, K., Werner, W., 1988. Eye-length measurement by interferometry with partially coherent light. Optics Letters 13, 186–188.
Fornage, B.D., Sneige, N., Ross, M.I., *et al.*, 2004. Small (less 2 cm) breast cancer treated with ultrasonographically (US) guided radiofrequency ablation: Feasibility study. Radiology 231, 215–224.
Freedman, B., Shpunt, A., Machline, M., Arieli, Y., 2012. Depth Mapping Using Projected Patterns. US Patent 8,150,142, April 3, 2012.
Fujimoto, J.G., De Silvestri, S., Ippen, E.P., *et al.*, 1986. Femtosecond optical ranging in biological systems. Optics Letters 11, 150–152.
García, J., Zalevsky, Z., García-Martínez, P., *et al.*, 2008. Three-dimensional mapping and range measurement by means of projected speckle patterns. Applied Optics 47, 3032–3040.
George, N.A., de Mul, F.F.M., Qiu, Q., Rakhorst, G.R., Schutte, H.K., 2008. New laryngoscope for quantitative high-speed imaging of human vocal folds vibration in the horizontal and vertical direction. Journal of Biomedical Optics 13, 064024.
Gerstner, K., Tschudi, T., 1994. New diode laser light source for absolute ranging two-wavelength interferometry. Optical Engineering 33 (8), 2692–2696.
Guiagliumi, G., Akasaka, T., Sirbu, V., Kubo, T., 2016. Optical coherence tomography. In: Topol, E.G., Teirstein, P.S. (Eds.), Textbook of Interventional Cardiology, seventh ed. Philadelphia: Elsevier, pp. 990–1012.
Haka, A.S., Shafer-Peltier, K.E., Fitzmaurice, M., *et al.*, 2005. Diagnosing breast cancer by using Raman spectroscopy. Proceedings of the National Academy of Sciences 102 (35), 12371–12376.
Haka, A.S., Volynskaya, Z., Gardecki, J.A., *et al.*, 2006. In vivo margin assessment during partial mastectomy breast surgery using Raman spectroscopy. Cancer Research 66 (6), 3317–3322.
Han, Y., Lu, D., Rao, R., Wang, Y., 2009. Determination of the complex refractive indices of aerosol from aerodynamic particle size spectrometer and integrating nephelometer measurements. Applied Optics 48, 4108–4117.
Hansard, M., Lee, S., Choi, O., Horaud, R., 2012. Time of flight cameras: Principles, methods, and applications. SpringerBriefs in Computer Science, Springer.
Hardarson, S.H., 2012. Retinal oximetry. PhD Thesis, University of Iceland, Reykjavik.
Hartmann, J., 1904. Objektivuntersuchungen. Z. Instrumentenkde 24, 1–21.
He, G.S., 2002. Optical phase conjugation: Principles, techniques, and applications. Progress in Quantum Electronic 26, 131–191.
Hee, M.R., Puliafito, C.A., Duker, J.S., *et al.*, 1998. Topography of diabetic macular edema with optical coherence tomography. Ophthalmology 105, 360–370.
Heintzenberg, J., Charlson, R.J., 1996. Design and applications of the integrating nephelometer: A review. Journal of Atmospheric and Oceanic Technology 13, 987–1000.
Hickham, J.B., Frayser, R., Ross, J.C., 1963. A study of retinal venous blood oxygen saturation in human subjects by photographic means. Circulation 27, 375–385.
Hogan, J.N., Wilson, C.J., 2009. Multiple Reference Non-Invasive Analysis System. US Patent 7,526,329. April 29, 2009.
Holm, S.P., 2014. Optical imaging of retinal blood flow: Studies in automatic vessel extraction, alignment, and driven changes in vessel oxymetry. PhD Thesis, University of Manchester.

Huang, D., Swanson, E.A., Lin, C.P., *et al.*, 1991. Optical coherence tomography. Science 254, 1178–1181.
Hu, P., Tan, J., Yang, H., *et al.*, 2011. Phase-shift laser range finder based on high speed and high precision phase-measuring techniques. In: 10th International Symposium on Measurement Technology and Intelligent Instruments. June 29–July 2, 2011, KAIST, Daejeon, Korea.
Ishii, Y., Onodera, R., 1991. Two-wavelength laser-diode interferometry that uses phase-shifting techniques. Optics Letters 16, 1523–1525.
Jelínková, H. (Ed.), 2013. Lasers for Medical Applications. Cambridge: Woodhead Publishing.
Jerlov, N.G., 1976. Marine Optics. Amsterdam: Elsevier.
Journet, B., Lourme, J.C., 2000. Laser range finding based on correlation method. In: 16th IMEKO World Congress, Vienna, Austria, September 25–28, 2000.
Journet, B.A., Poujouly, S., 1998. High-resolution laser rangefinder based on a phase-shift measurement method. Proceedings of SPIE 3520, 123–132.
Komachi, Y., Katagiri, T., Sato, H., Tashiro, H., 2009. Improvement and analysis of a micro Raman probe. Applied Optics 48, 1683–1696.
Komachi, Y., Sato, H., Aizawa, K., Tashiro, H., 2005. Micro-optical fiber probe for use in an intravascular Raman endoscope. Applied Optics 44, 4722–4732.
Kristjansdottir, J.V., 2014. Choroidal and retinal oximetry. MSc Thesis, University of Iceland. Reykjavik.
Laing, R.A., Cohen, A.J., Friedman, E., 1975. Photographic measurements of retinal blood oxygen saturation: Falling saturation rabbit experiments. Investigative Ophthalmology 14 (8), 606–610.
Lane, R.G., Tallon, M., 1992. Wave-front reconstruction using a Shack–Hartmann sensor. Applied Optics 31, 6902–6908.
Lange, R., 2000. 3D Time-of-flight distance measurement with custom solid-state image sensors in CMOS/CCD-technology. PhD Thesis, University of Siegen, Germany.
Lebedev, A.A., 2014. Optics herald. Rozhdestvensky Optical Society Bulletin 145, 11–15. St-Petersburg.
Lexer, F., Hitzenberger, C.K., Fercher, A.F., Kulhavy, M., 1997. Wavelength-tuning interferometry of intraocular distances. Applied Optics 36, 6548–6553.
Li, J., Ma, J., Wang, N., 2015. Application of retinal oximeter in ophthalmology. Chinese Journal of Ophthalmology 51 (11), 864–868.
Liang, J., Grimm, B., Goelz, S., Bille, J.F., 1994. Objective measurement of wave aberrations of the human eye with the use of a Hartmann-Shack wave-front sensor. Journal of the Optical Society of America 11, 1949–1957.
Liu, L., Gardecki, J.A., Nadkarni, S.K., *et al.*, 2012. Imaging the subcellular structure of human coronary atherosclerosis using 1-μm resolution optical coherence tomography (μOCT). Nature Medicine 17 (8), 1010–1014.
Lonnqvist, J., 1999. Apparatus and Method for Measuring Visibility and Present Weather. US Patent 5,880,836. March 9, 1999.
Ma, Z., Liu, A., Yin, X., *et al.*, 2010. Measurement of absolute blood flow velocity in outflow tract of HH18 chicken embryo based on 4D reconstruction using spectral domain optical coherence tomography. Biomedical Optics Express 1, 798–811.
Malik, B.H., 2011. Dual wavelength polarimetry for glucose sensing in the anterior chamber of the eye. PhD Dissertation, Texas A&M University, College Station.
Malik, B.H., Cote, G.L., 2010. Characterizing dual wavelength polarimetry through the eye for monitoring glucose corneal birefringence and ultimately enable the design and development of a polarization-based glucose monitoring system. Biomedical Optics Express 1, 1247–1258.
Manzanera, S., Prieto, P.M., Benito, A., *et al.*, 2015. Location of achromatizing pupil position and first Purkinje reflection in a normal population. Investigative Ophthalmology & Visual Science 56, 962–966.
Mazzarella, J., Cole, J., 2015. The anatomy of an OCT scan. Review of Optometry. https://www.reviewofoptometry.com/article/the-anatomy-of-an-oct-scan.
McDonald, M.B., Kaufman, H.E., Frantz, J.M., *et al.*, 1990. Excimer laser ablation in a human eye. Case report. Archives of Ophthalmology 107 (5), 641–642.
Mitchell, H.L., 2010. Nephelometer Instrument for Measuring Turbidity of Water. US Patent 7,663,751. February 16, 2010.
Mizrahi, A., Russell, L., Three wavelength integrating nephelometer. Available at: https://www.esrl.noaa.gov/gmd/aero/instrumentation/neph_desc.html.
Molebny, V., 2013. Wavefront measurement in ophthalmology. Aberrometry through the eyes of an engineer. In: Tuchin, V.V. (Ed.), Handbook of Coherent-Domain Optical Methods, vol. 2. New York, NY: Springer, pp. 315–361.
Molebny, V., 2017. Method of locating the visual axis objectively. Ophthalmology Visual Optics 37 (3), doi:10.1111/opo.12376.
Molebny, V.V., 1981. Optical Radar Systems. Moscow: Mashinostroenie, in Russian.
Molebny, V.V., Kamerman, G.W., Smirnov, E.M., *et al.*, 1998. Three beam scanning laser radar microprofilometer. Proceedings of SPIE 3380, 280–283.
Molebny, V.V., Kurashov, V.N., Pallikaris, I.G., Naoumidis, L.P., 1996a. Adaptive optics technique for measuring eye refraction distribution. Proceedings of SPIE 2930, 147–157.
Molebny, V.V., Pallikaris, I.G., Naoumidis, L.P., *et al.*, 1996b. Dual-beam dual-frequency scanning laser radar for investigation of ablation profiles. Proceedings of SPIE 2748, 68–75.
Molebny, V.V., Pallikaris, I.G., Naoumidis, L.P., *et al.*, 1997. Retina ray-tracing technique for eye-refraction mapping. Proceedings of SPIE 2971, 175–183.
Motz, J.T., Hunter, M., Galindo, L.H., *et al.*, 2004. Optical fiber probe for biomedical Raman spectroscopy. Applied Optics 43, 542–554.
Moyle, J., 2002. Pulse Oximetry, second ed. London: BMJ Books.
Mutto, C.D., Zanuttigh, P., Cortelazzo, G.M., 2012. Time-of-flight cameras and Microsoft Kinect™. New York, NY: Springer.
Navarro, R., Losada, M.A., 1997. Aberrations and relative efficiency of light pencils in the living human eye. Optometry and Vision Science 74, 540–547.
Navarro, R., Moreno-Barriuso, E., 1999. Laser ray-tracing method for optical testing. Optics Letters 24, 951–953.
Nehmadi, Y., Guterman, Y., 2016. System and Method for Providing 3D Imaging. US Patent 9,303,989. April 5, 2016.
Pallikaris, I., Papatsanaki, M., Stathi, E., *et al.*, 1990. Laser in situ keratomileusis. Lasers in Surgery and Medicine 10 (5), 463–468.
Park, H., Chodorow, M., Kompfner, R., 1981. High resolution optical ranging system. Applied Optics 20, 2389–2394.
Pittman, R.N., Duling, B.R., 1975. A new method for the measurement of percent oxyhemoglobin. Journal of Applied Physiology 38 (2), 315–320.
Platt, B.C., Shack, R.S., 2001. History and principles of Shack–Hartmann wavefront sensing. Journal of Refractive Surgery 17, S573–S577.
Pohlmann, K.C., 2011. Principles of Digital Audio, sixth ed. New York, NY: McGraw-Hill.
Poujouly, S., Journet, B., 2000. Laser range finding by phase-shift measurement: Moving towards smart systems. Proceedings of SPIE 4189, 152–160.
Reinstein, D., 2009. Ablations centred on visual axis may be more reliable. Eurotimes 14 (4), 21.
Riva, C.E., 1979. Blood Flow Measurement. US Patent 4,142,796. March 6, 1979.
Riva, C.E., 1983. Fundus Camera-Based Retinal Laser Doppler Velocimeter. US Patent. 4,402,601. September 6, 1983.
Riva, C.E., Feke, G.T., Eberli, B., Benary, V., 1979. Bidirectional LDV system for absolute measurement of blood speed in retinal vessels. Applied Optics 18, 2301–2306.
Riva, C.E., Ross, B., Benedek, G.B., 1972. Laser Doppler measurements of blood flow in capillary tubes and retinal arteries. Investigative Ophthalmology & Visual Science 11, 936–944.
Rovati, L., Minoni, U., Bonardi, M., F. Docchio, F., 1998. Absolute distance measurement using comb-spectrum interferometry. Journal of Optics 29 (3), J121–J127.
Saleh, M.D., Eswaran, C., 2012. An efficient algorithm for retinal blood vessel segmentation using h-maxima transform and multilevel thresholding. Computer Methods in Biomechanics and Biomedical Engineering 15 (5), 517–525.
Sali, E., Avraham, A., 2014. Three-Dimensional Mapping and Imaging. US Patent 8,717,417. May 6, 2014.
Sali, E., Avraham, A., 2016. Three-Dimensional Mapping and Imaging. US Patent 9,350,973. May 24, 2016.
Sasaki, O., Okazaki, H., 1986. Sinusoidal phase modulating interferometry for surface profile measurement. Applied Optics 25, 3137–3140.
Schmetterer, L., Kiel, J.W. (Eds.), 2012. Ocular Blood Flow. Heidelberg: Springer.
Schodl, R., 1976. Measuring Device for the Measurement of Fluid Flow Rates. US Patent 3,941,477.
See, C.W., V. Iravani, V., Wickramasinghe, H.K., 1985. Scanning differential phase contrast optical microscope: Application to surface studies. Applied Optics 24, 2373–2379.
Seiler, T., Kahle, G., Kriegerowski, M., 1990. Excimer laser (193 nm) myopic keratomileusis in sighted and blind human eyes. Refractive and Corneal Surgery 6 (3), 165–173.
Sekine, A., Minegishi, I., Koizumi, H., 1993. Axial eye-length measurement by wavelength-shift interferometry. Journal of the Optical Society of America A 10, 1651–1655.

Shack, R.V., Platt, B.C., 1971. Production and use of a lenticular Hartmann screen. Journal of the Optical Society of America A 61, 656.
Shapiro, S.L. (Ed.), 1984. Ultrashort Light Pulses. Picosecond Techniques and Applications, second ed. Berlin: Springer.
Shimotahira, H., Iizuka, K., Chu, S.C., *et al.*, 2001. Three-dimensional laser microvision. Applied Optics 40, 1784–1794.
Southwell, W.H., 1980. Wave-front estimation from wave-front slope measurements. Journal of the Optical Society of America 70, 998–1006.
Srinivasan, V.J., Radhakrishnan, H., Lo, E.H., *et al.*, 2012. OCT methods for capillary velocimetry. Biomedical Optics Express 3, 612–629.
Staszewski, R.B., Balsara, P.T., 2005. Phase-domain all-digital phase-locked loop. IEEE Transactions on Circuits and Systems II: Express Briefs 52 (3), 159–163.
Subhash, H.M., 2014. Smartphone based multiple reference optical coherence tomography (MRO™) system. Biomedical Optics. paper BT3A.72.
Subhash, H.M., Hogan, J.N., Leahy, M.J., 2015. Multiple reference optical coherence tomography for smartphone applications. SPIE Newsroom. doi:10.1117/2.1201503.005807.
Sugimoto, M., Nakamura, S., Inoue, Y., *et al.*, 2013. LT-PAM: A ranging method using dual frequency optical signals. International Journal on Smart Sensing and Intelligent Systems 6 (3), 791–809.
Surrel, Y., 1993. Phase stepping: A new self-calibrating algorithm. Applied Optics 39, 3598–3600.
Tabatabai, H., Oliver, D.E., Rohrbaugh, J.W., Papadopoulos, C., 2013. Novel applications of laser Doppler vibration measurements to medical imaging. Sensing and Imaging: An International Journal 14 (1), 13–28.
Taboada, J., Archibald, C.J., 1981. An extreme sensitivity in the corneal epithelium to far UV ArF excimer laser pulses. Proceedings of the Scientific Program of the Aerospace Medical Association. May 4, 1981. San Antonio, TX.
Trockel, S.L., Srinivasan, R., Braren, B., 1983. Excimer laser surgery of the cornea. American Journal of Ophthalmology 96 (6), 710–715.
Tuchin, V.V. (Ed.), 2009. Handbook of Optical Sensing of Glucose in Biological Fluids and Tissues. Boca Raton, FL: CRC Press – Taylor & Francis Group.
Varma, H.M., Valdes, C.P., Kristoffersen, A.K., *et al.*, 2014. Speckle contrast optical tomography: A new method for deep tissue three-dimensional tomography of blood flow. Biomedical Optics Express 5, 1275–1289.
Wang, C.C., Trivedi, S., Kutcher, S., *et al.* (2011). Non-contact human cardiac activity monitoring using a high sensitivity pulsed laser vibrometer. In: Proceedings of Conference on Lasers and Electro-Optics (CLEO), paper CWB6.
Wang, H., Jun, Xu, J., He, D., *et al.*, 2010. Real-time range imaging system based on a light-emitting diode array phase-shift range finder for fast three-dimensional shape acquisition. Optical Engineering 49 (7), 073201.
Yeh, Y., Cummins, H.Z., 1964. Localized fluid flow measurements with He–Ne laser spectrometer. Applied Physics Letters 4 (10), 176–178.
Zhang, Zh, 2010. LDA Application Methods. Laser Doppler Anemometry for Fluidic Dynamics. Heidelberg: Springer.
Zhou, Y., Zeng, N., Ji, Y., *et al.*, 2011. Iris as a reflector for differential absorption low-coherence interferometry to measure glucose level in the anterior chamber. Journal of Biomedical Optics 16 (1), 015004.

Relevant Websites

https://www.vision.abbott/us/homepage.html
Abbott.
https://wiww.asus.com/3D-Sensor/Xtion_PRO/
Asus Xtion Pro.
http://mesa-imaging.ch
HEPTAGON.
http://sante.ro/wp-content/uploads/2016/04/Ilumien-OPTIS-ENG-Brochure.pdf
Ilumien™ Optis™ PCI Optimization System.
http://www.icsightsound.com/pdf/i.scription_factsheet.pdf
iProfiler by Zeiss.
http://www.topcon-medical.eu/eu/products/40-kr-1w-wavefront-analyzer.html
TOPCON.