

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 3

Displays ■ Biomedical: Tissue Measurements ■ Imaging (Non-Medical)



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

Copyright © 2018 Elsevier Ltd. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers may always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

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

For information on all publications visit our
website at <http://store.elsevier.com>

		Working together to grow libraries in developing countries
www.elsevier.com • www.bookaid.org		

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

Jorge Ojeda-Castaneda

Universidad de Guanajuato

Fang-Chung Chen

National Chiao Tung University (NCTU)

Chau-Jern Cheng

National Taiwan Normal University

Lukas Chrostowski

University of British Columbia

Steve Cundiff

University of Michigan

Casimer DeCusatis

Marist College

Hui Deng

University of Michigan

Henry O. Everitt

Duke University

Mike Fiddy

University of North Carolina at Charlotte

Almantas Galvanaskas

University of Michigan

David Gershoni

Technion Institute of Technology

Junsang Kim

Duke University

Mackillo Kira

University of Marburg

Paul McManamon

Ladar and Optical Communications Inst (University of Dayton)

Mary-Ann Mycek

University of Michigan

Xingjie Ni

Penn State University

Christoph Schmidt

University of Göttingen

Mansoor Sheik-Bahae

University of New Mexico

Colin Sheppard

Italian Institute of Technology

Han-Ping David Shieh

National Chiao Tung University

Brian Vohnsen

University College Dublin

Xiushan Zhu

University of Arizona

CONTENTS OF VOLUME 3

Editorial Board	v
List of Contributors for Volume 3	ix
Editors in Chief	xi
Introduction	xiii

VOLUME 3

Displays – Introduction	<i>Han-Ping D Shieh</i>	1
Liquid Crystal Physics and Materials	<i>Huang-Ming Philip</i>	8
TFT Materials and Devices	<i>Po-Tsun Liu</i>	12
Color Formation of LCD – Spatial and Temporal	<i>Fang-Cheng Lin</i>	17
LCD Components	<i>Fang-Cheng Lin</i>	25
Projection Displays – LCD/MEMS/LCoS Based	<i>Fleming Chuang</i>	32
3D Displays, Stereoscopic/Autostereoscopic 3D	<i>Yi-Pai Huang and Chun-Ho Chen</i>	44
Wearable Displays	<i>Paul Yang</i>	51
View Angles of Liquid Crystal Displays	<i>Huang-Ming Philip Chen</i>	59
Organic Light-Emitting Diode (OLED)	<i>Chin H (Fred) Chen, Wen-Shi Wen, and Chin-Ti Chen</i>	64
Optical Characteristics of Display Devices	<i>Han-Ping D Shieh</i>	70
Reflective Display Technologies	<i>Han-Ping D Shieh</i>	79
In Situ Optical Tissue Diagnostics/Miniaturized Optoelectronic Sensors for Tissue Diagnostics	<i>Seung Yup Lee, Yooree Grace Chung, and Mary-Ann Mycek</i>	86
In Situ Optical Tissue Diagnostics/Laser Speckle-Based Spectroscopy Techniques in Biomedicine	<i>Karthik Vishwanath and Sara Zanfardino</i>	95
Photodynamic Therapy and Photobiomodulation: Can All Diseases be Treated with Light?	<i>Michael R Hamblin</i>	100
Resolution and Multiple Scattering in Imaging	<i>Edwin A Marengo, Zambrano-Nunez Maytee, and Edson S Galagarza</i>	136
Coherent Diffractive Imaging	<i>Lei Tian</i>	146
Nonuniqueness in Imaging	<i>Greg Gbur</i>	156
Phase Space and Imaging	<i>Markus Testorf</i>	164
Information Theory in Imaging	<i>FO Huck and CL Fales</i>	175
Inverse Problems and Computational Imaging	<i>M Bertero and P Boccacci</i>	187
Adaptive Optics	<i>C. Pernechele</i>	197
Phase Conjugation and Image Correction	<i>EN Leith</i>	205
Hyperspectral Imaging	<i>ML Huebschman, RA Schultz, and HR Garner</i>	211
Imaging Through Scattering Media	<i>AC Boccara</i>	219

Infrared Imaging	<i>K Krapels and RG Driggers</i>	229
Interferometric Imaging	<i>DL Marks</i>	241
Multiplex Imaging	<i>A Lacourt</i>	247
Photon Density Wave Imaging	<i>V Toronov</i>	256
Three-Dimensional Field Transformations	<i>R Piestun</i>	262
Volume Holographic Imaging	<i>G Barbastathis</i>	267
Wavefront Sensors and Control (Imaging Through Turbulence)	<i>CL Matson</i>	272

LIST OF CONTRIBUTORS FOR VOLUME 3

G Barbastathis

Massachusetts Institute of Technology, Cambridge, MA, USA

M Bertero

University of Genova, Genova, Italy

P Boccacci

University of Genova, Genova, Italy

AC Boccara

*Ecole Supérieure de Physique et Chimie Industrielles
and Centre National de la Recherche Scientifique, Paris,
France*

Chin H (Fred) Chen

*Shine Materials Technology Co., Ltd., Kaohsiung,
Taiwan, R.O.C.*

Chin-Ti Chen

*Chemistry Institute, Academia Sinica, Taipei, Taiwan,
R.O.C.*

Chun-Ho Chen

National Chiao Tung University, Hsinchu, Taiwan

Fleming Chuang

Coretronic Corporation, Taiwan, Republic of China

Yooree Grace Chung

University of Michigan, Ann Arbor, MI, United States

RG Driggers

*US Army Night Vision and Electronic Sensors
Directorate, Fort Belvoir, VA, USA*

CL Fales

NASA Langley Research Center, Hampton, VA, USA

Edson S Galagarza

*Northeastern University, Boston, MA, United States and
Technological University of Panama, Panama City,
Panama*

HR Garner

University of Texas, Dallas, TX, USA

Greg Gbur

*University of North Carolina at Charlotte, Charlotte,
NC, United States*

Michael R Hamblin

*Massachusetts General Hospital, Boston, MA, United
States; Harvard Medical School, Boston, MA, United
States; and Harvard-MIT Division of Health Sciences
and Technology, Cambridge, MA, United States*

Yi-Pai Huang

National Chiao Tung University, Hsinchu, Taiwan

FO Huck

*NASA Langley Research Center, Hampton, VA,
USA*

MI Huebschman

University of Texas, Dallas, TX, USA

K Krapels

*US Army Night Vision and Electronic Sensors
Directorate, Fort Belvoir, VA, USA*

A Lacourt

Université de Franche-Comté, Besancon, France

Seung Yup Lee

*Georgia Institute of Technology/Emory University,
Atlanta, GA, United States*

EN Leith

University of Michigan, Ann Arbor, MI, USA

Fang-Cheng Lin

National Chiao Tung University, Hsinchu, Taiwan

Po-Tsun Liu

National Chiao Tung University, Hsinchu, Taiwan

Edwin A Marengo

*Northeastern University, Boston, MA, United States and
Technological University of Panama, Panama City,
Panama*

DL Marks

*University of Illinois at Urbana-Champaign, Urbana,
IL, USA*

CL Matson

Air Force Research Laboratory, Kirtland, NM, USA

Zambrano-Nunez Maytee

*Technological University of Panama, Panama City,
Panama*

Mary-Ann Mycek

University of Michigan, Ann Arbor, MI, United States

C Pernechele

Astronomical Observatory of Padova, Padova, Italy

Huang-Ming Philip

National Chiao Tung University, Hsinchu, Taiwan

Huang-Ming Philip Chen

National Chiao Tung University, Hsinchu, Taiwan

R Piestun

University of Colorado, Boulder, CO, USA

RA Schultz

University of Texas, Dallas, TX, USA

Han-Ping D Shieh

National Chiao Tung University, Hsinchu, Taiwan

Markus Testorf

Thayer School of Engineering, Hanover, NH, United States

Lei Tian

Boston University, Boston, United States

V Toronov

University of Illinois at Urbana-Champaign, Urbana, IL, USA

Karthik Vishwanath

Miami University, Oxford, OH, United States

Wen-Shi Wen

Shine Materials Technology Co., Ltd., Kaohsiung, Taiwan, R.O.C.

Paul Yang

Sun Yat-Sen University, Guangzhou, China

Sara Zanfardino

Miami University, Oxford, OH, United States

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.

Displays – Introduction

Han-Ping D Shieh, National Chiao Tung University, Hsinchu, Taiwan

© 2018 Elsevier Ltd. All rights reserved.

Nomenclature

ACR	Ambient contrast ratio	LC	Laser crystallization
a-IGZO	Amorphous In-Ga-Zn-O	LCD	Liquid crystal display
ALD	Atomic layer deposition	LCOS	Liquid crystal on silicon
AOM	Acoustic optical modulator	LUMO	Lowest unoccupied molecular orbital
AR	Augmented reality	LPD	Local-primary-desaturation
BCE	Back channel etched	MILC	Metal-induced lateral crystallization
BEF	Brightness enhancement film	MIP	Memory-in-pixel
BRDF	Bidirectional reflectance distribution function	OCB	Optically compensated bend
CCFLs	Cold cathode fluorescence lamps	OLED	Organic light emitting diode
CCT	Correlated color temperature	OTFTs	Organic TFTs
CGL	Charge-generation layer	PEN	Polyethylene naphthalate
ChLCD	Cholesteric liquid crystal display	PECVD	Plasma-enhanced chemical vapor deposition
CIE	Commission International de l'Eclairage	PLD	Pulse laser deposition
CMFs	Color matching functions	PPI	Pixels per inch
CMYK	Cyan, magenta, yellow, black	PMMA	Polymethylmethacrylate
CRT	Cathode ray tube	PWM	Pulse width modulation
DBEF	Double brightness enhancement film	PVP	Poly-vinylphenyle
DLP	Digital light processing	PVA	Polyvinyl alcohol
DMD	Digital mirror display	QD	Quantum-dot
EIL	Electron injection layer	QR-LPD	Quick response liquid powder display
EL	Electroluminescence	R2R	Roll-to-roll
EPD	Electrophoretic display	SLM	Spatial light modulator
ESR	Enhanced specular reflector	SmC*	Chiral-smectic C
EWD	Electro-wetting display	SOI	Silicon-on-insulator
FOLED	Flexible OLED	SOP	System-on-panel
FoV	Field of view	SPC	Solid phase crystallization
FPD	Flat panel display	SRAM	Static random access memory
FSC	Field sequential color	TAC	Triacetyl cellulose
HIL	Hole injection layer	TFT	Thin film transistor
HOMO	Highest occupied molecular orbital	TTA	Triplet-triplet annihilation
IES	Internal enhancement layer	VA	Vertical alignment
IMOD	Interferometric modulator display	VFD	Vacuum fluorescent display
IPS	In-plane switching	VR	Virtual reality

A display can be defined as an object that transfers information to the viewer and/or an output unit that gives a visual representation of data. Displays can be divided into two types: emissive and non-emissive. By emissive we mean that the display surface gives off light. In the non-emissive, the display or a small portion of the display acts as a shutter or spatial filter which also can be reflective in nature. In the non-emissive mode, the display can pass light from a source in back of or in front of the shutter.

The dominant displays in the 20th century were Cathode Ray Tube (CRT). Since then, a non-emissive display – active matrix Thin Film Transistor (TFT) Liquid Crystal Displays (LCD), and potentially an emissive one-Active matrix OLED are dominated. Virtually all other types/forms of displays only have very niche applications or just disappeared due to cost-performance, form factor, health concerns, and others.

Emissive displays include the CRT, projection CRT, vacuum fluorescent display (VFD), field emissive display (FED), electroluminescent (EL) display, light emitting diode (LED), organic light emitting diode (OLED), plasma, Thin Film EL (TFEL) Displays, etc.

Most of emissive displays use certain types of emissive materials called phosphors. The phosphors are called cathodoluminescent, photoluminescent, and electroluminescent, depending whether the excitation action was due to the interaction of the phosphor with cathode rays, photons, or low-voltage electrons. Cathodoluminescent phosphors are used in CRTs – projection CRTs, VFDs, and FEDs. EL displays can use inorganic phosphors or organic phosphors such as those used in OLEDs. Under photoluminescence displays, we have “plasma displays,” where the excitation of the plasma in a localized region gives rise to UV photons which strike the phosphor screen in that region.

Under the heading of “shutter-type” we can consider LCDs that consist of AMLCDs, super twisted nematic liquid crystal displays (STN LCDs) and bistable LCDs. Another shutter-type is the liquid crystal on silicon (LCOS) display and the movable digital mirror display (DMD), and the digital light processing (DLP) projecting engines. Another spatial display type is both normal and dynamic inks. A dynamic ink can be a “polarizable” ink display in which a two shade (light/dark) ink sphere element can rotate under the presence of an electric field to show the light or dark side. The DMD movable mirrors reflect the digital information to the lens system and the unwanted rays are reflected away from the lens to a black light capture area.

In the 21st century, virtually all displays are in the form of “flat,” a very thin in thickness, a typical display system includes display panel and associated electronics to drive the display panel, timing controller, Gamma reference for color correction, signal conversion, and others as shown in Fig. 1. Within flat panel displays (FPD), their characteristics can be divided by luminance, driving method, display driving panel materials, processing temperature, non-emissive (liquid crystal), and emissive materials, as shown in Fig. 2.

Cathode Ray Tube

The CRT is a display that has been in use for over 100 years (the Braun CRT tube was developed in 1898). The CRT uses a phosphor screen which is bombarded by energetic electrons and emits visible light radiation. The single electron gun CRT was used for years in early monochrome television sets and computer displays. Monochrome green, amber, and white phosphors have been used in monitor tubes. CRT is not a flat panel display, due to the thickness, weight, difficult to scale up in size and most serious one of up to tens of hundreds driving voltage, CRTs have been totally displaced by FPD since the beginning of the 21st century.

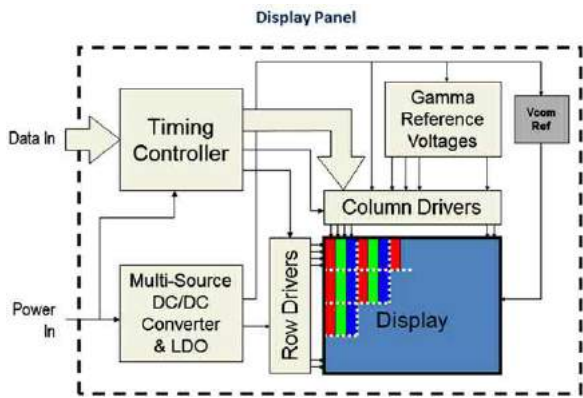


Fig. 1 A display system.

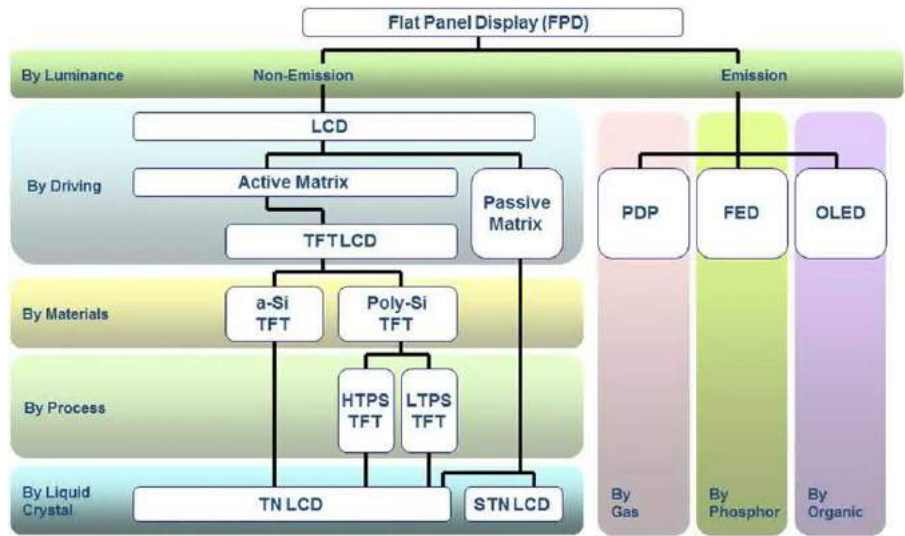


Fig. 2 Flat panel display (FPD) family and their characteristics.

Color CRT

The more ubiquitous method of obtaining a color display is the device called the shadow mask CRT (and the similar Trinitron CRT). This color CRT is designed in such a manner that through the use of an aperture mask, (seen in Fig. 3) the electrons from the red electron gun strike the red phosphor and likewise electrons from the green and blue electron guns strike the green and blue phosphors.

The combination allows the viewer to see a color image. This color separation action is seen in Fig. 4. Although Fig. 4 shows a delta tri-color design, there are also in-line systems in which the mask consists of an array of slots which are somewhat rectangular in nature. In this array, many slots can be on a line, or a mask made can be made of a series of wires as used in the Trinitron design.

The disadvantage of the shadow mask technique is that the mask can capture as much as 80% of the electrons from the electron guns. This yields lower screen luminance and lower contrast due to scattered electrons. As the screen size gets larger from 19 to 36 in. or more, additional temperature compensation methods must be used to reduce the effects of mask heating. Another method is using mask materials with lower coefficients of expansion such as INVAR. The phosphors used in the shadow mask color CRT are the red $Y_2O_3:S:Eu$ or, in some cases, the $Y_2O_3:Euphosphor$. The green phosphor is $ZnS:Cu$ and the blue phosphor is $ZnS:Ag$. Scale-up to size up to 40 in. as driving voltage needed in over 20,000 V. Shielding of high energy radiation also makes large size CRT very bulky and heavy in weight.

Vacuum Fluorescent Display

The VFD can be thought of as a low-voltage (12–40 V) CRT. Just as in the CRT, electrons from a cathode are controlled by a grid which allows them to be accelerated and strike the phosphor screen anode. In the operation of the VFD, electrons are emitted by the filament. The filament is a very thin tungsten wire coated with a tri-carbonate oxide (barium, strontium, and calcium

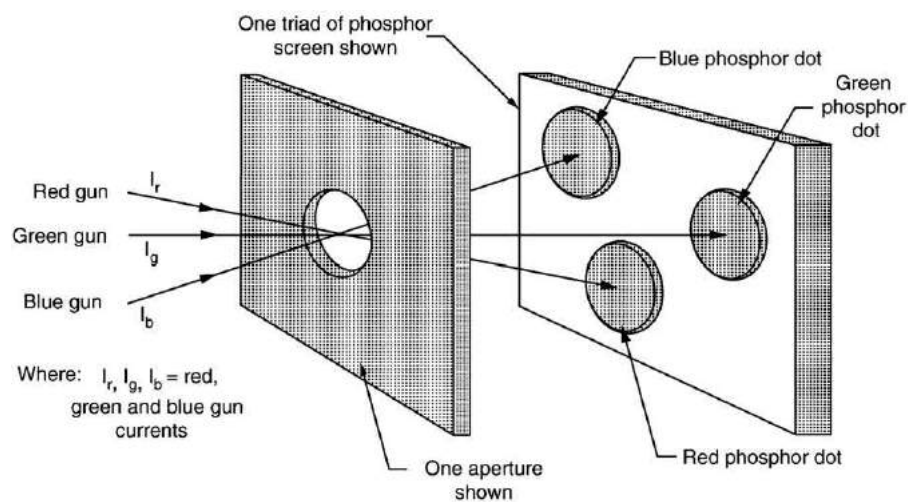


Fig. 3 The action of a shadow mask.



Fig. 4 A typical HDTV color CRT. Reproduced from Donofrio, R.L., 2002. The future of direct-view CRTs. SID Information Display Magazine 6, 10. Permission for reprint courtesy of the Society for Information Display.

carbonates) that are the same constituents of the CRT oxide cathode. By the application of a positive or negative voltage on this grid, electrons can either be cut off or accelerated to the phosphor on the anode. The VFD display is used in the auto industry and aging of the phosphor is very small, achieving 100,000 h to half of the initial luminance.

Field Emissive Displays

The FED is similar to the CRT and the VFD in that electrons from a source bombard a phosphor screen. In both the CRT and the VFD thermionic electrons are used. However, in the FED the electrons arise from field emission. Field emission devices create electrons though quantum mechanical tunneling from the emitter surface into the vacuum by the action of a large electrical field. This process is governed by the Fowler–Nordheim equation. The emitters can be of the Molybdenum Spindt type and also the poly-diamond or diamond-like carbon (DLC) coated emitters. We have the field enhancement with a sharp tip and reduced work function with the DLC. There has been renewed interest in FEDs, with the development of the “carbon nanotube” as an electron field emitting source. The carbon nanotube is a fullerene material, which is a distinct form of carbon. Fig. 5 shows that for this basic FED structure, the electrons from the Spindt emitter are controlled by the voltages of the anode, control electrode and emitter electrode.

Plasma Displays

Plasma displays used for television and information signs are photoluminescent displays. As shown in Fig. 6, the excitation of the plasma gas made of elements such as He, Ne, Ar, and Xe give rise to a glow discharge and the creation of vacuum ultraviolet photons that strike the screen and cause it to luminesce. In construction, the display is similar to the FED in that it is a matrix

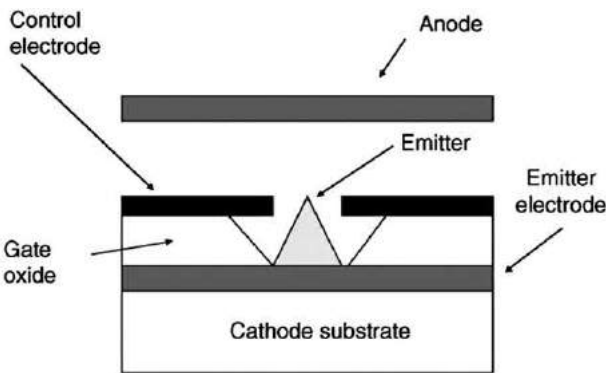


Fig. 5 Filed emission display cross-section. Reproduced from Green, P., Haven, D., 1998. Fundamentals of emissive displays. SID Short Course S-3. Permission for reprint courtesy of the Society for Information Display.

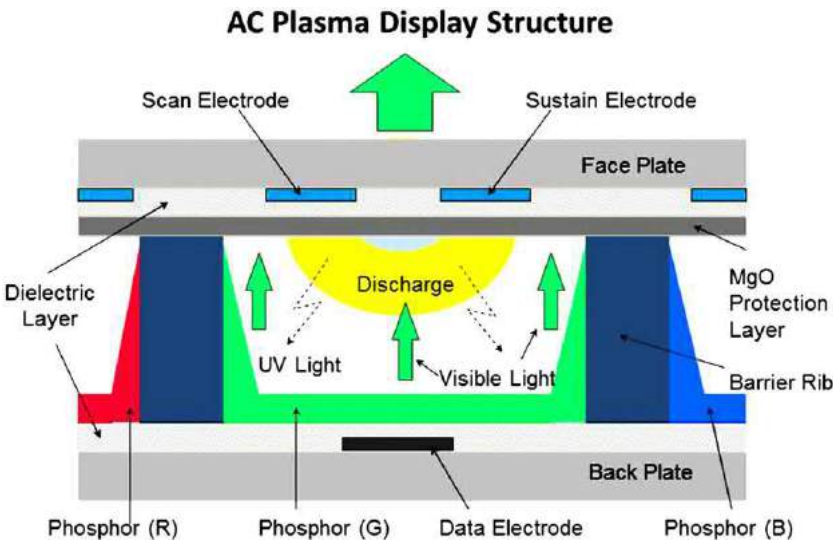


Fig. 6 A cross sectional view of typical AL plasma display.

addressed device. The AC design appears to be the more successful. Typical PDP phosphors for the triad are (Y,Gd)BO₃:Eu for the red, Zn₂ SiO₄:Mn for the green, and BaMgAl₁₀ O₁₇:Eu (called BAM for short) for the blue. Fig. 6 below shows the basic structure of an AC plasma display. One phosphor element is shown with barrier ribs, address and scan electrodes. The discharge of the plasma takes place in the region of the phosphor and excites the photoluminescent phosphor, which is shaped (in this example) along the side of the barrier ribs and over the address electrode.

Electroluminescence

Electroluminescence is the generation of (non-thermal) light by use of an electric field. High electric field EL units emit light through the impact of high energy electrons in luminescent centers in a semi-conducting material.

Inorganic LED Displays

Light-emitting diodes are being used as backlights for LCD displays, replacing some of the fluorescent light backlighting methods. Using a matrix array row and column connectors, we can digitally select a given color LED and pattern in an LED array and generate a color image. Red, green, and blue-UV/blue LEDs are now available, with an increasing demand for white LEDs.

Organic LEDs

The basic process of light generation in OLEDs is the formation of a P–N junction. Electrons injected from the N-type material and holes from the P-type material recombine at the junction and light is generated, as shown in Fig. 7. Additionally, OLEDs can be made on flexible (F) substrates and are called FOLEDs. The use of a transparent (T) cathode gives rise to a top emission OLED or a TOLED.

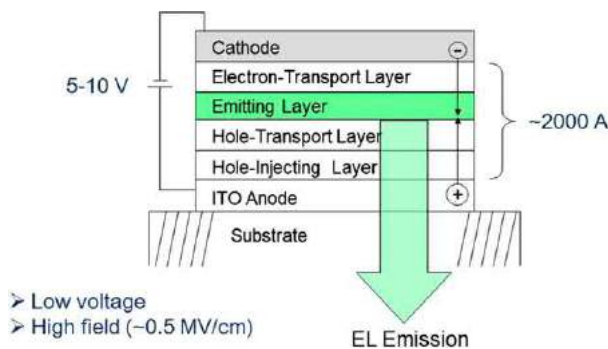


Fig. 7 A typical OLED configuration.

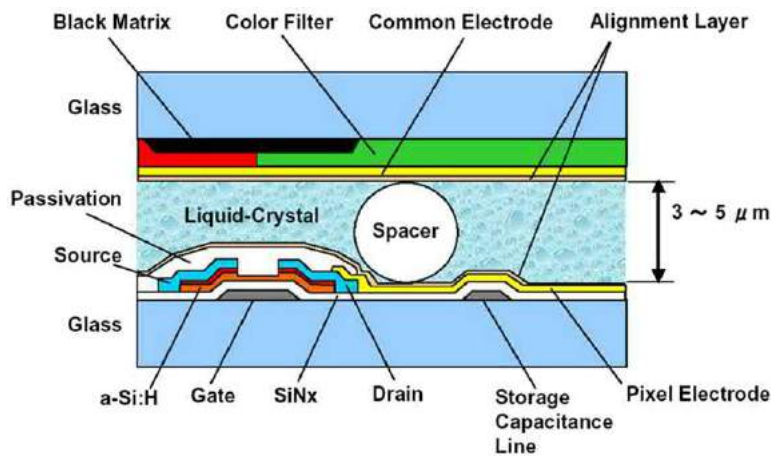


Fig. 8 A typical cross view of a TFT LCD panel.

Non-Emissive Displays – Active Matrix LCD – TFT

The active matrix liquid crystal display (AMLCD) device consists of a series of liquid crystal cells or windows whose characteristics can be controlled by an applied voltage to pass or restrict the passage of light through the windows via the external polarizers and the polarizing action of the LCD cells. AMLCD's are the dominated displays in the 21st century used almost in all area of mobile, laptop, monitors, TVs, and automotive displays (for driving information and entertainment). When the simple LCD is used without a transistor matrix, it is termed a passive display rather than an active one.

In order to show a color image, the LCD is designed to have three or more cells clustered per color pixel. The following (Fig. 8) is the transmission profiles of each of these red, green, and blue filters. The backlight used is a fluorescent lamp, but nowadays are white LED's or RGBW LED arrays.

Reflective LCD

For applications where there is suitable external light to view the display, no internal light source is used. Fig. 8 typical color LCD filter transmission characteristics. This reflective LCD consists of a polarizer and retarder plates stacked on the front side of the cell.

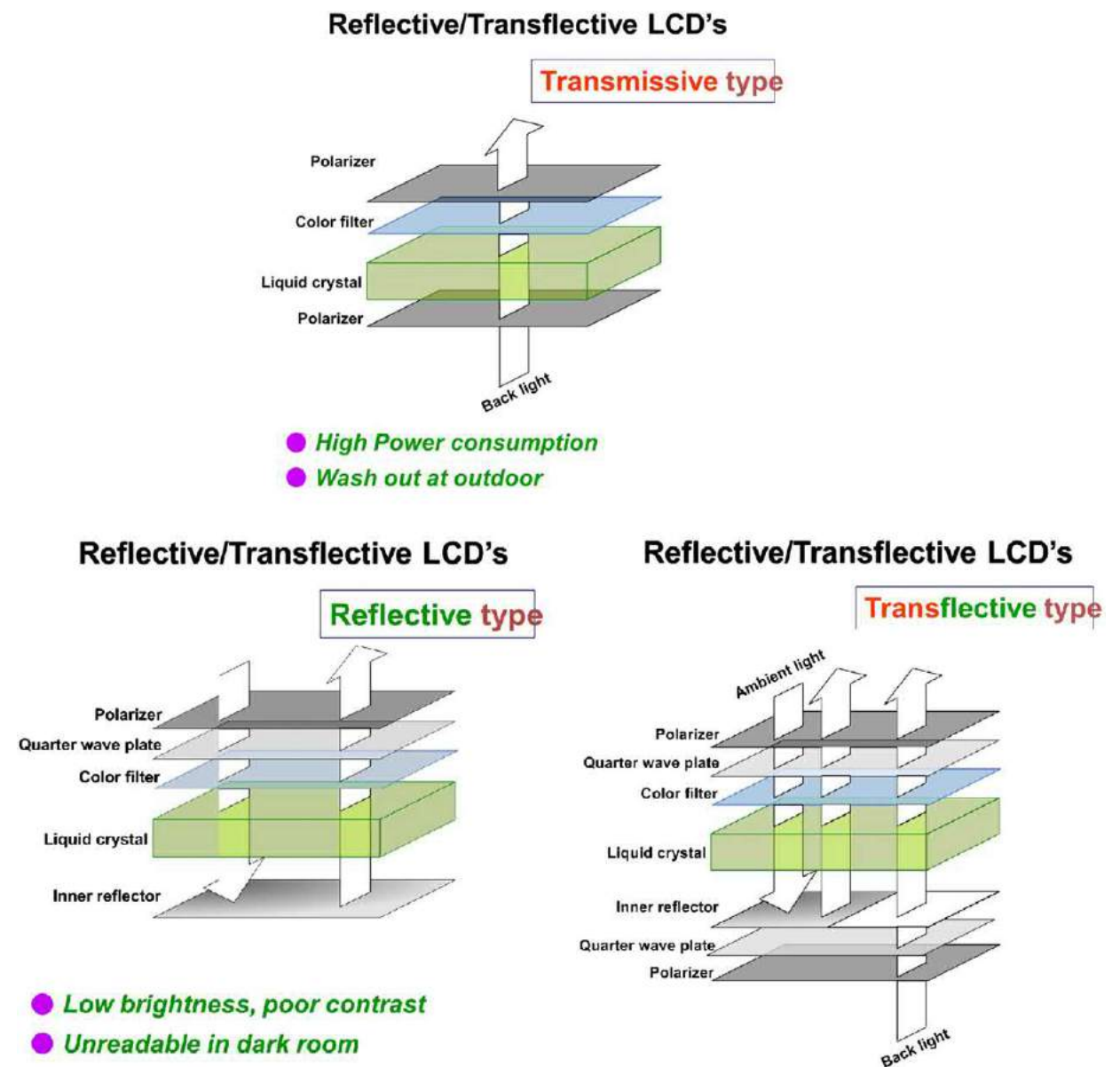


Fig. 9 Schematics of transmissive, reflective and transflective displays.

The cell with spacers contains the liquid crystal material and on the bottom of the cell is located a reflecting surface. It has been found that a dimpled reflecting surface makes a diffuse reflecting surface ([Fig. 9](#)).

Transflective LCD

The importance of the transflective device is that it has better contrast under high ambient illumination than the standard transmission back lighted LCD and fair contrast for low ambient illumination. The construction of this device consists of front and rear circular polarizers, back light, reflective elements, and novel filters. The back light is in the rear of the display. However, like the reflective LCD, internal reflectors are used on each pixel. The filters are designed spatially to be more transmissive for the reflective light than for the back light. Additionally the cell gap for the reflective light is smaller than the transmitted light region because the reflected light passes the cell twice whereas the transmitted light passes the cell once.

Further Reading

- de Gennes, P.G., Prost, J., 1993. *The Physics of Liquid Crystals*. Oxford: Clarendon Press.
- Fairchild, M.D., 2005. *Color Appearance Models*, second ed. John Wiley & Sons.
- Hecht, E., 1998. *Optics*, third ed. Addison Wesley Longman, Inc.
- Holliman, H., 2005. *3D Display System*.
- Lee, J.-H., Liu, D.N., Wu, S.-T., 2008. *Introduction to Flat Panel Displays*. 20. John Wiley & Sons.
- Shunsuke, Kobayashi, Mikoshiha, Shigeo, Lim, Sungkyoo (Eds.), 2009. *LCD Backlights* 21. John Wiley & Sons.
- Tsujimura, Takatoshi, 2017. *OLED Display Fundamentals and Applications*. John Wiley & Sons.
- Wu, S.T., Yang, D.K., 2001. *Reflective Liquid Crystal Displays*. SID Series in Displays Technology. Wiley and Sons Ltd.
- Yeh, Pochi, Gu, Claire, 2010. *Optics of Liquid Crystal Displays*. vol. 67. John Wiley & Sons.
- Yue Kuo, 2004. *Thin Film Transistors – Materials and Processes*. Volume 1: Amorphous Silicon Thin Film Transistors. Boston, MA: Kluwer Academic Publishers.

Liquid Crystal Physics and Materials

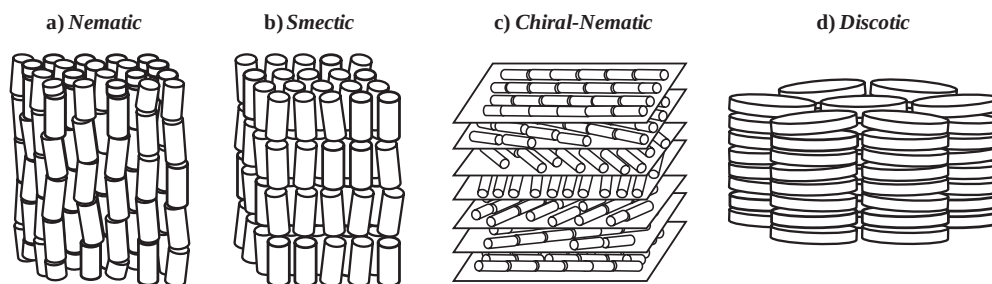
Huang-Ming Philip, National Chiao Tung University, Hsinchu, Taiwan

© 2018 Elsevier Ltd. All rights reserved.

Liquid crystals (LCs) possess a state of matter that has properties in between isotropic liquid and solid crystal phases. They can flow like a liquid, but simultaneously hold crystal-like molecular orientations. LC phases can be induced by various factors: the temperature variations, the level of solvent concentration, and the ratios of organic-inorganic composition. Three different types of LC phases are most mentioned in the literature: thermotropic, lyotropic, and metallotropic LC phases. The thermotropic LC phases exhibit a liquid-crystal phase transition as the temperature fluctuates. The lyotropic phases, on the other hand, show phase transitions as a function of both temperature and solvent concentration by amphiphilic molecules. The third type, metallotropic LCs, consist of organic and inorganic hybrid molecular system, which retain an LC transition that is relied not only on temperature variation and solvent concentration, but also on the ratio of organic-inorganic composition.

Thermotropic Liquid Crystal Phases

Thermotropic LCs indicate their corresponding mesophases at a certain temperature range. Two distinct characters, monotropic and enantiotropic mesophases, can be found during the thermal treatment. The monotropic LC mesophase is observed when the mesophase is shown upon cooling from the isotropic liquids. In contrast, the enantiotropic LC mesophase is defined when the mesophase appears in both heating from crystalline solid and cooling from isotropic liquid (de Gennes and Prost, 1993). Many thermotropic LCs exhibit not merely a single phase, but a variety of phases as temperature changed. For instance, a rod-like LC molecule (also called mesogen) may exhibit different smectic phases followed by the nematic phase as temperature increased on heating process. Various LC phases can be distinguished by their optical properties, or X-ray crystallography. The distinct textures, or defect domains under polarized optical microscope, are usually the fingerprint to trace their respective phases. X-ray crystallography is a useful tool to probe into the details of different smectic or discotic liquid crystal phases. In nematic phase, the amorphous character, unfortunately, can be observed in the small angle X-ray spectrum. Four different phases of thermotropic LCs are listed as followed.



Nematic Phase

Nematic phase has been widely studied and applied among other LC phases. The word “nematic” originates from the Greek ($\nu\eta\mu\alpha$ (nema), which means thread-like textures observing from the topological defects under polarized optical microscope. It was previously called Schlieren defects or disclination lines. The dark threads are line singularities which connect either two- or four-fold brushes point defects. In particular, only nematics have the distinct two-fold brushes. In the nematic phase, the calamitic (or rod-like) molecules can be self-assembled into a long-range directional order with their molecular long axes roughly parallel to each other without any positional order. As a result, the molecules can flow like a liquid with center mass positions of molecules randomly distributed, yet still keep their long-range directional order. In general, the mesogen is a rod-like molecule consists of rigid and flexible components. The rigid part can be aromatic rings, cyclic rings, or π -bonding structures. The flexible component is mostly aliphatic chain with different carbon chain length. The polar functional groups, such as nitrile (cyano-functional group, -CN) or fluoro- (-F), isothiocyanate (-NCS), can be attached in different positions to alter dielectric constants of mesogens. The optical axis in most nematics is uniaxial, except that a few nematic LCs with special molecular design are biaxial nematics. The nematic LCs can be aligned by an external magnetic or electric field. The alignment layer provided uniaxial molecular alignment in current LCDs is made by polyimide. The order parameter (S) is designed to determine the quality of LC molecular alignment and is usually defined based on the average of the second Legendre polynomial: $S = 1/2 [3 \langle \cos^2 \theta \rangle - 1]$, where θ is the angle between the liquid crystal molecular axis and the local director (i.e. preferred direction in liquid crystal media). Specifically, $S = 0$ represents a completely random and isotropic sample; and $S = 1$ is a perfectly aligned sample. S is in the order of 0.3–0.8 in general liquid crystal samples. For current LCDs applications, S needs to be at 0.6–0.7. The order parameter can be measured experimentally in a

number of ways, such as diamagnetism, birefringence, UV-vis spectroscopy, Raman scattering, NMR and EPR, etc. The UV-vis or IR linear dichroism would be the most effective tool in general.

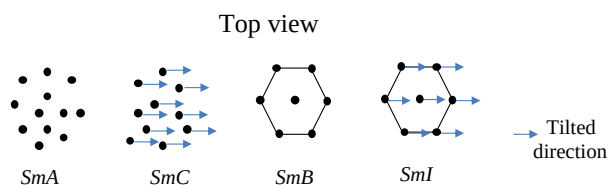
The theoretical treatments of LCs are mostly based on nematic LCs. Among them, the elastic continuum theory is a powerful method which is widely known for modeling LC devices under electric field. It discusses perturbations to a well-oriented continuum sample with strong boundary condition of parallel substrates. The distortions of the LC molecules in the continuum sample are described by the Frank free energy density. The response of the LC molecules can then be decomposed into three elastic constant terms, as the twist, splay, and bend distortions. The Maier-Saupe mean field theory provides a simple fluid nematic-isotropic phase transition model. Other models, such as Onsager hard-rod model and McMillan's model, address the Smectic A to nematic phase transition issues.

Nematic LCs are widely used as light modulators in the optical devices. Twist nematic LC device was first patented in 1970, which opened a gateway to the modern multi-billion LCD industry. The bulk molecular orientation following the direction of applied electric field is based on the dielectric constant of nematic LCs. In particular, the director of a positive dielectric constant material follows the applied electric field orientation, whereas the negative dielectric constant material stays perpendicular to the applied electric field. The in-plane-switching (IPS) or vertical alignment (VA) modes are developed for display applications, based on positive and negative dielectric constant LCs, respectively. Nevertheless, both positive and negative dielectric LCs can be utilized in the latest advance IPS mode (FFS mode) by manipulating the molecular alignment directions. The devices are noted as p-FFS or n-FFS mode to distinguish the difference. The p-FFS mode offers several attractive properties, such as high transmittance ($> 90\%$ transmittance at 5 V), low operation voltage, fast response time (12 ms at room temperature and 30 ms at -20°C), and no gray scale inversion. These advantages make p-FFS particularly desirable for outdoor applications. The n-FFS offers high transmittance over 93% in 1-dimensional (1D) pixel pattern and 85% in a 2D pixel pattern. The response time is slower than p-FFS mode (20 ms at room temperature). There is noticeable gray scale inversion which can be suppressed by 2D pixel pattern. The Gamma curve is single curve for the R, G, B color which make it easier for the driving design (Chen *et al.*, 2015).

Nematic liquid crystal mixtures are specially designed for display applications. Eutectic LC mixtures are prepared in order to accommodate wide temperature range to cover the device's application. The first commercial LC mixture E7, for instance, is made of four different molecules (3 biphenyl nitrile compounds and 1 terphenyl nitrile compound) in a fixed ratio. None of these four compounds has melting point below 24°C . The E7 mixture possesses nematic mesophase in the range of $-10 \sim 60.5^\circ\text{C}$. The lower temperature at -10°C indicates the eutectic point of this E7 mixture. For commercial nematic LC mixtures, the lower temperature can be reached around $-40 \sim -50^\circ\text{C}$ and the T_{ni} (nematic to isotropic temperature) is often varied from 80 to 120°C for display applications. In addition to application temperature, dielectric constant ($\Delta\epsilon$), birefringence (Δn), and rotational viscosity (γ_1), are the most important parameters to identify a proper LC for desired LC devices. In general, active-matrix (AM) MVA requires LC with $\Delta\epsilon = -3 \sim -5$, $\Delta n = 0.08 \sim 0.12$, and the rotational viscosity around 100 cp or lower. Conversely, the AM IPS mode requires the LC with $\Delta\epsilon = 7 \sim 11$, $\Delta n = 0.09 \sim 0.12$, and the rotational viscosity around 70 cp or lower. The faster rising and decay time could be achieved by using a thinner cell gap because the response time is at the ratio of cell gap squared ($\tau \propto d^2$). The response time can also be accelerated by warming up the device to elevated temperature, or by applying special driving voltage schemes (overdrive scheme). As the increasing demand of fast response time, Figure-of-Merit (FoM) grows to be one of the key parameters to evaluate the LC materials. The FoM is defined as $K(\Delta n)^2/\gamma_1$, where K is the elastic constant; Δn is the LC birefringence, and γ_1 is the rotational viscosity. High birefringence liquid crystal mixtures with low rotational viscosity are necessary for the high FoM LCs. They can be applied not only in the visible region, but in the infrared, GHz, or THz ranges. Gauza, *et al.* pointed out the importance of UV stability for the high birefringence liquid crystals. Mostly, the high birefringence liquid crystals consist of the rigid components, such as terphenyl or tolane moieties, which can enlarge its refractive index, but is susceptible to UV irradiation. Dąbrowski unveiled the material design and the corresponding birefringence (Dąbrowski *et al.*, 2013). The highest birefringence ($\Delta n = 0.79$) can be found in the LC molecule containing dibenzothiophene moiety. In addition, the highest Δn of isothiocyanate derivatives may reach 0.7 at 589 nm. As for nematic mixtures, the birefringence ($\Delta n = 0.5$) of LC mixtures can be well-prepared.

Smectic Phases

The smectic phases possess well-defined layers which are able to slide over one another as fluid. The word "smectic" originates from the Latin word "smecticus", meaning a soap-like texture. The smectic liquid crystals have positional order along with one long-range directional order. Since the layer structure is well-defined, any molecular tilt or distribution inside the layered structure is distinct to each other. For example, the molecules in Smectic A phase are oriented along the layer normal. In the same layer structure with molecules tilted away from the layer normal is called the Smectic C phase. Interestingly, the molecules self-aligned into a hexagonal packing and oriented along the layer normal is categorized as the Smectic B phase. As a result, various degrees of positional and directional orders are able to form different smectic phases.



The vast majority of LC research has been focused on chiral Smectic C for ferroelectric liquid crystals or scattering type Smectic A devices. The superior materials of Smectic LCs in the organic electronic devices have remained neglected for a while. Recently, the mobility of $13.9 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ has been reported from a solution prepared highly ordered Smectic E (SmE) mesophase in a bottom-gate bottom-contact FETs (Hsieh and Chen, 2015). Their intrinsic charge carrier mobility in the self-organized tilt orientation (SmE or SmI) is much superior to that of amorphous materials by two to three orders of magnitude.

Chiral Smectic and Nematic Phases

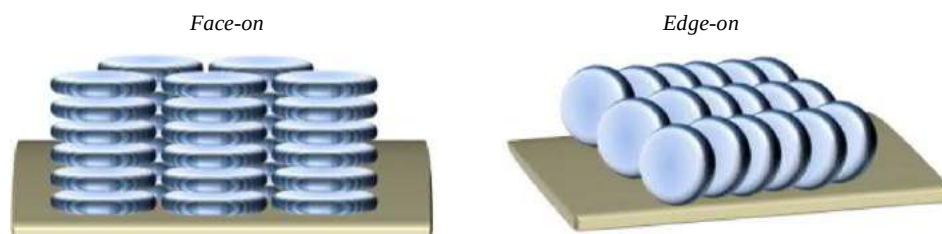
The chiral phases can be found in both nematic and smectic mesophases. The chiral-nematic phase exhibits chirality that was often called as the cholesteric phase as it was first observed from the cholesterol derivatives by Friedrich Reinitzer in 1888, and then confirmed in the publication by Otto Lehmann in 1889. This chiral-nematic phase exhibits a finite twisting angle between adjacent molecules perpendicular to the molecular director. The resulting helical layer structure is formed along the layer normal. Consisting of a helical stack of quasi-nematic layers, a molecular self-organization of chiral nematic structure is characterized by its helical pitch length p , and handedness. The pitch length p is defined as the distance over which the director rotates by 360° ; while the handedness describes the direction in which twisting of the nematic director occurs from one to another layer. The selective reflection in normal direction is described as $\lambda_R = p(n_e + n_o)/2$, in which n_e and n_o are the extraordinary and ordinary refractive indices, respectively. Consequently, the pitch and handedness of chiral-nematic liquid crystal systems can be manipulated to exploit unique optical properties, such as Bragg reflection, circularly polarized light, and low-threshold laser emission. The properties mentioned above can be utilized in a number of optical applications (Kopp *et al.*, 1998). Chiral-nematic LCs can be treated as 1-D photonic crystal when they aligned in the helical structure. Furthermore, Blue Phases (BPs) are found in the temperature range between chiral-nematic and isotropic liquid phases. They have three-dimensional cubic lattice structures assembled by the disclination line which is formed by crossing double helix twist cylinder structures. Coates and Gray (1973) first reported these thermodynamic phases in 1973. The theoretical predictions of the BPs hold icosahedral symmetry similar to quasicrystals were shown in 1981 (Seideman, 1990). The stabilization of blue phases over a temperature range of more than 60 K with room temperature (260–326 K) has been demonstrated in 2002 (Kikuchi *et al.*, 2002). The BPs stabilized at room temperature, that allows electro-optical switching with response times in the order of microsecond, opens the new path in fast light modulators or tunable photonic crystals for applications. The first Blue Phase mode LCD panel was developed in May 2008. Several drawbacks of blue phase LC for display applications were found in the high driving voltage, hysteresis, and light leakage in the dark state. Wu *et al.* has proposed a new electrode design to drastically reduce the driving voltage (Yoon *et al.*, 2010). A 12.7-in. LCD with 240 Hz frame rate color sequential display device was later realized by AUO in 2017 (Huang *et al.*, 2017). Its driving voltage was reduced to 15 V for single TFT addressing attributed to the protruded IPS electrodes structure. To solve hysteresis issue, Hsieh and Chen proposed using the fluoroacrylates monomers to ease the blue phase platelets texture formation under fast cooling and short annealing time before UV irradiation in 2015. The hysteresis was successfully suppressed and minimized (ΔV less than 0.8 V) in the PSBPLCs mixtures containing various fluoroacylates (Hsieh and Chen, 2015). Although numerous issues of BP LCD remain unsolved, its exciting features, such as sub-millisecond gray-to-gray response time, the alignment-free, optically isotropic voltage-off state, and large cell gap tolerance, etc. continue to attract scientists to explore new possibilities.

Chiral-smectic C (SmC*) phase exhibits a spontaneous electric polarization which was discovered by Meyer in the mid-1970s. The first surface-stabilized SmC* ferroelectric liquid crystal was realized by Clark and Lagerwall at the beginning of the 1980s (Clark and Lagerwall, 1980). Since then, the outstanding polar effects with sub-million second response time in chiral-smectic phases have drawn great attention in research, as well as seen in the current commercial micro-display products. The ferroelectric liquid crystals were studied extensively in various aspects, with much attention turning toward display application. FLC device is usually considered as bi-stable optical element traditionally. However, the continuous-director-rotation mode (CDR mode; or so-called half-V shape mode) FLC type does not have bi-stable character, which has the advantage of gaining undisturbed transmittance from temperature variation due to the suppression of SmA phase (Chen and Lin, 2009). The antiferroelectric phases are first called SmC* subphases because of their unclear properties and the smaller temperature range of stability. This phase becomes clear when the tilted molecular direction alternates from one layer to another. The fact that the molecular assembly structure is referred to anticlinic is to distinguish from the synclinic normal SmC phase, in which the director tilts in the same direction in adjacent layers. In addition to chiral smectic C, chiral smectic A can be tilted when electric field is applied to the system. The phenomenon is termed the “electroclinic effect” (Meyer and Pelcovits, 2002). The diffuse cone and the non-correlation models explain its field independence of birefringence. However, none of these models can fully describe the formation and temperature effects.

Discotic Phases

The first discotic liquid crystals (DLCs), benzene-hexa-*n*-alkanoates, were reported by Chandrasekhar *et al.* in the late 1970s. All kinds of discotic mesogens and metallomesogens have been reported and summarized (Chandrasekhar, 1993). In particular, disk-shape molecules orient themselves into a layer-like fashion known as columnar mesophases. The columns may be organized into rectangular or hexagonal arrays based on tilted structures of peripheral flexible chains attached to the aromatic cores. They are able to form nematic mesophase when the disk-shape mesogens do not stack into columns. Supramolecular organization of discotic

molecules can be controlled to give the macroscopic orientations either in planar uniaxial (face-on) or homeotropic alignment (edge-on) for different optoelectronic device applications as often seen in fast photoconductors for xerography, organic photovoltaics, organic light-emitting diodes, and chemical sensors. The negative birefringence property is unique in discotic liquid crystals. The hybrid orientation of discotic nematics film is applicable as optical compensation film improves the viewing angle of liquid crystal displays.



Discotic liquid crystals are novel organic semiconductor materials in the new generation. Interestingly, the mobility measured from FET transistors revealed that the higher values were found in crystalline phase right below discotic phase (van de Craats and Warman, 2001). A few critical reviews in 2007 have summarized the important aspects of recent researches in molecular design concepts, supramolecular structure, ordered thin films preparation, and electronic devices fabrication (Sergeyev *et al.*, 2007; Shimizu *et al.*, 2007).

Summary

The design and synthesis work of LCs were initially developed in depth by the Vorländer after Lehmann's report in 1889 and many others thereafter. The liquid crystals physics and materials still remained unpopular in early 20th century. Until the various device applications were demonstrated in the 1960s, the demands of novel LC materials and devices were sharply increased. The numerous varieties of physical and optical effects in liquid crystals have attracted many scientists and researchers to study the characteristics of LCs and apply them to various aspects of our lives. The fascinating anisotropic molecular self-assembly characteristics of liquid crystals have inspired many to manipulate their phases by energy, functional groups design, and surface alignments. The current interests in smectic liquid crystal research are focusing on de Vries/anti-de Vries type materials, tilting transitions without defect generation, and high-tilted anti-ferroelectric liquid crystals with perfect dark state and defect-free devices. In addition to these material developments, various functionalities can be integrated into different molecular assembly of LC mediums, such as organic solar cells, bio-sensors, chemical sensors, organic light-emitted diodes, etc. (Tyagi *et al.*, 2014; Popov *et al.*, 2016). It is worthy to further pursue the diverse liquid crystal systems from natural world and unfold many innovative possibilities in the future.

References

- Chandrasekhar, S., 1993. Discotic liquid crystals: A brief review. *Liq. Cryst.* 14, 3–14.
- Chen, H., Gao, Y., Wu, S.-T., 2015. n-FFS vs. p-FFS: Who wins? In: *SID Symposium Digest*, vol. 46, pp. 735–738.
- Chen, H.-M.P., Lin, C.-W., 2009. Free alignment defect, low driving voltage of half-V ferroelectric liquid crystal device. *Appl. Phys. Lett.* 95, 083501.
- Clark, N.A., Lagerwall, S.T., 1980. Submicrosecond bistable electrooptic switching in liquid crystals. *Appl. Phys. Lett.* 36, 899–901.
- Coates, D., Gray, G.W., 1973. Optical studies of the amorphous liquid-cholesteric liquid crystal transition: The "blue phase". *Phys. Lett.* 45, 115–116.
- Dąbrowski, R., Kula, P., Herman, J., 2013. High birefringence liquid crystals. *Crystals* 3, 443–482.
- de Gennes, P.G., Prost, J., 1993. *The Physics of Liquid Crystals*. Oxford: Clarendon Press.
- Hsieh, P.-J., Chen, H.-M.P., 2015. Hysteresis-free polymer stabilized blue phase liquid crystals comprising low surface tension monomers. *Liq. Cryst.* 42, 216–221.
- Huang, Y., Chen, H., Tan, G., *et al.*, 2017. New blue-phase liquid crystal optimized for color-sequential displays. In: *SID Symposium Digest*, vol. 48, pp. 486–489.
- Kikuchi, H., Yokota, M., Hisakado, Y., Yang, H., Kajiyama, T., 2002. Polymer-stabilized liquid crystal blue phases. *Nat. Mater.* 1, 64–68.
- Kopp, V.I., Fan, B., Viithana, H.K.M., Genack, A.Z., 1998. Low threshold lasing at the edge of a photonic stop band in cholesteric liquid crystals. *Opt. Lett.* 23, 1707–1709.
- Meyer, R.B., Pelcovits, R.A., 2002. Electroclinic effect and modulated phases in smectic liquid crystals. *Phys. Rev. E* 65, 061704.
- Popov, P., Honaker, L.W., Kooijman, E.E., Mann, E.K., Jákli, A.I., 2016. A liquid crystal biosensor for specific detection of antigens. *Sens. Bio-Sens. Res.* 8, 31–35.
- Seideman, T., 1990. The liquid-crystalline blue phases. *Rep. Prog. Phys.* 53, 659–705.
- Sergeyev, S., Pisula, W., Geerts, Y.H., 2007. Discotic liquid crystals: A new generation of organic semiconductors. *Chem. Soc. Rev.* 36, 1902–1929.
- Shimizu, Y., Oikawa, K., Nakayama, K.-i., Guillon, D., 2007. Mesophase semiconductors in field effect transistors. *J. Mater. Chem.* 17, 4223–4229.
- Tyagi, M., Chandran, A., Joshi, T., *et al.*, 2014. Self assembled monolayer based liquid crystal biosensor for free cholesterol detection. *Appl. Phys. Lett.* 104, 154104.
- van de Craats, A.M., Warman, J.M., 2001. The core-size effect on the mobility of charge in discotic liquid crystalline materials. *Adv. Mater.* 13, 130–133.
- Yoon, S., Kim, M., Kim, M.S., *et al.*, 2010. Optimisation of electrode structure to improve the electro-optic characteristics of liquid crystal display based on the Kerr effect. *Liq. Cryst.* 37, 201–208.

Further Reading

- Iino, H., Usui, T., Hanna, J., 2015. Liquid crystals for organic thin-film transistors. *Nat. Comm.* 6, 6828–7828.

TFT Materials and Devices

Po-Tsun Liu, National Chiao Tung University, Hsinchu, Taiwan

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Active-matrix (AM) displays have already accumulated unbeatable influence and created a brand-new life style in recent 20 years. In the active-matrix display operation scheme, thin film transistor (TFT) backplane technology is essential for LCD, OLED, eEPD, and even flexible displays. Taking AMLCD as an example, in each display pixel the liquid crystal layer functions as a light shutter, in which the switch is controlled by the TFT. TFT devices play a role of conducting electric current to charge the LC capacitor and storage capacitor. When the TFT device is operated at “on” state, the pixel signal voltage is delivered and charging the LC capacitors during tens of micron seconds (μs). With TFT operated at “off” state, pixel signal voltage will be stored in the LC capacitors and prevent it from being leaky for several mini seconds (ms). For AMOLED operation, TFT acts as an electric current driver to light “on” and “off” the OLED pixel besides a switch.

A TFT device structure comprises several material layers, including metallic gate electrode, gate dielectric film, semiconductor channel, and source/drain electrodes, which is similar to metal-oxide-semiconductor field effect transistor (MOSFET). In the turn-on operation, charged carriers (electron or hole) induced by gate bias voltage are conductive in the semiconductor channel and move from source to drain terminal under the influence of drain-to-source voltage. The carrier mobility is thereby dependent on the semiconductor channel layer, which can be classified into inorganic and organic materials. The inorganic semiconductor materials widely cover hydrogenated amorphous silicon (a-Si: H), polycrystalline silicon (poly-Si), and amorphous oxide semiconductors (AOSs), while the organic ones mainly include small or large molecular polymers according to the present technical development (Table 1). As the TFT device structure, there are three types of structures commercially fabricated: (1) bottom-gate back channel etch (BCE) structure, (2) bottom-gate etch stopper (ES) structure and (3) top gate (TG) structure (Fig. 1). The selected steps of TFT fabrication processes are schematically depicted as followed (Fig. 2). For realistic applications in flat-panel displays

Table 1 Comparison of TFT devices with different semiconductor layers

Characteristic	a-Si TFT	poly-Si TFT	TAOS TFT (transparent amorphous oxide semiconductor)	Organic TFT
Stability in ambient	Excellent	Excellent	Excellent	Poor
Uniformity	Good	V _{th} Dispersion	Good	V _{th} Dispersion
Performance	$\mu_n = 0.7$ ($\text{cm}^2/\text{V}\cdot\text{s}$)	$\mu_n = 200(\text{cm}^2/\text{V}\cdot\text{s})$ $\mu_p = 100(\text{cm}^2/\text{V}\cdot\text{s})$	$\mu_n = 1\text{--}100$ ($\text{cm}^2/\text{V}\cdot\text{s}$)	$\mu_p = 0.1\text{--}5$ ($\text{cm}^2/\text{V}\cdot\text{s}$)
Design/Device	NMOS	CMOS	NMOS	PMOS
Off current	low	high	lowest	fair
Processing temperature	Low	High	Low	Low
Flexibility	X	X	0	0
Transparent	X	X	0	0
Process	3–5 masks	7–9 masks (ELC + Implantation)	3–5 masks	3–5 masks

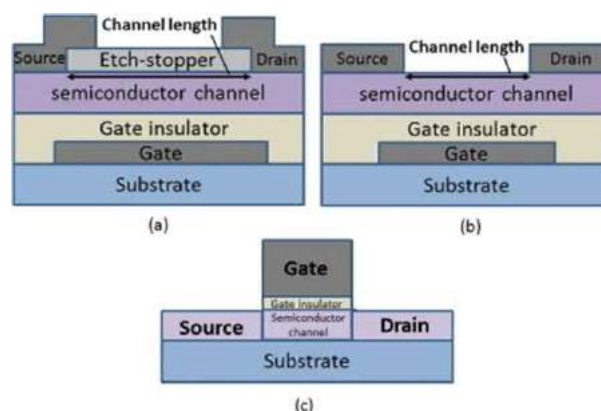


Fig. 1 TFT devices with (a) bottom-gate ES, (b) bottom-gate BCE, and (c) top-gate (TG) structure.

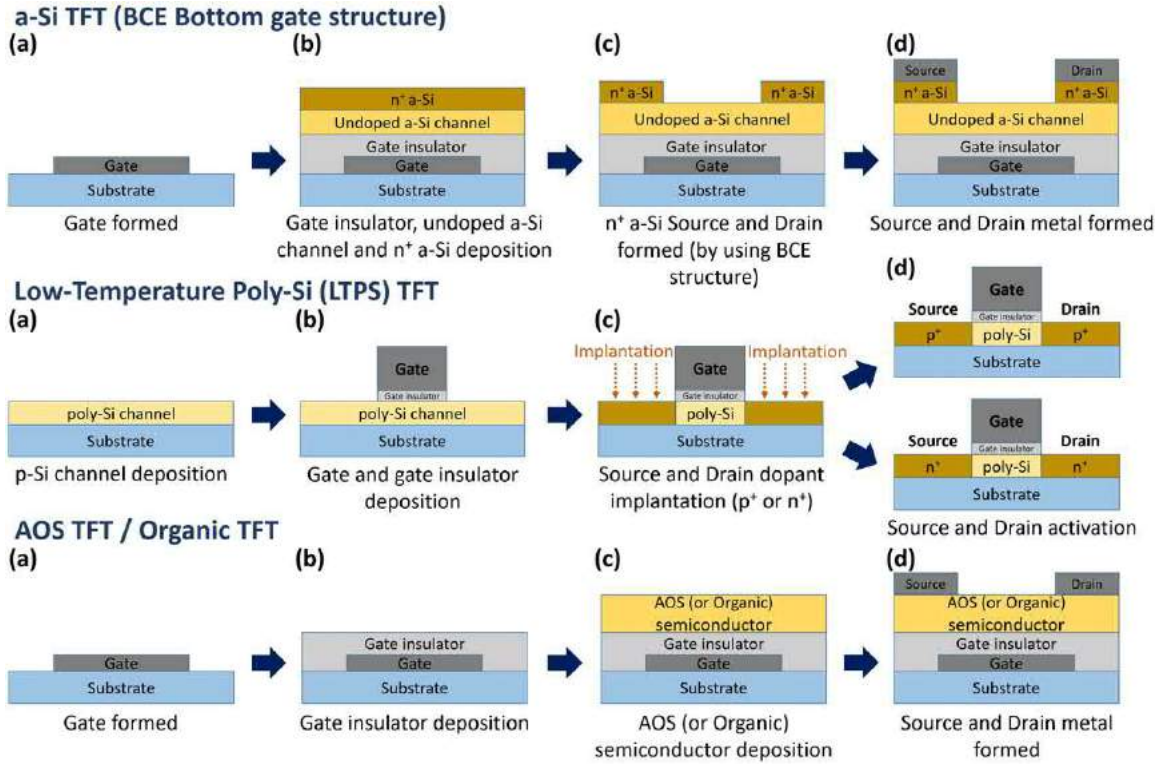


Fig. 2 Selected steps of fabrication processes for a-Si TFT, LTPS TFT, and AOS TFT/Organic TFT.

(FPDs), the TFT device has to meet several requirements, such as high turn-on current (drain current I_D), high field-effect mobility μ_{FE} (cm^2/Vs), low threshold voltage (V_{th}), low leakage current, low leakage current from gate terminal, low parasitic (overlap) capacitance, and low photosensitivity to ambient/backlight.

a-Si:H Thin Film and a-Si:H TFT

In the present FPD industry, the engineering of a-Si:H TFT is suitable for large-area and high-resolution AMLCD products. The plasma-enhanced chemical vapor deposition (PECVD) reactor is widely used in the deposition of a-Si:H thin film layer with the reactants of SiH_4 and H_2 gases around 300°C for TFT device fabrication. The network of a-Si:H thin film lacks long-range arrangement order so that bonding deviations occurs and perturbs the energy of sp^3 bonding and anti-bonding states in a specific co-valent bond compared to its crystalline case (*c*-Si). Therefore, a-Si:H film is full of a great deal of dangling bonds and weak Si-Si bonding associated with band tail electron states in energy band gap, leading to charge-trapping centers. Dangling bonds with electron energy states near Fermi level are the major contributor of these defects. Hydrogen is the most effective passivation agent for dangling bonds. Also, the uneven distribution of donor-like and acceptor-like trap states in energy band gap causes the slightly n-type behavior for an undoped a-Si:H film, which can be observed experimentally (Yue Kuo, 2004). The field-effect electron mobility is theoretically 1, because of the scattering effects of these band tail states.

There are two types of a-Si:H TFT devices adopted commercially in the present FPDs: the bottom-gate (inverted staggered) TFT with etch stop (ES) layer and backchannel etched (BCE) structure (Ban *et al.*, 1996; Maeda *et al.*, 1992). The BCE a-Si:H TFT reduces one photo-mask step for array fabrication compared to the ES a-Si:H TFT. Also, the gate insulator SiN_x , undoped a-Si:H channel, and n^+ a-Si:H layer as source/drain regions are deposited sequentially without unloading processes. However, the BCE a-Si:H TFT needs a thicker a-Si:H layer to have enough process margin for complete removal of n^+ a-Si:H on the top of the a-Si:H channel. Light illumination on a-Si:H will give rise to the decrease of both the dark conductivity and photoconductivity while the degradation can be recovered by a subsequent thermal annealing at around 200°C . In addition, the gate-to-source bias voltage applied to a-Si:H TFT will cause the increase of threshold voltage with increasing time. With a subsequent thermal annealing at 200°C for a few hours, the degraded characteristics of TFT can be restored to its original state (Yue Kuo, 2004; Street, 1991). Although a-Si:H TFT suffers several issues on electrical stability, the a-Si:H TFT array is nowadays a popular backplane technology in the large-area AMLCDs due to the features of a-Si:H film uniformity and low-cost fabrication.

Poly-Si Thin Film and Poly-Si TFT

Polycrystalline silicon thin-film transistors (poly-Si TFTs) have been widely used in many potential applications including pixel driving elements of active matrix organic light emitting diode (AM-OLED), and integrated peripheral driving circuits and addressing elements of active-matrix liquid crystal displays (AMLCDs) (Walker *et al.*, 2003; Shimoda *et al.*, 1998; Lewis *et al.*, 1990; Oshima and Morozumi, 1989). Recently, poly-Si TFTs are very attractive for system on top of the panel (SOP) as devices performances improve further. The degree of circuit integration will continue to increase as device performances improve further. The entire system will include memories, such as SRAM and nonvolatile FLSAH Memories, solar cells, and touch sensors as well as driver circuits for AMLCDs (Oh *et al.*, 2000; Young *et al.*, 1996; Cao *et al.*, 1994). Poly-Si TFTs have the potential advantages of silicon-on-insulator (SOI) MOSFETs such as simple fabrication process, good device-to-device isolation, high circuit density, and high device performance as well as the possibility to be applied in vertical 3-D integration. Poly-Si TFTs technology has been receiving more attention because it is a promising mean of achieving 3-D integration, which has been utilized in various 3-D circuits (Hayashi *et al.*, 1996; Kuriyama *et al.*, 1998; Wang *et al.*, 2000). A grain boundary is the interface between two grains, or crystallites, in a polycrystalline material. However, grain boundaries of poly-Si material cause trap and tail states which put strong influences on device characteristics including an increase in threshold voltage (V_{TH}), decrease in mobility, increase in off-state leakage current, decrease in on-state current, poor subthreshold characteristics and degradation in device reliability (Matore, 1984; Levinson *et al.*, 1982). The trap states in the grain boundaries is thought to be associated with the presence of dangling and strained bonds (Lewis *et al.*, 1989). To solve these problems, some crystallization methods have been introduced to enlarge the grain size. Poly-Si thin film with large crystalline grains can be obtained using a variety of techniques: solid phase crystallization (SPC) (Hatalis and Greve, 1988), laser crystallization (LC) (Smith *et al.*, 1997) and metal-induced lateral crystallization (MILC) (Lee *et al.*, 1995; Lee and Joo, 1996; Jin *et al.*, 1998) of amorphous silicon (*a*-Si). Because crystallization of poly-Si channel plays a main role in carrier mobility and uniformity of poly-Si channel, a robust crystallization method is required for poly-Si TFTs in future various applications.

SPC is a classical method of achieving poly-Si by furnace at 600°C for 24 h. SPC produces a large-grain poly-Si, but its crystallization temperature is too high for the ordinary glass substrate. Laser crystallization (LC) method is popular for low temperature poly-Si (LTPS) TFTs. A new low temperature (500°C) crystallization method for poly-Si TFTs was developed: Metal-Induced Lateral Crystallization (MILC). The *a*-Si film in the channel area of a TFT was laterally crystallized from the source/drain area, on which an ultrathin nickel layer was deposited before annealing. The enhanced crystallization rate at this low temperature has been attributed to the silicate formation of $NiSi_2$. This low-temperature budget process with low metal contamination may be used to overcome problems associated with the SPC approach (Lewis *et al.*, 1989). Further improvement in the poly-Si TFTs performances has been recently obtained by the introduction of excimer laser crystallization (ELC), leading to electron field-effect mobility $> 100 \text{ cm}^2/\text{V s}$. Excimer lasers emit in the UV region (output wavelengths 193 nm, 248 nm and 308 nm for ArF, KrF and XeCl gas mixtures, respectively) with a short pulse duration (10–30 ns). The combination in Si of strong optical absorption of the UV light and small heat diffusion length during the laser pulse implies that high temperatures can be developed in the Si-surface region, causing melting, without appreciable heating ($< 400^\circ\text{C}$) of the substrate. This makes the ELC-process compatible with glass substrates, one of the major advantages of this technique. Such achievements have finally opened the door to the fabrication of complex CMOS circuits based on poly-Si TFTs and some major display manufacturers have already on the market AMLCDs with integrated polysilicon TFTs drivers (Smith *et al.*, 1997). The technologies of poly-Si TFTs will become more and more important in future applications including 3D integration of active devices and SOP. More researches investigate the related new technologies and underlying mechanisms in thin-film devices with scaling down dimension are worthy to study.

AOS Thin Film and AOS TFT

With the high demand for high-resolution image ($> 2000 \times 4000$ pixels) and fast-frame rate ($> 120 \text{ Hz}$), *a*-Si:H TFT is not an appropriate device for next generation FPD displays due to its low mobility ($< 1 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$) and high photosensitivity (low band gap about 1.7 eV). On the other hand, poly-Si TFT which has high mobility ($> 100 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$) suffers from a problem with the variation of electrical properties due to grain boundary. Amorphous oxide semiconductors (AOS) including InGaZnO (Nomura *et al.*, 2004), InZnSnO (Fuh *et al.*, 2014), AlZnSnO (Teng *et al.*, 2013), InZnO (Liu *et al.*, 2009) and InWO (Liu *et al.*, 2016a), are materials with high electron mobility larger than $10 \text{ cm}^2/\text{Vs}$ and high transmission about 80% in visible portion of the electromagnetic spectrum even when their films are deposited at room temperature. The high transparency of AOSs enables the realization of emerging transparent display technology, in which users can be allowed to view what is shown on a glass video screen while still being able to see through it. Among AOS materials, amorphous In-Ga-Zn-O (*a*-IGZO) has been attractive. For the In_2O_3 - Ga_2O_3 -ZnO ternary system, the incorporation of cations with large ionic valance such as Ga^{3+} to high conductive oxides such as In_2O_3 and ZnO is effective to control the carrier concentration due to their strong metal-oxygen bonds (Nomura *et al.*, 2004). The mobility of InGaZnO_4 is primary determined by the In_2O_3 content fraction because only In^{3+} meets the electron configuration criterion $(n-1)d^{10}ns^0$ ($n \geq 4$) of heavy post transition metal cation for ionic AOS (IAOS) among the three cations. The large ionic valence ion of Ga^{3+} combines with high conductive oxides such as In_2O_3 and ZnO to control the carrier concentration effectively because of the strong metal-oxygen bonds. Ga^{3+} also suppresses carrier generation via oxygen vacancy formation because Ga ion forms stronger bond with oxygen than Zn and In ions (Nomura *et al.*, 2004).

AOS thin films can be formed by a variety of thin film deposition technologies, including sputtering, pulse laser deposition (PLD), and solution-based deposition methods (Fortunato *et al.*, 2012). For the large-area manufacturing, the sputtering technology is favorable. Three types of AOS TFT device aforementioned can be adopted for FPD applications. There are many factors for the electrical stability of AOS TFT devices, such as the oxide semiconductor material, the material composition, content of oxygen deficiency, the device structure, the gate insulator, the environment for stressing, the electrical biasing conditions, and the intensity of the illuminated light. AOS TFTs are generally sensitive to the oxygen or moisture during the bias stability test, resulting in threshold voltage shift. Post-deposition annealing process can be implemented to enhance the resistance of AOS channel to electrical stress and oxygen/moisture uptake from environment (Fuh *et al.*, 2014). The annealing process provides the energy for structure relaxation of AOS thin film, which reduces the content of oxygen deficiencies in AOS channel. Electrical trap states are originated from the oxygen deficiencies. A better transfer characteristic and electrical stability of AOS TFT can be resultantly achieved since both trap states in the AOS channel and at the interface of AOS/gate insulator are reduced effectively by post-deposition annealing. On the other hand, the channel passivation methods can be used to isolate the absorption oxygen on the channel surface and enhance electrical stability (Park *et al.*, 2008a,b; Pei *et al.*, 2009; Liu *et al.*, 2016b). Novel AOS materials and manufacture processes have been proposed for overcoming many critical challenges. It is a promising technology for next-generation FPD displays.

Organic Thin Film and OTFT

Organic semiconductors (OSCs) are another TFT device technology actively studied over the past two decades. Organic electronics are lighter, more flexible, and potentially lower manufacture cost than their inorganic counterparts. Steady improvement in the electrical performance and the stability of OSCs led the technology development to utilize various flexible substrates, such as glass, plastic, and even paper (Kim *et al.*, 2004; Moore, 2002; Klauk *et al.*, 2003; Tsumura *et al.*, 1986; Horowitz, 1998; Lee *et al.*, 2006; Schon and Batlogg, 2001; Lee and Subramanian, 2003). The organic semiconductors that are used are pentacene, poly (3-hexylthiophene) (P3HT), poly (3-alkylthiophene) (P3AT), and poly (3-octylthiophene) (P3OT), which can be deposited by vacuum evaporation and solution processing techniques (Shekar *et al.*, 2004). Their performance critically depends on the efficiency with which charge carriers (electrons and/or holes) move within the π -conjugated semiconductor materials. Most of the reported organic TFTs (OTFTs) were built as the bottom gate structure for the concern in deposition of active material on the insulator instead of at the bottom. Two categories of gate dielectrics are commonly used in organic transistors: inorganic and organic. Examples of inorganic dielectrics include silicon dioxide and silicon nitride, typically used in traditional silicon integrated circuits and a-Si:H TFT display backplanes. The solution processable organic dielectrics, Poly-vinylphenyle (PVP) and polymethylmethacrylate (PMMA) are typical candidates for organic gate dielectrics (Shekar *et al.*, 2004). The reported field-effect mobility of vacuum evaporation deposited pentacene-based TFTs have been quoted in the range of 0.1–1.5 cm²/Vs which is comparable to devices using a-Si:H as the semiconductor material. However, OTFTs typically suffer from several issues including high-voltage operation, and thermal stability. The proceeding development in the enhancement of field-effect mobility and thermal stability are still in progress.

Summary

TFT is a critical device to switch a pixel on and off in FPD technology. Nowadays, a-Si:H TFT array is mainly adopted as a backplane for large-area display applications. To cope with the increasing demand of image resolution with high frame rate, poly-Si TFT with high-field-effect mobility and driving current is gradually and extensively required in FPD industry. Additionally, AOS TFTs and organic TFTs have been studied as possible alternatives for Si-based TFTs to lower production costs and processing temperature for flexible display applications. At present, leading manufacturers have already demonstrated many prototypes of AMLCD and AMOLED panels with AOS TFT backplanes. Also, some of them have announced the mass production for innovative applications.

References

- Ban, A., Nishioka, Y., Shimada, T., Okamoto, M., Katayama, M., 1996. A simplified process for SVGA TFT-LCDs with single-layered ITO source bus-lines. *SID 96 Digest*, pp. 93–96.
- Cao, M., Zhao, T., Saraswat, K.C., Plummer, J.D., 1994. A simple EEPROM cell using twin polysilicon thin film transistor. *IEEE Electron Device Lett.* 15, 304–306.
- Fortunato, E., Barquinha, P., Martins, R., 2012. Oxide semiconductor thin-film transistors: A review of recent advances. *Adv. Mater.* 24, 2945–2986.
- Fuh, C.S., Liu, P.T., Huang, W.H., Sze, S.M., 2014. Effect of annealing on defect elimination for high mobility amorphous indium-zinc-tin-oxide thin-film transistor. *IEEE Electron Device Lett.* 35 (11), 1103–1105.
- Hatalis, M.K., Greve, D.W., 1988. Large grain polycrystalline silicon by low-temperature annealing of low-pressure chemical vapor deposited amorphous silicon films. *J. Appl. Phys.* 63, 2260–2266.
- Hayashi, F., Ohkubo, H., Takahashi, T., *et al.*, 1996. A highly stable SRAM memory cell with top-gated P–N drain poly-Si TFT's for 1.5 V operation. In: *IEDM Tech. Dig.*, pp. 283–286.

- Horowitz, G., 1998. Organic field-effect transistors. *Adv. Mater.* 10 (5), 365–377.
- Jin, Z., Bhat, G.A., Yeung, M., Kwok, H.S., Wong, M., 1998. Nickel induced crystallization of amorphous silicon thin films. *J. Appl. Phys.* 84 (1), 194–200.
- Kim, J.H., Moon, D.G., Han, J.I., 2004. Organic TFT array on a paper substrate. *IEEE Electron Device Lett.* 25 (10), 702–704.
- Klauk, H., Halik, M., Zschieschang, U., *et al.*, 2003. Pentacene organic transistors and ring oscillators on glass and on flexible polymeric substrates. *Appl. Phys. Lett.* 82 (23), 4175–4177.
- Kuriyama, H., Ishigaki, Y., Fujii, Y., *et al.*, 1998. A C-switch cell for low-voltage and high-density SRAM's. *IEEE Trans. Electron Devices* 45, 2483–2488.
- Lee, J.B., Subramanian, V., 2003. Organic transistors on fiber: A first step toward electronic textiles. In: *IEDM Tech. Dig.*, pp. 8.3.1–8.3.4.
- Lee, K.S., Smith, T.J., Dickey, K.C., *et al.*, 2006. High resolution characterization of pentacene/polyaniline interfaces in thin-film transistors. *Adv. Funct. Mater.* 16 (18), 2409–2414.
- Lee, S., Jeon, Y., Joo, S., 1995. Pd induced lateral crystallization of amorphous Si thin film. *Appl. Phys. Lett.* 66, 1671–1673.
- Lee, S.W., Joo, S.K., 1996. Low temperature poly-Si thin-film transistor fabrication by metal-induced lateral crystallization. *IEEE Electron Device Lett.* 17, 160–162.
- Levinson, J., Shepherd, F.R., Scanlon, P.J., *et al.*, 1982. Conductivity behavior in polycrystalline semiconductor thin film transistors. *J. Appl. Phys.* 53, 1193–1202.
- Lewis, A.G., Huang, T.Y., Wu, I.W., Bruce, R.H., Chiang, A., 1989. Physical mechanisms for short channel effect in polysilicon thin film transistors. In: *IEDM Tech. Dig.*, pp. 349–352.
- Lewis, A.G., Wu, I.-W., Huang, T.Y., Chiang, A., Bruce, R.H., 1990. Active matrix liquid crystal display design using low and high temperature processed polysilicon TFTs. In: *IEDM Tech. Dig.*, San Francisco, CA, pp. 843–846.
- Liu, P.T., Chang, C.H., Chang, C.J., 2016a. Suppression of photo-bias induced instability for amorphous indium tungsten oxide thin film transistors with bi-layer structure. *Appl. Phys. Lett.* 108 (26), 261603.
- Liu, P.T., Chang, C.H., Fuh, C.S., Liao, Y.T., Sze, S.M., 2016b. Effects of nitrogen on amorphous nitrogenated InGaZnO (a-IGZO: n) thin film transistors. *IEEE/OSA Journal of Display Technology* 12 (10), 1070–1077.
- Liu, P.T., Chou, Y.T., Teng, L.F., 2009. Environment-dependent metastability of passivation-free InZnO TFT after gate bias stress. *Appl. Phys. Lett.* 95, 233504.
- Maeda, H., Fujii, K., Yamagishi, N., *et al.*, 1992. A 15-in.-diagonal full-color high-resolution TFT-LCD. *SID 92 Digest*, pp. 47–50.
- Matare, H.F., 1984. Carrier transport at grain boundaries in semiconductors. *J. Appl. Phys.* 56, 2605–2631.
- Moore, S.K., 2002. Just one word: Plastics. *IEEE Spectrum* 39 (9), 55–57.
- Nomura, K., Ohta, H., Takaji, A., *et al.*, 2004. Room-temperature fabrication of transparent flexible thin-film transistors using amorphous oxide semiconductors. *Nature* 432, 488–492.
- Oh, J.H., Chung, H.J., Lee, N.I., Han, C.H., 2000. A high-endurance low-temperature polysilicon thin-film transistor EEPROM cell. *IEEE Electron Device Lett.* 21, 304–306.
- Oshima, H., Morozumi S., 1989. Feature trends for TFT integrated circuits on glass substrates. In: *IEDM Tech. Dig.*, Washington, DC, p. 157.
- Park, J., Kim, S., Kim, C., *et al.*, 2008b. High-performance amorphous gallium indium zinc oxide thin-film transistors through N₂O plasma passivation. *Appl. Phys. Lett.* 93, 053505.
- Park, W.J., Shin, H.S., Du Ahn, B., *et al.*, 2008a. Investigation on doping dependency of solution-processed Ga-doped ZnO thin film transistor. *Appl. Phys. Lett.* 93, 083508.
- Pei, Z.L., Pereira, L., Goncalves, G., *et al.*, 2009. Room-temperature cosputtered HfO₂ – Al₂O₃ multicomponent gate dielectrics. *Electrochem. Solid State Lett.* 12, G65.
- Schon, J.H., Batlogg, B., 2001. Trapping in organic field-effect transistors. *J. Appl. Phys.* 89 (1), 336–341.
- Shekar, B.C., Lee, J., Rhee, S., 2004. Organic thin film transistors: Materials, processes, and devices. *Korean J. Chem. Eng.* 21 (1), 267–285.
- Shimoda, T., Ohshima, H., Miyashita, S., *et al.*, 1998. High resolution light emitting polymer display driven by low temperature polysilicon thin film transistor with integrated driver. In: *Proceedings ASID*, Seoul, Korea, pp. 217–220.
- Smith, P.M., Carey, P.G., Sigmon, T.W., 1997. Excimer laser crystallization and doping of silicon films on plastics substrates. *Appl. Phys. Lett.* 70, 342–344.
- Street, R.A., 1991. Hydrogen diffusion and electronic metastability in amorphous silicon. *Physica B* 170 (69),
- Teng, L.F., Liu, P.T., Wang, W.Y., 2013. Electrical performance enhancement of Al-Zn-Sn-O thin film transistor by supercritical fluid treatment. *IEEE Electron Device Lett.* 34 (9), 1154–1156.
- Tsumura, A., Koezuka, H., Ando, T., 1986. Macromolecular electronic device: Field effect transistor with a polythiophene thin film. *Appl. Phys. Lett.* 49 (18), 1210–1212.
- Walker, A.J., Nallamothu, S., Chen, E.H., *et al.*, 2003. 3D TFT-SONOS memory cell for ultra-high density file storage applications. In: *VLSI Symp. Tech. Dig.*, pp. 29–30.
- Wang, H., Chan, M., Jagar, S., Wang, Y., Ko, P.K., 2000. Submicron super TFTs for 3-D VLSI applications. *IEEE Electron Device Lett.* 21 (9), 439–441.
- Young, N.D., Harkin, G., Bunn, R.M., McCulloch, D.J., French, I.D., 1996. The fabrication and characterization of EEPROM arrays on glass using a low-temperature poly-si TFT Process. *IEEE Trans. Electron Devices* 43, 1930–1936.
- Yue Kuo, , 2004. *Thin Film Transistors-Materials and Processes: Amorphous Silicon Thin Film Transistors*. 1. Boston: Kluwer Academic Publishers.

Further Reading

Fortunato, G., Mariucci, L., Carluccio, R., Pecora, A., Foglietti, V., 2000. Excimer laser crystallization techniques for polysilicon TFTs. *Applied Surface Science* 154, 95–104.

Color Formation of LCD – Spatial and Temporal

Fang-Cheng Lin, National Chiao Tung University, Hsinchu, Taiwan

© 2018 Elsevier Ltd. All rights reserved.

Structure of the Human Eye

The human eye is an adaptive lens. Fig. 1 shows the cross-section of the human eye (Webvision, 2017). Light goes into the pupil of the eye and continues through the lens which acts the same as a camera focusing light onto the retina. The retina contains two types of photoreceptors, rods and cones, named for their shapes. The rods are more numerous, roughly 125 million, and are more sensitive than the cones to detect small amounts of light, such as starlight. However, they are unable to distinguish color. Six to seven million cones perform in bright light, giving detailed colored views, but they fail at low light levels (Hecht, 1998). Cones are much more concentrated in the central yellow spot known as the “macula.” In the center of the macula is the “fovea centralis,” a 0.2 mm diameter rod-free area with very thin, densely packed cones. Also, the fovea is the area of the eye responsible for sharpest vision.

The distribution of rods and cones in the retina is not uniform, as illustrated in Fig. 2 (Osterberg, 1935). The density of cones falls rapidly to a constant level at about 10–15 degrees from the fovea. At about 15–20 degrees from the fovea, the density of the rods reaches a maximum. Because there is only one pigment type for rods, humans only see objects as shades of gray. In contrast, there are three types of cones which refer to as the short-wavelength (S-cone), middle-wavelength (M-cone), and long-wavelength (L-cone) sensitive cones, as shown in Fig. 3 (CIE 170-1:2006, 2017). This is why humans perceive vivid colors. While receiving light, the retina converts it into nerve signals, and then the nerve signals are sent through the optic nerve, at the back of the eye, to the brain. Hence, in the optical nerve region which located about 13–18 degrees of the retina, there is no any receptor to sense light stimuli and is so-called “blind spot.”

Color Matching Functions and Tristimulus

A color matching experiment was conducted in 1931 to describe the color that the human eye perceives, as illustrated in Fig. 4. Based on this study, a set of color matching functions (CMFs): $\bar{r}(\lambda)$, $\bar{g}(\lambda)$, and $\bar{b}(\lambda)$ was derived as shown in Fig. 5(a) (UPENN, 2017). The CMFs were transformed to $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, and $\bar{z}(\lambda)$ using mathematics to avoid the negative values appearing on the CMFs, as shown in Fig. 5(b). This is the well-known CIE XYZ color space established by the Commission Internationale de l’Éclairage (CIE) in 1931. Consequently, the color perceived can be quantified by the CIE as three tristimulus: X, Y, and Z. The color (X, Y, Z) is affected by three components: illuminant (P), an object’s reflectance factor (R), and the CMFs of the human eye, ($\bar{x}(\lambda)$, $\bar{y}(\lambda)$, and $\bar{z}(\lambda)$), and is given in Eq. (1).

$$X = \int P(\lambda)R(\lambda)\bar{x}(\lambda)d\lambda; Y = \int P(\lambda)R(\lambda)\bar{y}(\lambda)d\lambda; Z = \int P(\lambda)R(\lambda)\bar{z}(\lambda)d\lambda \quad (1)$$

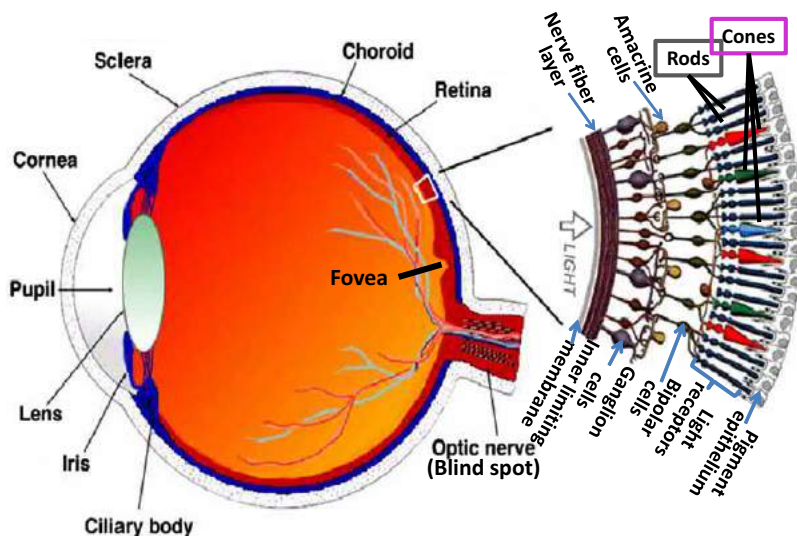


Fig. 1 A drawing of a cross-section through the human eye with a schematic enlargement of the retina including rod and cone light receptors (Webvision, 2017). The photoreceptors, cones and rods, receive the incident light and then send the response signals to ganglion cell, and finally optic nerves collect the signals to the brain.

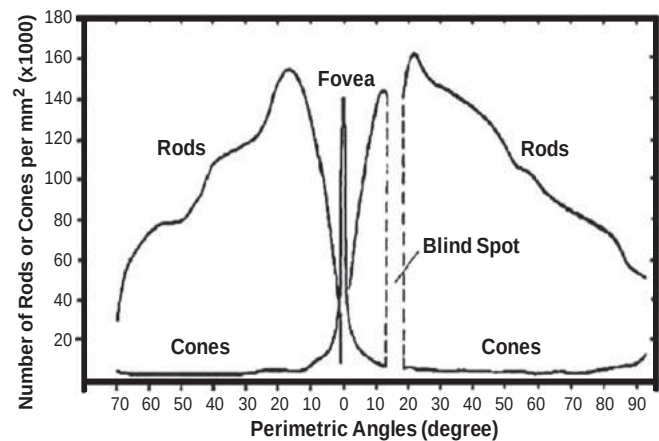


Fig. 2 Distribution of rods and cones along a line in the retina (Osterberg, 1935). The cones are much more concentrated in the macula, with an extremely high density in the center of the macula-fovea centralis. Oppositely, rods widely spread at the retina except for the center of the fovea, with a sharp decline in the fovea. An area, located about 13–18 degrees of the retina, existing no photoreceptor cells is the blind spot, where are full of the optic nerves and called optic disc and blind spot.

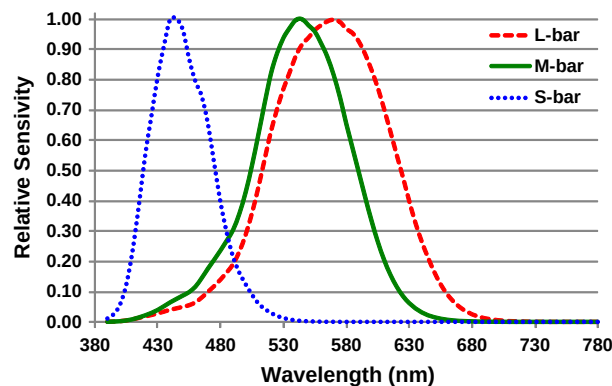


Fig. 3 Normalized response sensitivity of three cone-types – L-, M-, and S-cones, named after their responses at long, medium, and short wavelengths (CIE 170-1:2006, 2017).

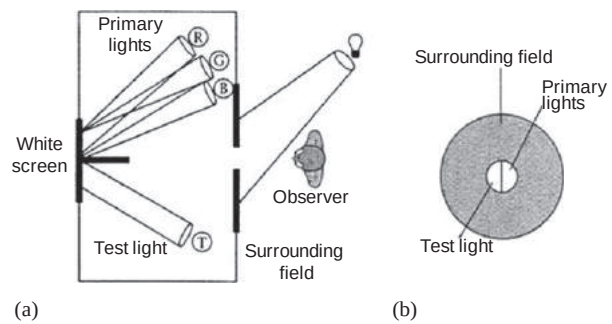


Fig. 4 Color matching experiment. (a) The observer sees a white screen divided into two halves, one-half illuminated by a test light, the other half illuminated by a mixture of one or more primary lights (here three are shown). (b) The observer can adjust the intensity of the primary light(s) to try to make the two halves of the screen match (UPENN, 2017).

Colors in Liquid Crystal Displays

Liquid crystal displays (LCDs) have become the dominant display technology for applications ranging from small mobile devices, notebooks, car navigation systems up to large-size television sets. This high penetration level can be attributed to a host of advantageous features, including high resolution, high brightness, light-weight, and a thin form factor. In an LCD, a vivid color

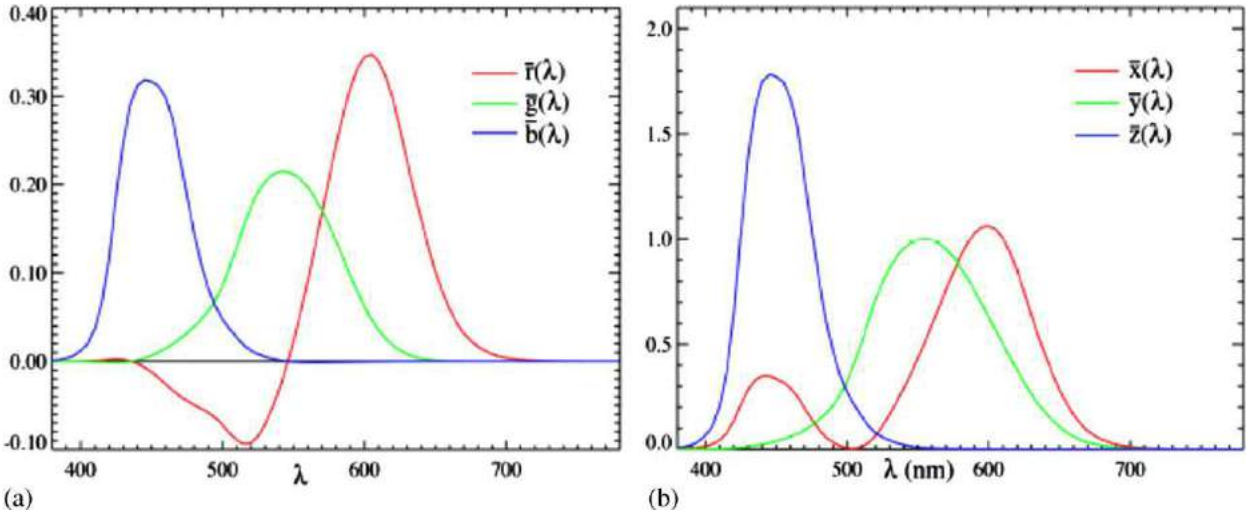


Fig. 5 Color matching functions (CMFs) from (a) human experiments, $\bar{r}(\lambda)$, $\bar{g}(\lambda)$, and $\bar{b}(\lambda)$ and (b) transformation using mathematics, $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, and $\bar{z}(\lambda)$, to prevent from appearing negative values (Wikipedia, 2017).

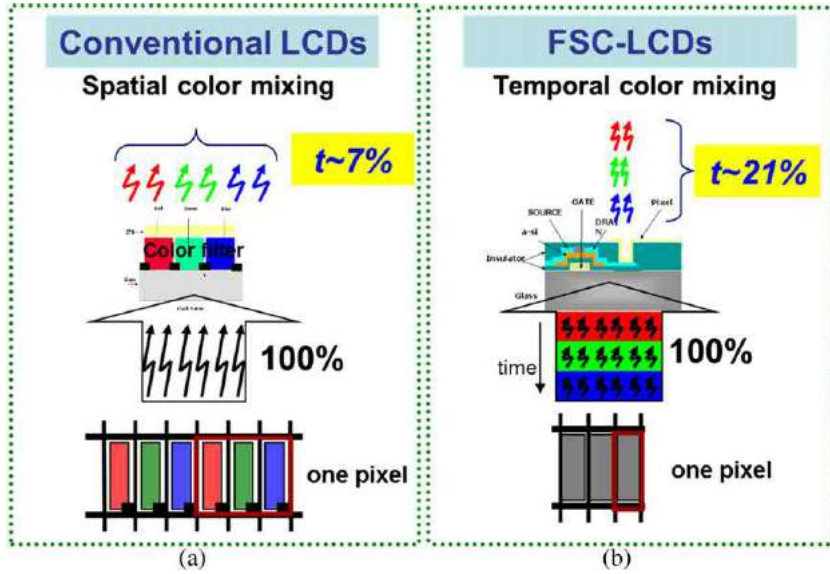


Fig. 6 Display mechanisms of (a) a conventional color filter type LCD by spatial color synthesis, with a poor optical throughput of 7% and (b) a field sequential color LCD by temporal color synthesis, with a triple optical throughput of 21%. Getting rid of color filters, the field sequential color LCD has a triple horizontal panel resolution compared with the color filter type LCD.

performed can be classified into two types: spatial color synthesis by micro color filters (typically red (R), green (G), and blue (B)) and temporal color synthesis by rapidly flash the primaries (RGB) time-sequentially, as shown in Fig. 6.

Spatial Color Synthesis

A traditional TFT-LCD is commonly illuminated using a constant full-on backlight which consists of conventional cold cathode fluorescence lamps (CCFLs) or white light emitting diodes (W-LEDs). The backlight polarization state changes to the linear state after propagating through a polarizer on the TFT-array glass. Different voltages between these two glasses control LC orientations and create different phase retardations due to the LC birefringence property. When the linear polarized lights propagate through the LC cell, the polarization states are modulated by differing LC orientations. Finally, lights from different sub-pixels continue to propagate into red, green, and blue color filters in various intensities and become colored. After passing through the analyzer, a color image is finally created through human eye spatial synthesis. In a commercial 10-bit LCD, the intensity of each subpixel (i.e., R, G, and B subpixels) can range over 1024 shades of gray. Thus, this 10-bit LCD combining

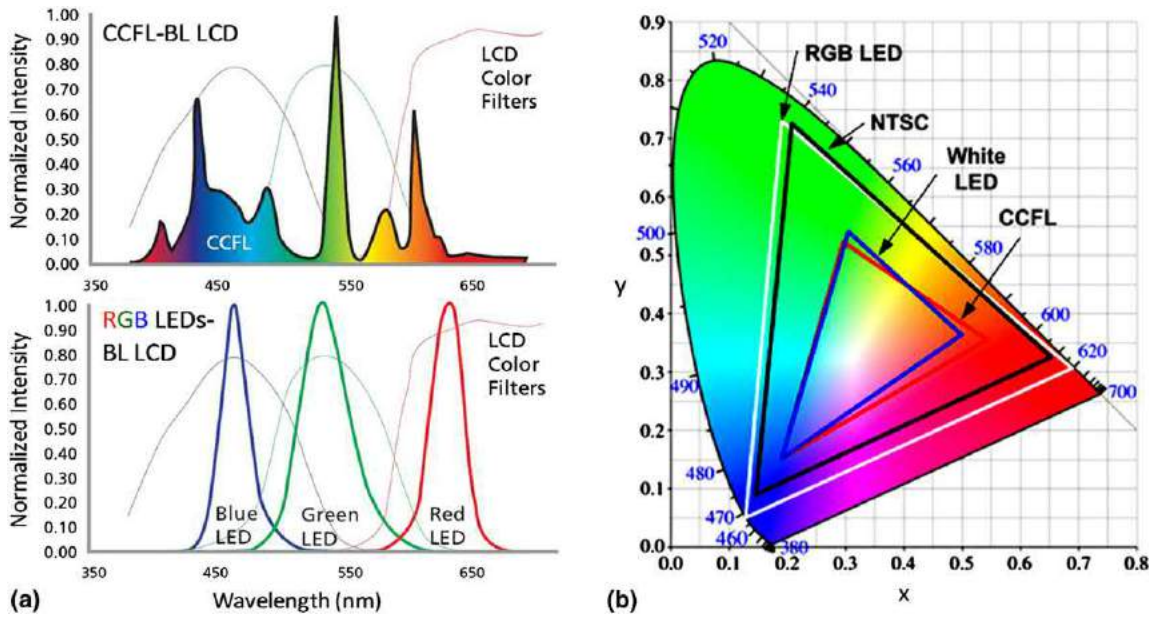


Fig. 7 (a) Typical color filter transmittance spectra with illuminant spectra of a CCFL backlight (top) and an RGB-LED backlight (bottom). (b) LCD color gamut with backlights employing the RGB LEDs, CCFLs and white LEDs (LED, 2017). An RGB-LED backlight has narrow emission spectra to fit the spectra of color filters of an LCD, resulting in less color cross-talk among different color filters. Thus, an LCD using the RGB-LED backlight reaches larger color gamut than that of using CCFL or W-LED backlight.

different intensities of R, G, and B subpixels can present a palette with 1.07 billion colors ($1024 \text{ shades of red} \times 1024 \text{ shades of green} \times 1024 \text{ shades of blue}$).

LEDs are praised for several advantages, such as being mercury-free, having the small form factor, a long lifetime, high power efficiency, wide color gamut, and fast switching operation. Therefore, LEDs are being applied to the LCD backlight system and not only save power but also enable the panel lighter and thinner. Due to the fast response of LEDs, it is straightforward to switch the LEDs backlight on/off, thus the advanced technologies of scanning backlight driving (ClearLCD, 2017), high dynamic range (Seetzen *et al.*, 2003), and field sequential color (Hasebe and Kobayashi, 1985) are swiftly developed resulting from this new option for backlight systems.

Although the LCD has been applied in wide applications, some drawbacks, such as low optical throughput and limited color gamut, are still required to be overcome. The generated light propagates through an optical stack comprised of a set of polarizers, color filters, and diffusers. Substantial proportions of light are absorbed and scattered by these components. Overall, approximately only 7% of backlight reaches the front-of-screen image, more than 90% of the backlight is consumed. Additionally, the spectra of CCFL and W-LED widely covers all visible wavelengths (e.g., 400–700 nm), and the transmittance spectra of color filters are not pure peaks at red, green, and blue, as shown in Fig. 7(a). Consequently, the primary colors perceived by human eyes in a conventional LCD are often made up of several different frequencies resulting in color gamut shrinkage (Power Electronics, 2017). Typically, the color gamut of a traditional TFT-LCD using the CCFL or W-LED backlight is 72% NTSC only. Compared with a CCFL or a W-LED backlight module, an RGB-LED backlight has proper and narrow emission spectra to fit the spectra of color filters of an LCD. Thus, an LCD using the RGB-LED backlight reaches larger color gamut, typically as high as 120% NTSC, as shown in Fig. 7(b).

The issues of limited color gamut and low light efficiency are attributed to use wide-band CCFL/W-LED backlight and use color filters absorbing roughly 70% backlight intensity. As the world is increasingly aware of its ecological footprint and the harm that is causing to our planet, there is a great incentive to reduce the power dissipation of state-of-the-art LCDs. Hence, Hasebe and Kobayashi (1985) proposed a temporal color synthesis, field sequential color (FSC) technology, to present a vivid color image without using color filters. Using R, G, and B LEDs which are a narrower spectrum and near point light source to replace a conventional CCFL and W-LED backlight modules. The color gamut of an FSC-LCD using RGB-LEDs as a backlight easily reaches more than 120% NTSC.

Temporal Color Synthesis

Field sequential color LCDs (FSC-LCDs) without color filters was proposed to enhance light efficiency and color gamut. Differing from the spatial color mixing of the conventional color-filter type LCD, the R, G, and B fields for each frame in an FSC-LCD are sequentially and rapidly displayed. The three-primary system requires a minimum refresh rate of 180 Hz to prevent luminance flicker. The human eye can perceive full-color images by the temporal color mixing mechanism. Because the FSC-LCD gets rid of

color filterless, the advantages are low material cost, three times possible screen resolution, large color gamut, and high light efficiency (low power consumption) (Huang *et al.*, 2011).

However, FSC-LCDs notoriously suffer from color breakup (Jarvenpaa *et al.*, 2005), which occurs when relative velocities between displayed objects and the eyes of the viewer exist. In that case, the constituent color components do not overlap entirely at the retina, and the individual color fields may become visible in the front-of-screen image (Fig. 3(a)). This color breakup phenomenon may occur both during smooth motion pursuit, but particularly during saccadic eye movements which may result in large retinal color displacements. Color breakup artifacts are highly annoying, causing some viewers severe viewing discomfort and in some cases even nausea.

Static Color Breakup

Static color breakup phenomenon happens when observing static images. While humans watch a static image, eyes will move around the image to gather details. During or after a saccade, the static color breakup is perceived. The mechanism of static color breakup is illustrated in Fig. 8 (Jarvenpaa *et al.*, 2005; Yohso and Ukai, 2006). Fig. 8(a) shows three static white bars with a black background showing on an FSC-display. When perceiving the image, human eyes sweep and gaze between white bars with saccadic movements, like the white line in Fig. 8(a). In a conventional FSC-LCD using an RGB color sequence, the display shows red, green, and blue fields sequentially. When eyes move in saccade, the field images of different colors are projected onto the retina separately, as illustrated in Fig. 8(b). The effective factors of static color breakup are saccade velocity, image difference between objects and background, and display filed rate.

Dynamic Color Breakup

The other type of color breakup, dynamic color breakup, happens while perceiving dynamic images (Kurita and Kondo, 2000). Human eyes will pursue a moving object at the same velocity to focus the object on the fovea, as shown in Fig. 9. When a white bar

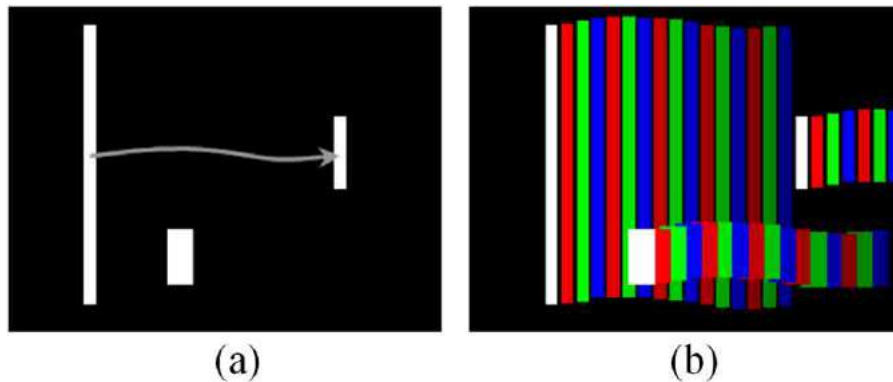


Fig. 8 Mechanism of the static color breakup. (a) Images on an FSC display and a saccade path, and (b) observed color breakup during or just after saccade (Jarvenpaa *et al.*, 2005).

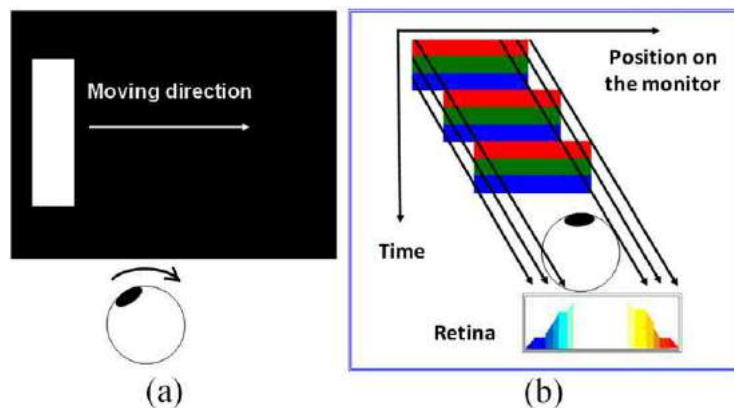


Fig. 9 Mechanism of the dynamic color breakup. (a) Human eyes synchronized with the moving image and (b) the eye-trace integration on the retina. The extra color field will be observed at the bar edge.

is moving from left to right on an FSC-LCD using an RGB color sequence, an observer perceives it through pursuit movement. Fig. 9(b) shows the spatial-temporal relation. The horizontal axis is the position of the FSC-LCD, the vertical axis is time, and oblique black lines indicate the eye-trace line. After eye integration, the three primary colors are projected onto the retina separately and causes the colorful band appeared on the edge of the white bar. From the observing mechanisms of these two types of color breakup, static color breakup is more severe and troublesome than dynamic color breakup because of faster eye velocity.

Color breakup suppression has been implemented on digital light processing (DLP) projectors by inserting additional mono color fields (Langendijk *et al.*, 2006) or increasing the field rate to 540 Hz or more (Mori *et al.*, 1999). Hence, most traditional color breakup solutions for FSC-LCDs involved either increasing the field rate or inserting additional mono-color fields (RGBY, RGBCY or RGBD). The black fields insertion method (360 Hz RGBKKK) has been reported to narrow the color breakup width further to be less perceptible (Koma and Uchida, 2003). Among these three methods, 360 Hz RGBKKK was concluded as the best color breakup suppression method. Although RGB-LED backlights can be switched very rapidly, the response time of liquid crystals prevents the implementation of many of the above methods (e.g., field rate > 300 Hz) for large-screen FSC-LCDs. In spite of the emergence of new fast panel technologies, it is unlikely that affordable panels with frame rates higher than 240 Hz will be widely available in the foreseeable future. In recent years, Lin *et al.* proposed Stencil-FSC (Lin *et al.*, 2009, 2015) and local-primary-desaturation (LPD) (Lin *et al.*, 2011; Zhang *et al.*, 2011) methods to effectively suppress this color breakup issue and make the FSC-technology potential to be practical.

Stencil-FSC:

The “stencil” is a unique technique of painting, and its main concept is to paint the base color first and paint the detail parts in the following. Sometimes multiple stencil layers are utilized for a single image to add colors or create the depth illusion. Additionally, in the human visual system, the luminance contrast sensitivity function is significantly higher than the chromatic contrast sensitivity functions in both temporal and spatial frequencies. It indicates that the visual system is more sensitive to small image changes in luminance contrast compared to chromatic contrast (Fairchild, 2005). Based on stencil concept and the human visual properties, the Stencil-FSC methods incorporate the local color backlight dimming technology. The LC panel performs as a monochrome stencil mask to produce the outline of an image, then the spatially modulated color LED backlight “paints” colors on the mask to create a multi-color field, as shown in Fig. 3(b)–(d). By doing so, the most luminance is gathered in the multi-color field and reduces the luminance of the residual field image to suppress color breakup successfully. For practical applications, the Stencil-FSC methods drive at 240 Hz (four field images), 180 Hz (three field images) and 120 Hz (two field images) field rates to effectively suppress color breakup for large-sized TFT-LCD applications (Fig. 10). Currently, 180 Hz Stencil-FSC has been implemented on a 65-in. color filter-less MVA LCD with 32×24 RGB-LED backlight resolution. Using the IEC 62087 video according to the average of APLs (average picture levels) curve to evaluate display power consumption, 180 Hz Stencil-FSC (52-Watt in average) averagely reduced power consumption to 20% compared with a 65-in. conventional color filter LCD (250-Watt).

However, 180 Hz Stencil-FSC induces green desaturation because redundant red and blue colors are mixed in the green-based color field (Huang *et al.*, 2009). This problem is especially apparent in an image that contains abundant green information, as illustrated by the magenta circle of Fig. 11. For an image, *Lotus*, the first field LC signals are taken from green color; in this case, the blue and red color LC signals are lower than green LC signals. Therefore, even though the red and blue backlight intensities under the purple circles are low, the blue and red lights also propagate through the green-based field and contribute redundant luminance resulting in a reduction of green color saturation. Consequently, a modified algorithm was proposed, 180 Hz min-based Stencil-FSC, the average color difference of CIEDE2000 of the *Lotus* is decreased from 2.9 to 0.3 (Huang *et al.*, 2009).

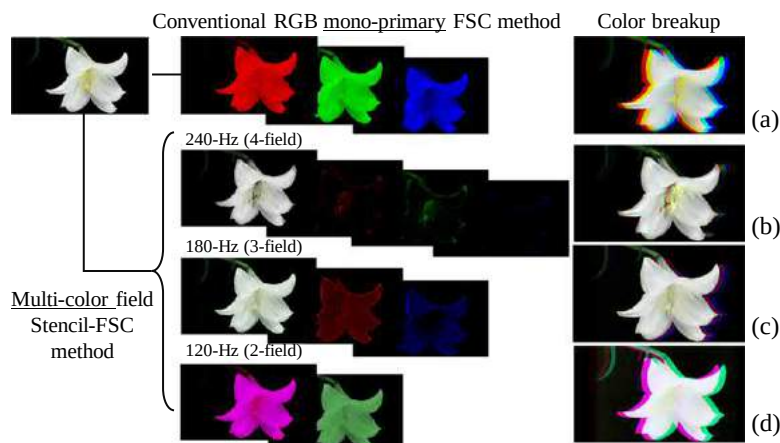


Fig. 10 Schematic plot and corresponding simulated color breakup images of (a) current RGB-FSC method and proposed Stencil-FSC methods in (b) 240 Hz, (c) 180 Hz, and (d) 120 Hz. Stencil-FSC produces a multi-color field image to accumulate most image information, resulting in less intensity of the remaining primary color fields. Thus, color breakup is reduced.

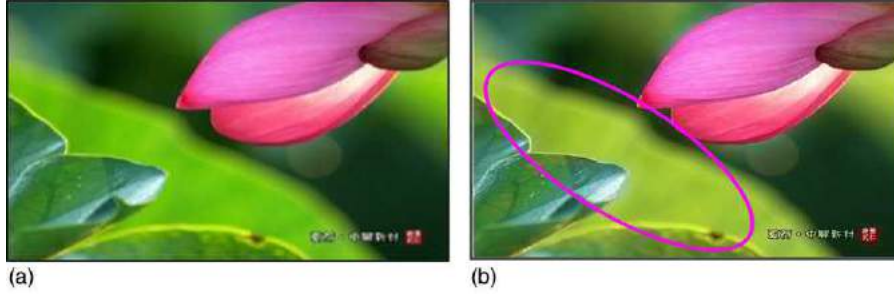


Fig. 11 Reduction of green color saturation using 180 Hz Stencil-FSC. (a) Target image, *Lotus* (provided by Jacky Lee). (b) Simulated image with an average color difference of 2.9 in CIEDE2000 (Huang *et al.*, 2009).

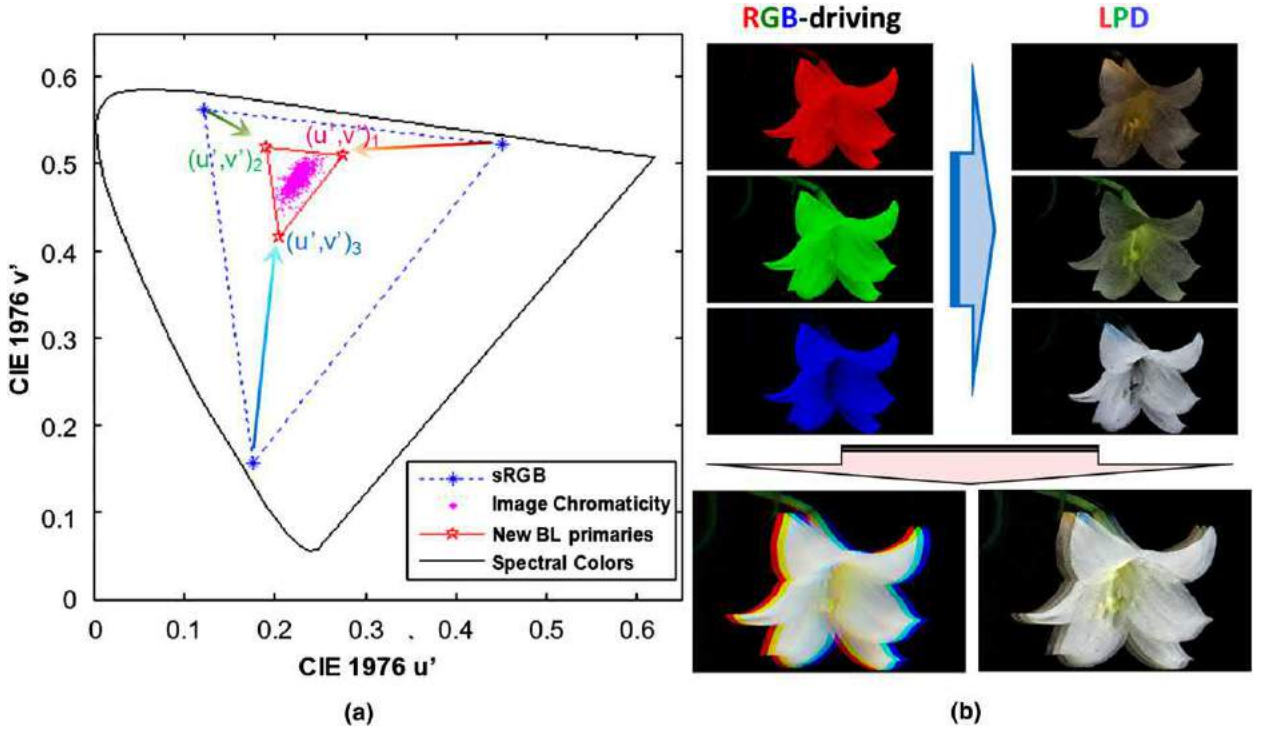


Fig. 12 (a) The color triangles of sRGB and LPD backlight primary colors in the CIE 1976 $u'v'$ uniform chromaticity diagram. (b) The three field images using the conventional RGB-driving (left) and the LPD method (right), and their color breakup images.

However, the intensity of the first field image is not as high as that of green-based one. Thus, the color breakup suppression would be slightly decreased.

To overcome the shortcomings in the previously addressed methods, a novel algorithm, the LPD process, was further proposed to redistribute the original three highly saturated primary color backlight fields, e.g., R, G, and B, into less saturated ones. Color breakup is suppressed with the reduced color differences between each field.

LPD:

Traditional FSC-LCDs utilize the maximum display gamut to display fully saturated R, G, and B field images sequentially at an 180 Hz field rate to create a full-color image. Without considering color contents of a target image, the color breakup phenomenon is very severe, as shown in Fig. 12(b)_left. In contrast, using the LPD method with a field rate of 180 Hz, the RGB-LED backlight module was segmented and independently controlled to provide a small but sufficient color gamut for each input image. The pixel chromaticity distributions of each backlight segment are first analyzed in the CIE 1976 $u'v'$ uniform color space. Next, the backlight primary colors enclosing all pixels chromaticities are dynamically desaturated, as illustrated in Fig. 12(a). By doing so, the color of the three backlight fields are desaturated by mixing another color(s) in the same image field (Fig. 12(b) right). The color differences between each field image are reduced. Thus color breakup is significantly reduced, as shown in the right of Fig. 12(b).

The color breakup reduction concept of Stencil-FSC is to display a multi-color image accumulating most image information in the first single field, thus lowering the intensities of the residual field images. Even human eyes have a fast relative velocity with an object on an FSC screen, the separated color fields are dark, thus suppressing color breakup. The algorithm is simple with only one convolution needed for backlight intensity evaluation. However, the 180 Hz Stencil-FSC method faces a trade-off between image fidelity and color breakup suppression. In contrast, the LPD method utilizes a small but sufficient color gamut to display three new desaturating field images depending on a target image and then reducing color differences between each field images. This is how LPD maintains high image fidelity and suppresses color breakup. However, three convolutions are required to evaluate backlight intensities for three field images when using LPD. This computation complexity causes a high hardware cost to execute LPD. Despite these minor issues, the FSC technique is still highly promising to apply on future display scenarios, such as ultra-low power LCD, light-weighted augmented reality and virtual reality displays, and low heat and high brightness vehicle head-up display.

See also: Displays – Introduction

References

- CIE 170-1:2006, 2017. Cone fundamentals for various field sizes and observer ages. Available at: <http://www.cis.rit.edu/mcsl/online/cie.php> (accessed 01.07.17).
- ClearLCD featuring Aptura™, 2017. Dynamic Scanning Backlight makes ClearLCD™ TV come alive. Available at: <https://wenku.baidu.com/view/d63879126edb6f1aff001f0f.html> (accessed 01.07.17).
- Fairchild, M.D., 2005. Color Appearance Models, second ed. John Wiley & Sons.
- Hasebe, H., Kobayashi, S., 1985. A full-color field sequential LCD using modulated backlight. In: Proceedings of SID Symposium Digest, vol. 16, pp. 81–83.
- Hecht, E., 1998. Optics, third ed. Addison Wesley Longman, Inc.
- Huang, Y.-P., Lin, F.-C., Shieh, H.-P.D., 2011. Eco-displays: The color LCD's without color filters and polarizers. J. Display Technol. 7 (12), 630–632.
- Huang, Y.P., Lin, F.C., Tsai, C.C., Shieh, H.P.D., 2009. Stencil field-sequential color method for color breakup suppression on 180 Hz LCD-TV. In: Eurodisplay Symp. Digest Tech Papers, vol. 29.
- Jarvenpaa, T., 2005. Measuring color breakup of stationary image in field-sequential-color. J. Soc. Info. Display 13, 139–144.
- Koma, N., Uchida, T., 2003. A new field-sequential-color LCD without moving-object color break-up. J. Soc. Info. Display 11, 413–417.
- Kurita, T., Kondo, T., 2000. Evaluation and improvement of picture quality for moving images on field-sequential color displays. In: International Display Workshop, pp. 69–72.
- Langendijk, E.H.A., Swinkels, S., Eliav, D., Chorin, M.B., 2006. Suppression of color breakup in color-sequential multi-primary projection displays. J. Soc. Info. Display 14, 325–329.
- Lin, F.C., Chang, C.W., Huang, Y.P., Shieh, H.P.D., 2015. High-image reproduction by the low-field-rate stencil-LPD method for field-sequential-color LCDs. J. Display Technol. 11 (12), 1069–1075.
- Lin, F.C., Huang, Y.P., Wei, C.M., Shieh, H.P.D., 2009. Color breakup suppression and low power consumption by stencil-FSC method in field-sequential LCDs. J. Soc. Info. Display 17 (3), 221–228.
- Lin, F.C., Zhang, Y., Langendijk, E.H.A., 2011. Color breakup suppression by local primary desaturation in field-sequential color LCDs. J. Display Technol. 7 (2), 55–61.
- Mori, M., Hataida, T., Ishikawa, K., *et al.*, 1999. Mechanism of color breakup in field-sequential-color projectors. J. Soc. Info. Display 7, 257–259.
- Osterberg, G.A., 1935. Topography of the layer of rods and cones in the human retina. Acta Ophthalmol Suppl 13, 1–103.
- Power Electronics, 2017. LED backlighting for LCDs requires unique drivers. In: Power electronics technology. Available at: http://powerelectronics.com/power_management/led_drivers/led-backlighting-lcd-power-efficiency-0512/ (accessed 01.07.17).
- Seetzen, H., Whitehead, L.A., Greg, W., 2003. A high dynamic range display using low and high resolution modulators. In: Proceedings of SID Symposium Digest, vol. 34, pp. 1450–1453.
- UPENN, 2017. Some characteristics of early color vision. Available at: http://www.ling.upenn.edu/courses/ling525/color_vision.html (accessed 01.07.17).
- Webvision, 2017. Simple anatomy of the retina. Available at: <http://webvision.med.utah.edu/sretina.html> (accessed 01.07.17).
- Wikipedia, 2017. CIE 1931 color space. Available at: http://en.wikipedia.org/wiki/CIE_1931_color_space (accessed 01.07.17).
- Yohso, A., Ukai, K., 2006. How color break-up occurs in the human-visual system: the mechanism of the color break-up phenomenon. J. Soc. Info. Display 14 (12), 1127–1133.
- Zhang, Y., Langendijk, E.H.A., Hammer, M., Lin, F.C., 2011. A field sequential color display with local primary desaturation backlight scheme. J. Soc. Inf. Display 17 (3), 242–248.

LCD Components

Fang-Cheng Lin, National Chiao Tung University, Hsinchu, Taiwan

© 2018 Elsevier Ltd. All rights reserved.

A thin film transistor liquid crystal display (TFT-LCD) is composed of two main structures: an LC panel and a backlight module (Schadt and Helfrich, 1971), as shown in Fig. 1. As liquid crystals (LCs) do not emit light directly, TFT-LCDs must rely on additional light sources (e.g., cold cathode fluorescence lamps, CCFLs or light emitting diodes, LEDs) which are generally behind the LC cell and are called "backlight." The components of an LC panel, a backlight module, and periphery driving are introduced, respectively.

LC Panel

The function of an LC panel is like a light valve to modulate the light intensity from a backlight. The LC panel is basically a sandwich structure with LC molecules filled between two glass substrates: a color filter glass and a TFT-array glass. The top glass, color filter glass, is composed of a polarizer (analyzer) and a color filter layer. In the color filter, each pixel segment is made up of three primary colors (i.e., red, green, and blue) sub-pixels approximate 318 micrometers in size for a 4K-UHD 55-in. LCD. The LCs control the light transmission using different molecular orientations that vary by applied voltages. The bottom glass, TFT-array glass, consists of a polarizer and TFTs. The number of TFTs is three times more than the number of pixels due to the requirement for red, green, and blue sub-pixels. The specific components of an LCD are described as below.

Glass

Two glasses (top color filter-glass and bottom TFT-glass) are used as substrates to maintain the stability of liquid crystals. The requirements of an LCD glass are much higher than house-decoration one. Being capable of withstanding a critical manufacturing environment (i.e., high temperature, strong acid chemicals usage), the LCD glass requires the features of low thermal expansion coefficient, chemically resistant, high surface uniformity. Also, alkali metal oxide cannot exist in the glass to prevent the TFT short circuit. To make a light weight and thin panel, the thickness of each glass would be less than 0.5 mm which is also the mainstream in 2016. Finally, the generation of TFT-LCDs manufacturing what we are familiar with is the size of a mother glass (without fabricating and cutting). The 10th-generation means a 2.85 m × 3.05 m mother glass, as shown in Fig. 2.

Polarizer

Each glass has a linear polarizer film. The one in the TFT glass is used to create a polarized light from a non-polarized backlight; the other one in the color filter glass is used to analyze the polarization state of the light passing through the LC cell to control the light intensity for each sub-pixel. Hence, the LCs can modulate the light polarization and control how much light can propagate through the LC panel. A current polarizer form is made from polyvinyl alcohol (PVA) plastic with an iodine doping (Shurcliff,

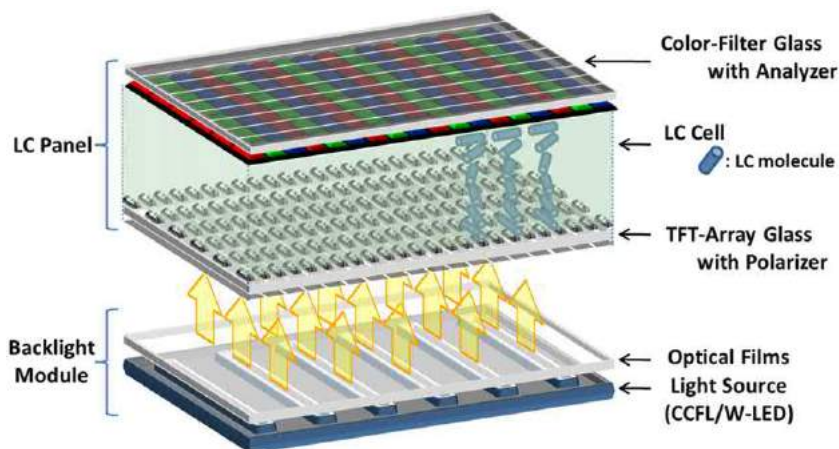
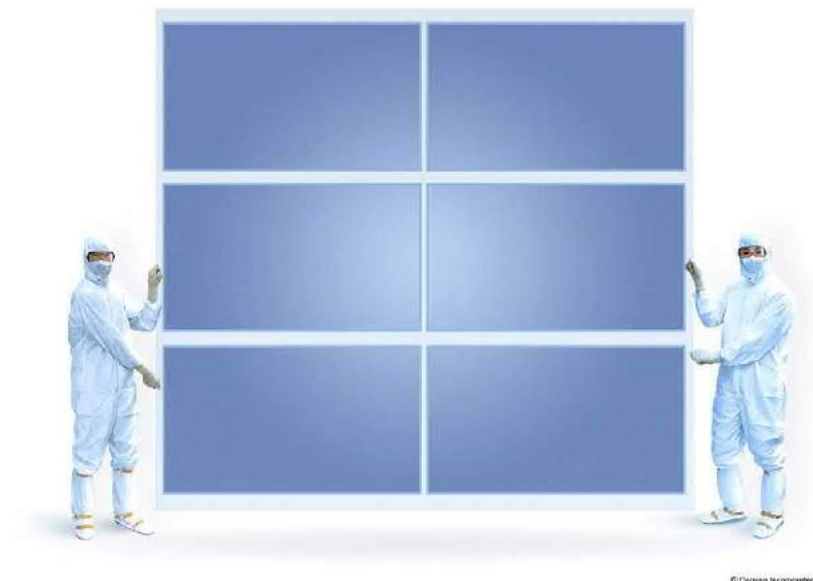


Fig. 1 Traditional LCD structure: an LC panel and a backlight module. The LC panel acts as a light valve to control the transmittance of the light from a backlight source; the backlight module provides the non-self-emitting LC panel with the light source, which is usually CCFLs or white-LEDs.

CORNING

Artwork is showing six 65-inch panels on a Gen 10 glass substrate



© Corning Incorporated 2007

Fig. 2 A 10th-generation mother glass (2.85 m × 3.05 m) presented by Corning which co-located facility with Sharp in Japan in 2007. Available at: <http://mms.businesswire.com/bwapps/mediaserver/ViewMedia?mgid=118723&vid=5&download=1>.

1962; Gunning and Foschaar, 1983). Each side of PVA has a triacetyl cellulose (TAC) layer to increase the mechanical property of PVA. The transmission of a polarizer for LCDs is about 45% or less for visible wavelength.

Transparent Electrode

Transparent electrodes are necessary for the display. The liquid crystal is between the top and bottom electrode glass layers. The backlight needs to propagate through the bottom electrode, color filters, and the top electrode to be displayed. If the electrode is not transparent, it will reduce the aperture ratio, thus reducing the luminance of a display. Indium tin oxide (ITO) is commonly coated on the glass to make a thin and transparent conductive layer in an LCD. The ratio of indium and tin is around 90–10%, and the thickness of ITO layer is typically 100–300 nm. ITO film in the visible light range, the transparency is usually greater than 80%.

Color Filter

Typically, the display is the use of three primary colors (i.e., red, green, and blue) of the additive color mixture to achieve a full-color representation. To precisely show a vivid color image, each pixel has a set of RGB-color filter, names as R, G, and B sub-pixels, as shown in Fig. 3. There is a black matrix in between each sub-pixel to prevent the light crosstalk from neighbor sub-pixels causing a desaturation color. The transmission of the traditional RGB color filter for LCDs is about 30% only. Recently, some companies released 4-primary (RGBY by Sharp) (Yoshida, 2013) and 6-primary (RGBCYM by Panasonic) (Chan and Uei, 2011) color filters to enlarge the color gamut of an LCD. To increase the brightness, RGB with a white sub-pixel, RGBW, was also proposed, especially for a portable display application. Finally, sub-pixel rendering algorithm considering human visual system has also been developed for increasing the apparent resolution of a portable display, such as PenTile technology (Elliott, 2007).

Liquid Crystal (LC)

LCs possess a state of matter that has properties in between isotropic liquid and solid crystal phases. LC phases can be induced by three factors: the temperature variation, the level of solvent concentration, and the ratios of the organic-inorganic composition. In LCD technologies, the temperature variation is the most common used. The finding of LCs is that in 1888, Friedrich Reinitzer, an Austrian botanist, discovered the compound of cholesteryl benzoate has an intermediate state of matter in between lattice structure crystal and isotropic liquid at different temperatures. Next year, Reinitzer sent a letter to Otto Lehmann, a Germany physicist, about his findings. Lehmann used a polarized microscopy to observe this “new phase,” and found its property of birefringence (Δn). Then, they confirmed the existence of this new phase and named it as “Liquid Crystal.” Depending on the geometric

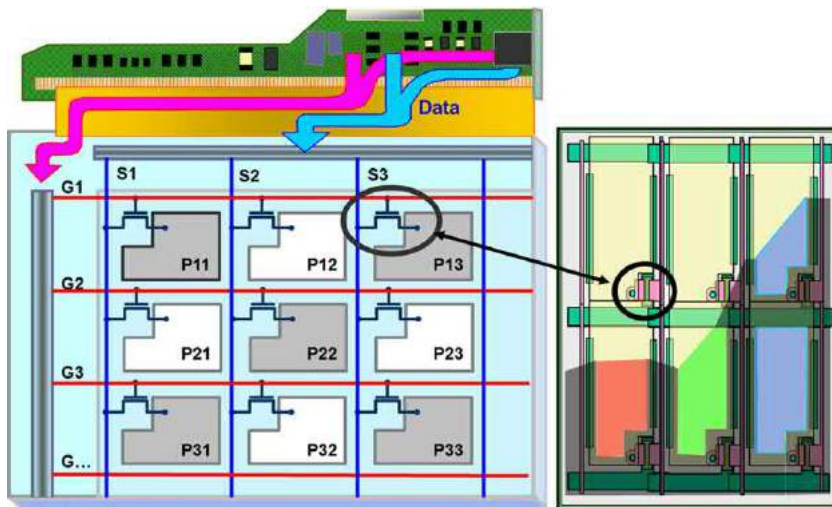


Fig. 3 Schematic diagram of an active-matrix TFT-LCD. The black circles represent a TFT switch in each sub-pixel. The gate lines control the TFT switch on to allow the designed signals in each column sent by the data lines to pass into liquid crystals to control the transmittance of each sub-pixel.

structure of LCs, they are divided into three types of phases– smectic phase, nematic phase, and cholesteric phase (Reinitzer and Chemie, 1888; Lehmann, 1908).

LCs are filled in-between two glass substrates with a cell gap (d). Because LC has optical birefringent property and is varied by different applied voltages, the LC cell can modulate the light polarization status and control the intensity of a backlight. In an LCD, the cell gap is in $\sim \mu\text{m}$ level and uses photo spacers to maintain the cell gap uniformity. The thinner cell gap with a high birefringence index makes the faster LC response, and then a higher frame rate can be achieved to increase the image motion clarity. A fast response LC is required to reach a high frame rate LCD for suppressing motion blur or reaching a field sequential color display. Several research groups have proposed fast response LC modes, such as ferroelectric LC, π cell, optically compensated bend (OCB), and blue phase LC (Meyer *et al.*, 1975; Bos *et al.*, 1984; Yamaguchi *et al.*, 1993; Castles *et al.*, 2012; Chen and Wu, 2014). These modes all can reach a sub-millisecond level but still need more efforts to make them commercialized.

Thin Film Transistor (TFT)

The TFT element is like a switch, and the LC element is like a capacitor. Controlling the TFT switch in on or off state can determine the voltage stored in the capacitor to update or keep the same status. An active-matrix TFT-LCD generally composes of a 2D driving circuit of addressing electrodes. Each pixel has three sub-pixels, and each sub-pixel has a TFT to be an electrical switch and maintain a particular voltage for LC molecules, as shown in Fig. 3. There are three main technologies for TFT manufacturing, including amorphous silicon (a-Si) (Powell, 1989), low temperature poly-silicon (LTPS) (Serikawa *et al.*, 1989), and oxide-TFT (Kamiya and Hosono, 2010).

Currently, most TFT-LCD utilizes a-Si technology because of high uniformity and low manufacturing cost by its mature process. Because the leakage current is small, a-Si is sufficient to be a switch. However, the mobility of a-Si TFT is too low ($\sim 1 \text{ cm}^2/\text{V s}$) to be applied on a high-resolution panel. In contrast, using LTPS process, the mobility is the highest ($\sim 100 \text{ cm}^2/\text{V s}$) among these three kinds of TFTs. It is appropriate to integrate circuit on a glass substrate and achieve system-on-panel (SOP), thus making a narrow bezel, high-resolution panel. However, the uniformity is insufficient, and processing cost is greater than others. Thus, LTPS is currently applied on small- and medium-sized panels. For an oxide-TFT, the mobility is in-between a-Si and LTPS with high uniformity. Therefore, oxide-TFT is suitable for a high resolution and large-sized LCD panel application. In 2004, additionally, Prof. Hosono proposed InGaZnO (IGZO) to be a TFT material (Kamiya and Hosono, 2010). The process of IGZO is compatible with ITO electrodes and becomes a transparent oxide-TFT or called transparent amorphous oxide semiconductor, TAOS.

Backlight Module

Light Sources

The light sources of a backlight are cold cathode fluorescence lamps (CCFLs) or white light emitting diodes (W-LEDs). Because LEDs are praised for several advantages, such as being mercury-free, having the small form factor, a long lifetime, high power efficiency, wide color gamut, and fast switching operation. Therefore, CCFLs are phased out, and point-like W-LEDs become the mainstream for backlight applications. Using LEDs not only saves power but also enables the panel lighter and thinner

(Sakai, 2004). Additionally, the R-, G-, B-LEDs packed in a single chip have also been presented as a backlight source to increase the color gamut of an LCD. This RGB-LED backlight makes an LCD is more compatible with an active matrix organic light emitting diodes display in image color performance. The advanced technologies of scanning backlight driving, high dynamic range (Seetzen et al., 2003), and field sequential color are swiftly developed resulting from this new option for backlight systems.

The LED backlight of an LCD has two types depending on its placements– bottom-lit and edge-lit or side-lit. In a bottom-lit backlight, the LEDs are directly placed under an LC cell. By doing so, the high dynamic range or local dimming technologies can be applied to enhance the dynamic range of an image and simultaneously reducing power consumption. Additionally, the color saturation is also increased while dynamically and independently controlled RGB-LED signals depending on an input image content. However, this bottom-lit backlight requires an enough space (thickness, from 25 to 40 mm) to make a uniform backlight intensity distribution from a point source (from LED chips) to a uniform plane light distribution. Consequently, this makes an LCD bulky. The bottom-lit backlight used to be applied to a large-sized LCD TV. In contrast, a side-lit LCD places its light sources in one or two edges of an LCD bezel. With an enhanced specular reflector or a light guide, light from the edge is redirected to go into an LC cell normally. Therefore, edge-lit backlight module does not need a thick space to mix the light, thus making whole LCD panel thin (from 1 to 10 mm) and promising for portable display applications, such as smartphones, tablets, and laptops.

Optical Films

To increase the light efficiency and make a uniform light propagate through the LC panel, some optical films needs to be utilized, including the backlight reflector film, enhanced specular reflector (ESR), diffuser, brightness enhancement film (BEF), double brightness enhancement film (DBEF), as shown in Fig. 4 (Yi and Emmons, 2007).

Both bottom-lit and edge-lit types backlight, putting an ESR at the bottom of backlight module is necessary to reflect backlight. While the light is coupled from a light guide, a diffuser can scatter and make the light more uniform. However, the light spreads out and is wasted. To overcome this issue, a BEF, attached to the front surface of a backlight unit, is added to collect the light. Only light with a small incident angle can pass through to LC panel; light with a large incident angle will be totally reflected (total internal reflection) to ESR and redirect back to the BEF. Therefore, the uniform scattering light from a diffuser is collected by the BEF.

The non-polarized light is absorbed about 50% to become polarized light by a polarizer in the bottom-glass. To prevent this 50% waste, a DBEF can be utilized behind the bottom-glass. The DBEF can select the proportion of light which polarization direction is parallel to the polarizer on the bottom-glass (i.e., P1 in Fig. 4) and bounce the perpendicular proportion light without being absorbed and depolarized by the diffusion and scattering. (P2 in Fig. 4). This P2 polarized light becomes P1 after re-bouncing by the ESR and passes through the bottom polarizer. Using the BEF and DBEF can increase the brightness, reduce power usage or length the battery usage.

Quantum-Dot (QD) Film

In addition to the brightness enhancement films, a QD film is proposed to increase color saturation. In general, quantum dots are tiny particles, only several nanometers in size. If the material is made into the scale of the quantum dots, the electrons are easily excited to change the energy level, and then combine with the holes and finally emit light. The wavelength of light emitting energy is proportional to the size of quantum dots. Larger quantum dots emit longer wavelengths resulting in emission

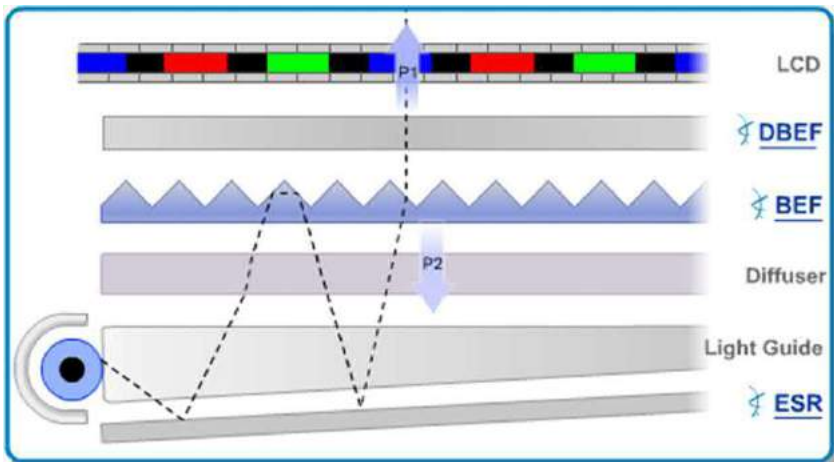


Fig. 4 Optical components of a backlight module from 3M Vikuiti™. When the light coupled from a light guide in an edge-lit backlight module, the optical components are utilized to increase light efficiency and uniformity.

colors such as orange or red. Smaller quantum dots emit shorter wavelengths resulting in colors like green or blue. Therefore, we can manipulate the quantum dot size to stimulate the red, green or blue light. Quantum-dot film (Luo *et al.*, 2014) is utilized to increase image color saturation by narrowing the spectral linewidths of red, green, and blue lights, as shown in Fig. 5. By modulating the size of CdSe, CdS, and CdTe molecules in nano-meter scale, the light spectra are narrowed, thus enlarging the color gamut of an LCD.

Periphery Component

TFT Driver Integrated Circuit

Basic functions of a driver IC (Pappas *et al.*, 2009) are used to drive TFT (gate driver) and supply the corresponding voltage to LCs (data driver). The main circuit structure of the LCD is shown in Fig. 6, including (a) gate driver IC to control TFTs on and off, (b) data driver IC to supply the corresponding voltage to LCs, (c) pulse width modulation to control backlight brightness, (d) DC

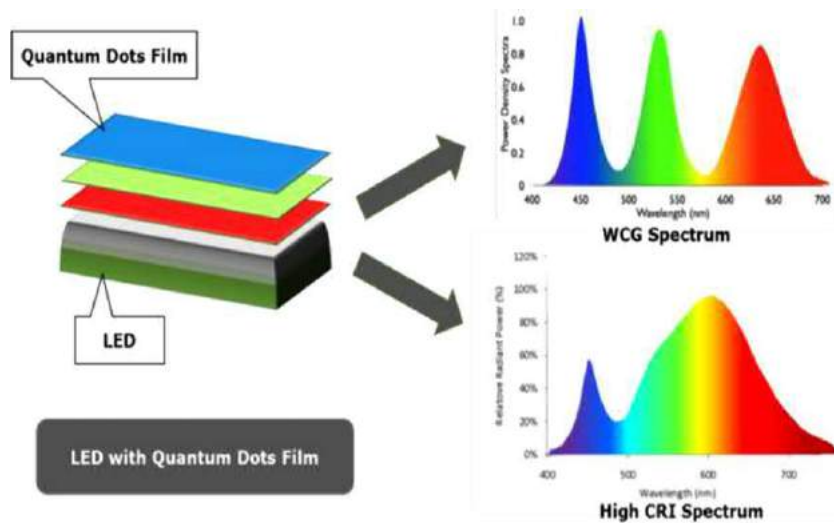


Fig. 5 Quantum-dot (QD) films are utilized in a W-LED backlight system. QD-backlight can reach a wider color gamut compared with W-LED backlight. Data are from 2016 HIS.

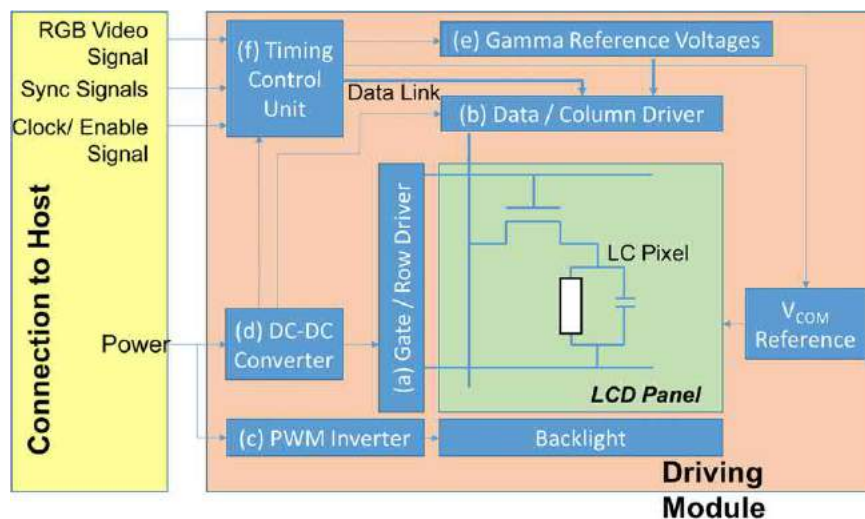


Fig. 6 Driving structure of an LCD, including (a) gate driver IC to switch TFTs on and off, (b) data driver IC to supply the corresponding voltage to LCs, (c) pulse width modulation to control backlight brightness, (d) DC transformer circuit to provide the LCD with the required voltage, (e) gamma reference circuit to correct gamma effect, and (f) timing controller (T-Con) to control the sync signals in horizontal and vertical, and data transmission timing signal. Reproduced from Pappas, I., Siskos, S., Dimitriadis, C.A., 2009. Active-matrix liquid crystal displays – Operation, electronics and analog circuits design. InTech. doi: 10.5772/9686.

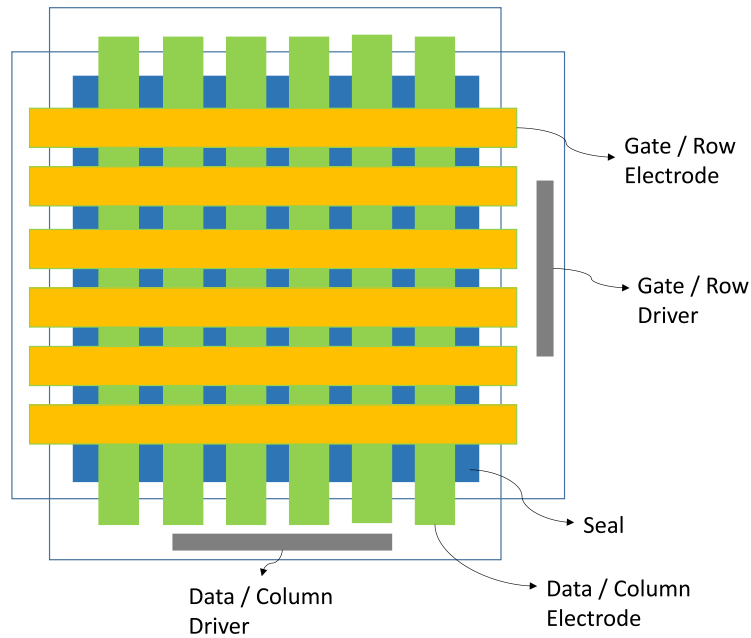


Fig. 7 Passive matrix driving method, which the Y-direction electrode is scanned time sequentially, and the brightness of a pixel is determined by the difference in driving voltage between the X-electrode and the Y-electrode. Reproduced from Pappas, I., Siskos, S., Dimitriadis, C.A., 2009. Active-matrix liquid crystal displays – Operation, electronics and analog circuits design. InTech. doi: 10.5772/9686.

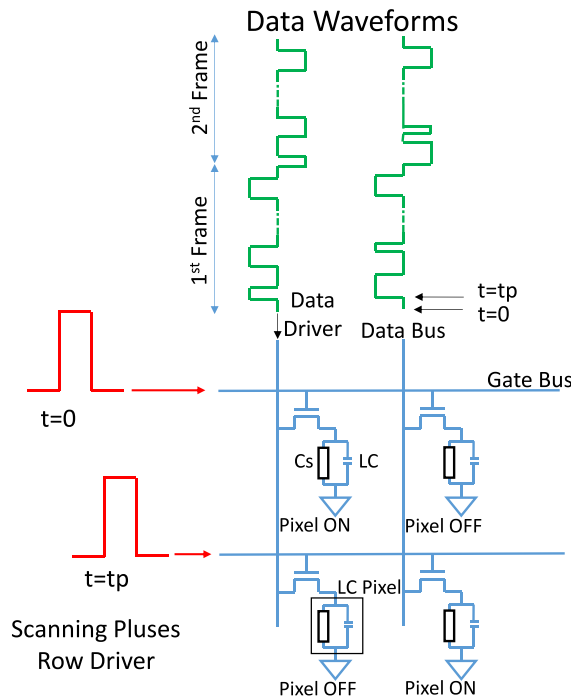


Fig. 8 Active matrix driving method. The data driving circuit sends a signal to each pixel, and a gate driver controls the TFT to open or close. Reproduced from Pappas, I., Siskos, S., Dimitriadis, C.A., 2009. Active-matrix liquid crystal displays – Operation, electronics and analog circuits design. InTech. doi: 10.5772/9686.

transformer circuit to provide the LCD with the required voltage, (e) gamma reference circuit to correct gamma effect because the human eye's sense of light intensity is nonlinear with a gamma 2.2–2.5, and (f) timing controller (T-Con) is used to control the horizontal sync signals and vertical sync signals, and data transmission timing signal.

The TFT driving method can be divided into two kinds of architectures: (1) Passive matrix driving method (PMLCD): A pixel is arranged at the cross of an X- and Y- electrodes, as shown in Fig. 7. The Y-direction electrode is scanned time sequentially, and the

brightness of a pixel is determined by the difference in driving voltage between the X-electrode and the Y-electrode. The disadvantage of a passive matrix LCD is that each pixel has a relatively refreshing rate. Therefore, it is generally applied to a lower resolution display, such as a wearable device.

(2) Active matrix driving method (AMLCD): It usually has two components at each pixel: a switch TFT to control on and off and a capacitor to store the voltage, as shown in Fig. 8. The data driving circuit sends a signal to each pixel, and a gate driver controls the TFT to open or close. When a pixel of the TFT is turned on, then the pixel immediately have a voltage applied to charge the storage capacitor (Cs); when the TFT of a pixel is turned off, then this pixel has no applied voltage, and the LC can maintain its status by the voltage provided by its storage capacitor. This method provides higher response time (refreshing rate) than a PMLCD. Therefore, it is usually used in the high-level products, such as TVs or smartphones.

Trends of LC Displays

Although LCDs are still dominating to industrial, medical, and consumer electronics markets, the AMOLED is occupying the market of small-sized displays, especially in smartphones and AR/VR applications. To make LCDs continue to dominate in these markets, LCDs need to have features of a thinner, lighter, curvable, and flexible profile, lower power consumption/higher efficiency. The display industries also have to enhance LCD performances, such as brighter luminance, higher refreshing rate, higher image resolutions, and a larger color gamut. Therefore, the materials of thinner glasses, fast response LCs, higher TFT mobility and transparency, higher efficiency and color saturation of light sources are worth studying. After having a breakthrough in materials, scientists can implement existing technologies, such as local dimming and field sequential color techniques, on a new product to satisfy the consumers. The display industries also can propose advanced functions to make LCDs more useful, such as dynamically controlling refreshing rate to different operation scenarios. In a touch scenario, the refreshing rate is high to reduce the touching latency and make it more practical; in a reading mode, almost showing a static image, the refreshing rate can be adapted to lower frequency and reduce power consumption, thus lengthening the battery life. Integrating all efforts, the mature manufacturing process of LCDs still provides high quality and cost-effective to produce. Even though AMOLEDs have unique properties, but with featuring these advantages for varieties of display products, LCDs are still promising to hold the share of display markets not only in large-sized but also in small-sized display applications.

See also: Displays – Introduction. Liquid Crystal Physics and Materials

References

- Bos, P.J., Koehler, K.R., Beran, 1984. The pi-cell: A fast liquid-crystal optical-switching device. *Molecular Crystals and Liquid Crystals* 113 (1), 329–339.
- Castles, F., Day, F.V., Morris, S.M., *et al.*, 2012. Blue-phase templated fabrication of three-dimensional nanostructures for photonic applications. *Nature Materials* 11, 599–603.
- Chan, C.-C., Uei, G.-F., 2011. Multi-primary color display and color filter. US Patent 8040027.
- Chen, Y., Wu, S.T., 2014. Recent advances on polymer-stabilized blue phase liquid crystal materials and devices. *Journal of Applied Polymer Science* 131 (13), 40556.
- Elliott, C.H.B., 2007. Color display pixel arrangements and addressing means. US Patent 7307646.
- Gunning, W.J., Foschaar, J., 1983. Improvement in the transmission of iodine-polyvinyl alcohol polarizers. *Applied Optics* 22 (20), 3229–3231.
- Kamiya, T., Hosono, H., 2010. Material characteristics and applications of transparent amorphous oxide semiconductors. *NPG Asia Materials* 2, 15–22.
- Lehmann, O., 1908. Liquids crystals. *Berichte der Deutschen Chemischen Gesellschaft* 41, 3774.
- Luo, Z., Xu, D., Wu, S.T., 2014. Emerging quantum-dots-enhanced LCDs. *Journal of Display Technology* 10 (7), 526–539.
- Meyer, R.B., Liebert, L., Strzelecki, L., Keller, P., 1975. Ferroelectric liquid crystals. *Journal de Physique Lettres* 36, 69–71.
- Pappas, I., Siskos, S., Dimitriadis, C.A., 2009. Active-matrix liquid crystal displays – Operation, electronics and analog circuits design. InTech. doi: 10.5772/9686.
- Powell, M.J., 1989. The physics of amorphous-silicon thin-film transistors. *IEEE Transactions on Electron Devices* 36 (12), 2753–2763.
- Reinitzer, F., 1888. Beiträge zur Kenntniss des Cholesterins. *Monatshefte Chemie* 9 (1), 421–441.
- Sakai S., 2004. A thin LED backlight system with high efficiency for backlighting 22-in. TFT-LCDs. In: *Proceedings of SID Symposium Digest*, vol. 35, pp. 1218–1221.
- Schadt, M., Helfrich, W., 1971. Voltage dependent optical activity of a twisted nematic liquid crystal (TN-LCD). *Physical Review Letters* 27, 561.
- Seetzen, H., Whitehead, Lorne A., Ward, Greg, 2003. A high dynamic range display using low and high resolution modulators. In: *Proceedings of SID Symposium Digest*, vol. 34, pp. 1450–1453.
- Serikawa, T., *et al.*, 1989. Low-temperature fabrication of high-mobility poly-Si TFTs for large-area LCDs. *IEEE Transactions on Electron Devices* 36 (9), 1929–1933.
- Shurcliff, W.A., 1962. Polarized light production and use. Cambridge, Mass.: Harvard U.P.
- Yamaguchi, Y., Miyashita, T., Uchida, T., 1993. Wide-viewing-angle display mode for the active-matrix LCD Using bend-alignment liquid-crystal cell. In: *Proceedings of SID Symposium Digest*, vol. 24, pp. 277–280.
- Yi, J.R., Emmons, R.M., 2007. Thin direct-lit backlight for LCD display. US Patent 7220036.
- Yoshida, Y., Tomizawa, K., Miyazaki, A., 2013. Multi primary color display device. US Patent 8451302.

Relevant Website

<http://mms.businesswire.com/bwapps/mediaserver/ViewMedia?mgid=118723&vid=5&download=1>
Corning.

Projection Displays – LCD/MEMS/LCoS Based

Fleming Chuang, Coretronic Corporation, Taiwan, Republic of China

© 2018 Elsevier Ltd. All rights reserved.

Projection Display

A projection display system also called digital projector, projects an enlarged image on a screen. Projectors are commonly used in presentations, especially for the large size display, say, more than 100" diagonal. As shown in Fig. 1, a basic projection display includes light source which supplies the light, illumination optics which delivers the light and transforms the light beam into rectangular shape and homogenizes light onto microdisplay, and projection optics to project the image of the illuminated microdisplay and enlarge onto the screen.

The illumination optics might comprises the lens array or light pipe (or rod) and several mirrors, relay lens depends on the type of illumination method is used. In some cases, polarization of the light prior to image generation may be required for LC related microdisplay, say LCD and LCoS.

Light Source

The majority of today's projectors use HID lamp (High-intensity discharge lamps (HID lamps) are a type of electrical gas-discharge lamp which produces light by means of an electric arc between tungsten electrodes housed inside a translucent or transparent fused quartz or fused alumina arc tube) as the most common types of projector light source. In the past few years, the light source of projector move from the traditional projector lamp to solid state light source, such as LED and lasers. These new projector light sources are more efficient, environmentally-friendly and longer lasting than their predecessors.

The HID uses combination of rare earth metal salts and mercury vapor to deliver light.

The critical improvement of light lamp is to minimize the Arc gap of the lamp and increase the power to deliver more light in such a small Arc gap. As the smaller the Arc gap is, as higher the light collection efficiency of illumination optics can be.

Used in smaller projectors (so called, pico-projector or pocket-projector), LEDs are essentially applied. LED light sources are mercury-free and can power off and on instantly. The output lumens for the top line LED projectors are targeting for 3K lumens. The life time of LED projectors can last up to 20,000 h at least. It reaches its maximum brightness in several seconds and requires no cooling down period as shut-down it. In comparison old mercury lamps requires several minutes to boost-up to maximum brightness and it can be damaged if unplugged before the cooling down period ends.

In addition to the light source itself, it normally comprises the light collection optics and prisms as well as optical component needed for homogenization, and in some cases, polarization of the light prior to image generation. The imager module is the active component that generates the image from the incoming light.

In early 2010, Sony announced the development of a highly-efficient RGB laser light source with 10,000 h life-span. Since then, Laser projector became the primary developing direction for projection players. Also respectable is the lasers' considerably higher brightness, better contrast and wider color gamut compared to traditional projector lamp.

Microdisplay

Currently there are three types of microdisplay that are widely used in projection display, namely digital mirror devices (DMDs), transmissive liquid crystal displays (LCDs) panels and liquid crystal on silicon (LCOS). The Microdisplay is a key part of the system since most of the optical limitations are influenced by its properties (dimensions, polarization, acceptance angle, etc.), hence those will affect its design structures.

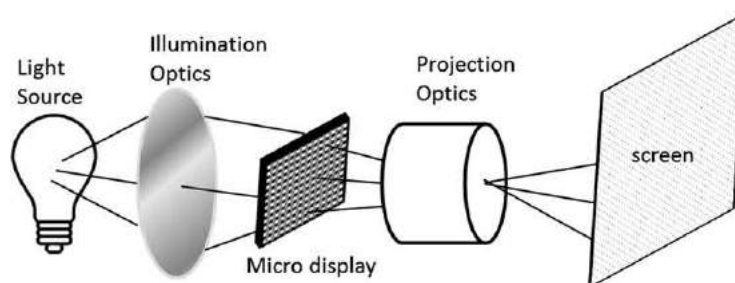


Fig. 1 Basic structure of a projection display.

Projection display – LCD

In 1989, Epson announced the first LCD projector. This projector was developed using three HTPS LCD (HTPS is an abbreviation of High Temperature Poly-Silicon, an active matrix transmissive LCD) panels. The classical structure of 3-LCD projector is shown in Fig. 3 which uses UHP lamp as a light source. The lamp projects white light onto a combination of mirrors that split the light into its three basic video colors. The three color images are combined using a cross dichroic prism (x-prism) (Robinson *et al.*, 2005a) (Cross dichroic prism (X-prism) is combination of 4 triangular prisms. It has function of combining R, G and B three color beams to form the color image) to form a full-color image consisting of millions of colors. The image passes through a projection lens and is projected onto a screen.

The LCD panel applied in projection is HTPS LCD. It has a thin-film transistor (TFT) generated by poly-silicon in each pixel. These pixel transistors act as a conduction switch by changing the scan line's voltage. They are produced in the same way as semiconductors. They are small and highly reliable because they can easily be miniaturized (pixilation or high open area ratio) and drivers can be generated on substrates by processing at a high temperature (Fig. 2).

HTPS LCD panel only transmits longitudinal waves of light. The most popular and effective polarization converter system is a strip half-wave plate. Polarization converter is equipped behind the second lens array (Lens B in Fig. 3). The lens array plate, comprising an array of lenslet, is used to divide light source of individual lenslet into small portion. By passing these focused second light sources through the half-wave plate (Fig. 4), the unified polarization state is formed and the divided light source is superimposed on to the LCD panel by condenser lens and field lens.

Below the projection lens (in Fig. 3), the Cross Dichroic Prism is combination of four triangular prisms. It has function of combining R, G, and B three color beams from respective LCD panels to form the color image. The R/G/B panel should be well aligned and lateral color of projection lens is also required with well control to achieve the crispy pixel imaging.

The 3-LCD projector has three panels to serve for R/G/B separately. Therefore, it does not have possible artifact problems-rainbow effect (Brennesholtz *et al.*, 2008) (The rainbow effect is a visual artifact possible on many single-chip DLP projectors. Because there is only one image chip, color is created sequentially with a rotating color wheel. The sequential color action of slower color wheels can be seen by some people). That principally exists in DLP projector if the frame fresh rate is not high enough. The distinguished advantage of LCD over DLP is the light collection efficiency. LCD can deliver higher lumen output ratio (amount of lamp brightness) in same level color performance and power consumption.

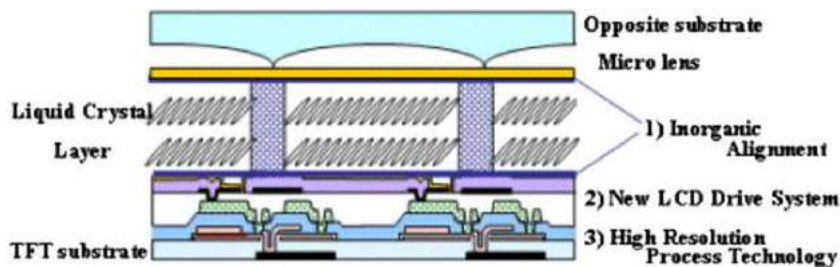


Fig. 2 First HTPS LCD panel using inorganic alignment layer. Available at: <https://phys.org/news/2005-02-world-https-lcd-panel-inorganic.html>.

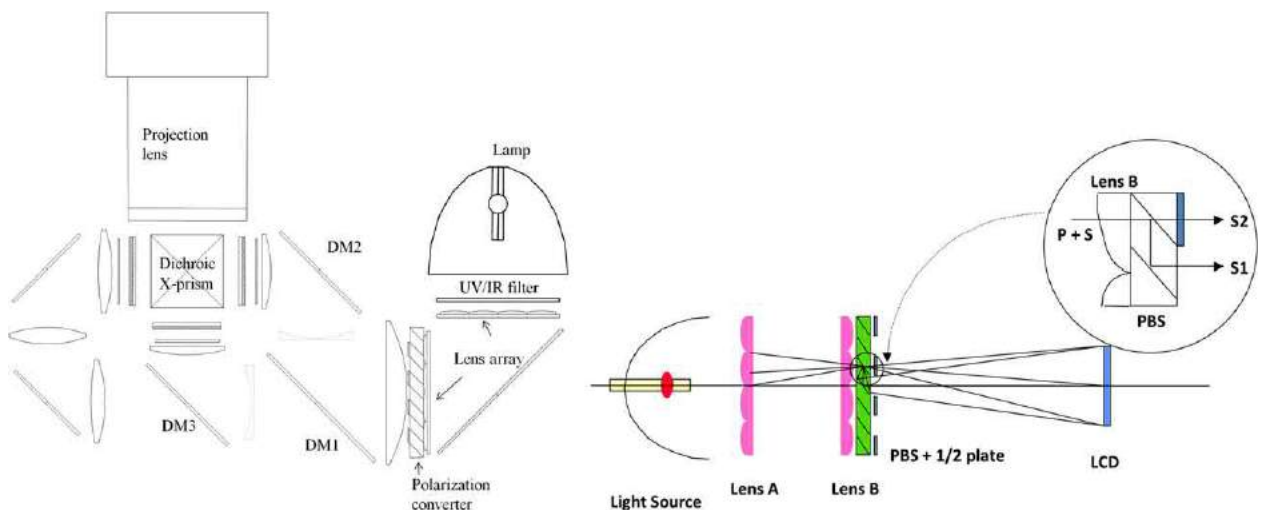


Fig. 3 Structure of LCD projector and Polarization converter (see “Relevant Websites” section).

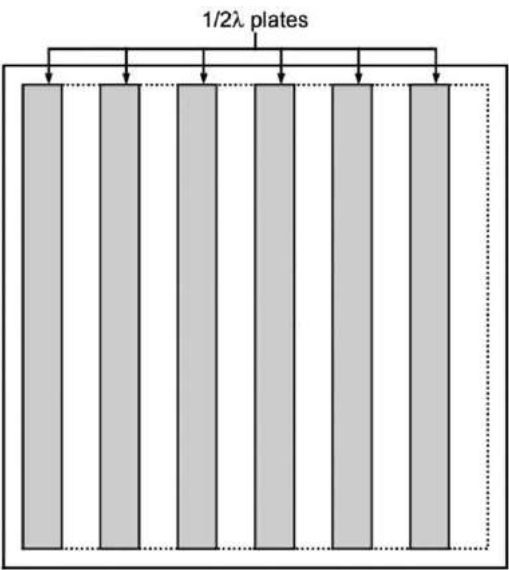


Fig. 4 Stripe Half-Wave-Plate.

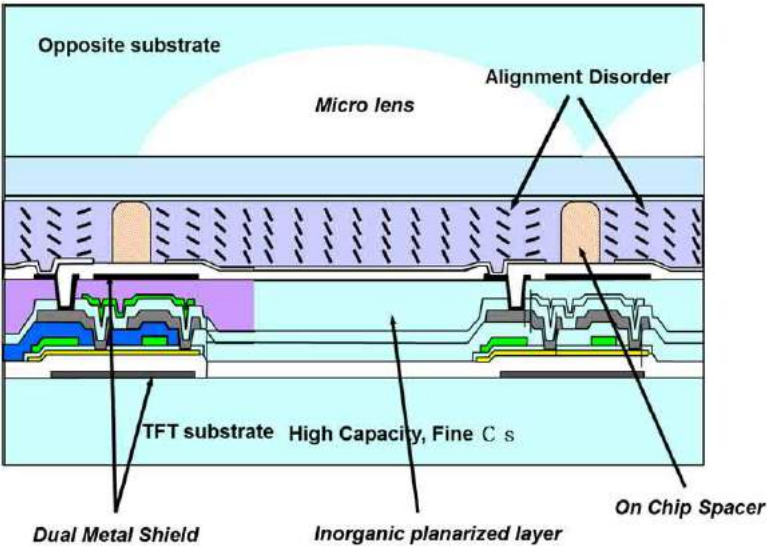


Fig. 5 Typical LCoS structure.

It is known as the screen door effect and resulted in such a sharp image that it was possible to see the individual pixels. With today's advanced resolution and micro-lens equipped in front of LCD panel, the pixels are so small that it becomes harder to notice the screen door effect even though it still exists.

An LCD projector uses three panels that light passes through to yield the full-colored image. Because of this, it is possible for any panel to fall out of alignment which will cause one of the three stacked images to be shifted slightly; this is called convergence. If it happens, then the on-screen image will be slightly fuzzy.

Projection Display – LCoS

Liquid crystal on silicon (LCoS or LCOS) is a miniaturized reflective active-matrix liquid-crystal microdisplay using a liquid crystal layer on top of a silicon backplane. LCoS was initially developed for projection display, but it is also used for wavelength selective switching, structured illumination, near-eye displays and optical pulse shaping, etc. By way of comparison, HTPS LCD projectors are using transmissive LCD, allowing light to pass through the liquid crystal; however, LCoS is reflective LCD. A typical LCoS structure is shown in Fig. 5.

In an LCoS panel, a CMOS chip controls the voltage on square reflective aluminum electrodes buried just below the chip surface, each controlling one pixel with an independently addressable voltage. Typical LC cells are about 1–3 micrometers thick and pixel pitch range from sub-ten microns, as small as 2.79 μm (see “Relevant Websites” section). A common voltage for all the pixels is supplied by a transparent conductive layer made of indium tin oxide (ITO) on the cover glass. The pixels are generally coated with a reflective aluminum layer, and a polyimide alignment layer. The liquid crystal industry can take advantage of the existing silicon technology to produce high resolution microdisplays that is easy and inexpensive to fabricate.

In LCOS, liquid crystals are applied to a reflective mirror substrate. The input light is always modulated into longitudinal polarization light (s-wave). The PBS (Robinson *et al.*, 2005a) (Polarizing Beam splitters (PBS) are used to split unpolarized light into two polarized parts. PBS is Beam splitters designed to split light by polarization state) is the most common light switching optical element to control the light in or out. (Fig. 6).

There are methods to increase the contrast of LCoS and LCD projection system. If a PBS is used as the exit analyzer, the isolating QWP (Quarter-wave plate (QWP) consists of a carefully adjusted thickness of a birefringent material such that the light associated with the larger index of refraction is retarded by 90° in phase (a quarter wavelength) with respect to that associated with the smaller index) can work with a retarder stack filter to enhance the ANSI contrast (Robinson *et al.*, 2005a). By applying the Wire grid polarizer into system, it could increase the NA and keeping high contrast at the same time (Brennesholtz and Stupp, 2008; Robinson *et al.*, 2005b).

There are a few typical optical architectures which have been using in the three-panel LCoS projection system. In the system, each panel separately modulates red, green, and blue light. such as, Philips color prism (US Patent 6,330,113), three PBS plus x-prism (Armitage *et al.*, 2006a), three wire-grid polarizer plus x-prism (Armitage *et al.*, 2006b), three PBS plus three lenses (US Patent 7,198,373), off Axis Illumination (Bone, 2001) and ColorQuad (US Patent 6,183,091), etc. You may see the details in the respective references for above diversified architectures. The most popular type is three PBS plus x-prism, shown in Fig. 7. They used dichroic mirrors to separate the three colors, each of which is directed to the corresponding LCoS panel with a separate PBS.

Skew rays are corrected by QWPs between the PBS and the panel. The reflected images are combined into a full color image in an x-prism prior to the projection lens. This design is conceptually simple and allows the optimization of each color channel separately.

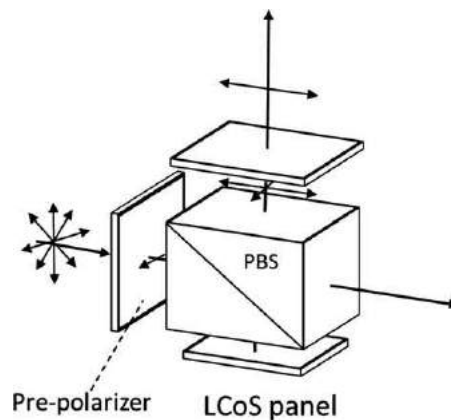


Fig. 6 On and off for PBS and LCoS.

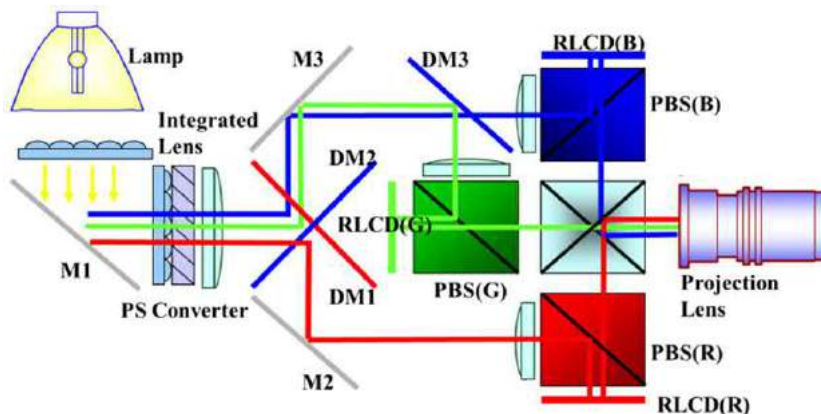


Fig. 7 3-PBS/x-prism LCoS projection system.

Projection Display – MEMS

The DMD™ chip is a micro-electro-mechanical systems (MEMS) array of fast digital micromirrors, monolithically integrated onto and controlled by an underlying silicon memory chip (Hutchison, 2007). (Left of Fig. 8).

The operation principal of DLP projector is shown in Fig. 8 (right). Grayscale patterns are created by programming on/off duty cycle of each mirror, and multiple light sources can be multiplexed to create full RGB color images. Each mirror represents one pixel. The light, rather than passing through the panel, is reflected from it. The mirrors move back and forth, varying the amount of light that reaches the projection lens from each pixel of the primary colors in a rapid rotating sequence to for the true color image.

Optical systems for DLP could be divided into single-panel and three-panel projection applications. (Fig. 9).

Optical systems for single-panel projection applications can be grouped into two main architectures, Telecentric and Non-Telecentric system. Telecentric architectures are defined by locating the exit pupil of the illumination system (entrance pupil of the projection lens) at or near infinity from the device surface. The chief rays of every bundle incident on every mirror then are essentially parallel to each other. For the illumination system, this provides uniform angles of incidence across the entire field, creating uniform black levels for the dark field. Non-Telecentric architectures differ from Telecentric in that the exit pupil of the illumination path is located a short, finite distance from the device, and the entrance pupil of the projection lens must be coincident with it. Some designs use a field lens (or lenses) instead of a prism directly in front of the device to perform the angle separation. The field lens must be on axis with the remainder of the projection lenses, but is shared by the illumination path. This presents some challenges in illumination design since these field lenses also are part of the illumination path, but are off-axis to the DMD and tilted to the illumination path. By designing the illumination pupil near the stop of the projection lens and folding the illumination path there, a very compact system can be designed.

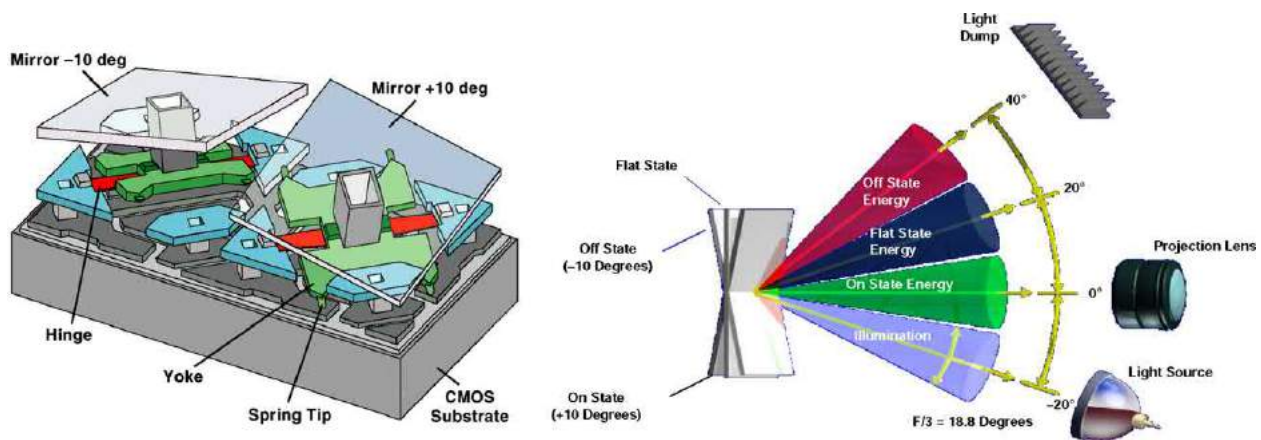


Fig. 8 Structure of DMD chip and simplified optical function. (Courtesy: Texas Instruments Inc.).

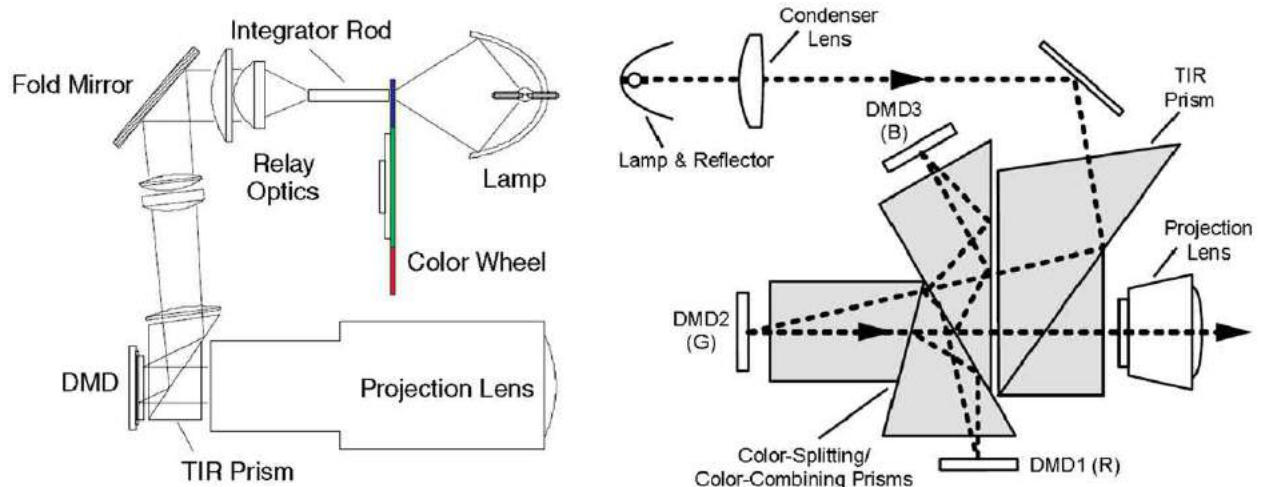


Fig. 9 Generic Optical System: single panel and 3 panels. (Courtesy: Texas Instruments Inc. (Im, H., Kim, W., Chung, J., 2010. A new transmissive type tilting actuator for a smooth projected television image. Proceedings IMechE 224 Part C: J. Mechanical Engineering Science 1327–1333.)).

A three-chip DLP projector uses a prism to split light from the light source, and each primary color of light is then routed to its own DLP chip, then recombined and routed out through the lens. It completely eliminates the “Rainbow” effect of single-chip DLP. 3-Chip DLP Projectors are used for high-performance, high brightness applications in large rooms and large audience venues. 3-Chip DLP technology is currently considered the top of the line technology for digital projection (Im *et al.*, 2010).

New Technologies, Challenges, and Applications in Projection Display Industry

There are many important improvements of the projection display undergoing, no matter what kind of microdisplay is applied. For example, color by applying new solid state light source (LED or Laser), product to meet new HDR standardization, component to improve the dynamic range, image resolution enhancement, UST lens, etc.

4K × 2K: True 4K and 4k by Smooth Picture

As the most advanced digital projection technology for cinema, with 4k digital cinema solutions deployed, DLP and LCoS have their own products. Both are based on 3-chip solutions. By utilizing three 1.4" 4K DMD chips, the manufactures could deliver the Cinema projector's brightness as high as 20K lumens. JVC and Sony have revealed its new native 4K HDR projector, which features a newly-developed laser light source with a claimed brightness of 3000 lumens. However, the price of true 4K 3 LCoS projector is very high no matter DLP or LCoS.

A pixel shifting technique which focuses on producing a smooth picture on the lower resolution panel, DLP and LCoS have their own approaches. So called e-shift by JVC, the operation principle is explained how they work (in Fig. 10). In this diagram at the left-corner of figure, the refractive index of the e-Shift device is changed through use of an electric image shift-switching signal. Following, the path of light emitted by the light sources (such as the D-ILA display device) changes, following the changed refractive index. By changing the path, one pixel is shifted diagonally (vertically and horizontally). By shifting pixels, the vertical and horizontal resolution is substantially doubled. It shows “quadruple” the amount of effective pixels, but it cannot be addressed that small portion of the pixel to display unique information. It has to be the entire pixel and when you put together the two sub frames the entire pixel from the non-shifted and shifted image have to be color coordinated to look like the original image.

JVC has 4K e-shift D-ILA front projector by three full HD resolution D-ILA devices. For the DLP camp, so called “Smooth Picture” technique is developed for the similar pixel shift purpose as well (Hutchison, 2007). For example, the transmissive type of Smooth picture module is operated by vibration to tilt the light from the DMD of optical system (equipped between projection lens and DMD). A voice coil motor (VCM) type actuator for smooth picture has to satisfy performance requirements such as higher driving force, lower power consumption, and lower cost. In updated development (Im *et al.*, 2010), TI developed the 0.67" 2.5K × 1.6K DMD for single chip projector. By applying smooth picture module, the high quality 4K picture is achieved. Several companies have introduced the 4K DLP projectors into market.

The e-shift device would also be available for reflective type (Hutchison and Neal, 2008) instead of transmissive type mentioned above. Compared to transmissive type, it requires a very good quality mirror and more precisely tilt angle control.

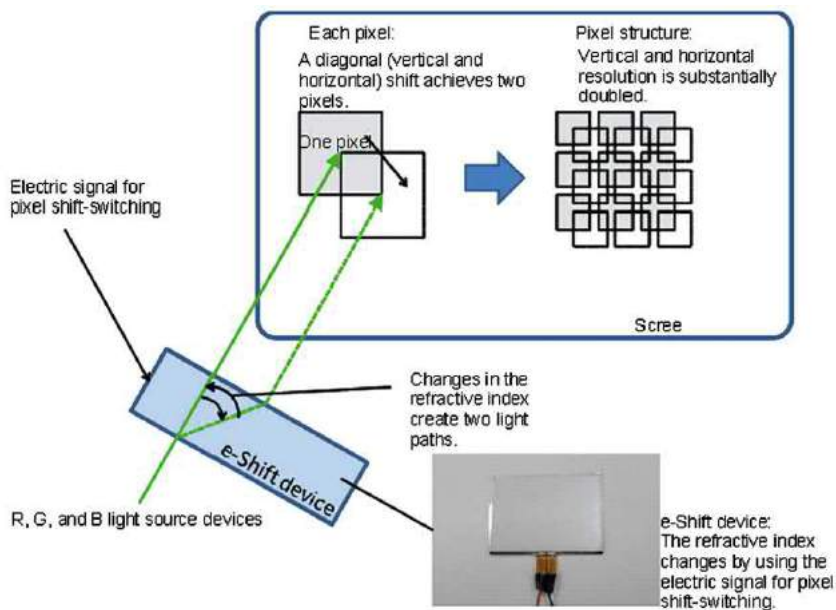


Fig. 10 Pixel shift-switching using the e-Shift device.

Ultra-Short-Throw

The terminology, Throw Ratio, refers to the ratio of the distance to the screen (throw) to the screen width (Fig. 11). A larger throw ratio corresponds to longer distance of a projector from the screen. For an Ultra-Short-Throw (UST), it is very short distance the image is thrown onto the screen, and it has a large effect on screen size. Often in home projector and individuals lacking the space in the room, projector to equip with an ultra-short-throw projection lens is quite popular in the market. As the throw ratio is as below as 0.35, we could call it UST. The critical issue for the projection lens design is that it requires a very large field of view; therefore, design and manufacturing become challenging.

Some special design techniques were developed in UST projector. In popular lens structure of projection system, the stop is positioned to make the system Telecentric or non-Telecentric to match the light beam of illumination system. It is always located close to focal point of the rear part (the lens elements at right of stop) of projection lens. Generally, multiple glass and plastic lenses are combined and large-diameter lenses are used to achieve the target performance. Thus, the wider the viewing angles, the larger the lenses and the more the number of the lenses. The requirements to widen the viewing angles of a projector are inversely proportional to reducing the projector's size and weight causing the manufacturing difficulties, especially; the aspherical surface is mostly applied for reducing the aberrations of lens system.

Some optical structures applied the reflecting mirror at most front position of projection lens was developed. There are two types of mirror: convex mirror and concave mirror. To reduce the size of lens element, forming the intermediate stop or intermediate image between the mirror and front lens element is also designed by many developers. Some distinguished design examples from the published patent are described below.

US Patent 7,009,765 (Fig. 12): One intermediate image and one intermediate stop to reduce the size of front lens group. No mirror is applied.

US Patent 20130162957 (Fig. 13): No intermediate image or stop applied, using convex mirror. The mirror and first lens will be large.

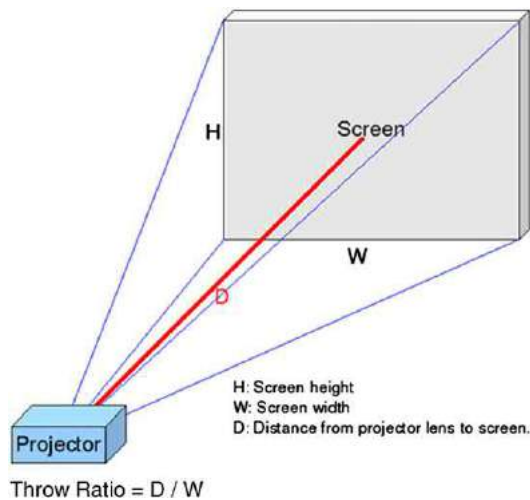


Fig. 11 Definition of Throw Ratio.

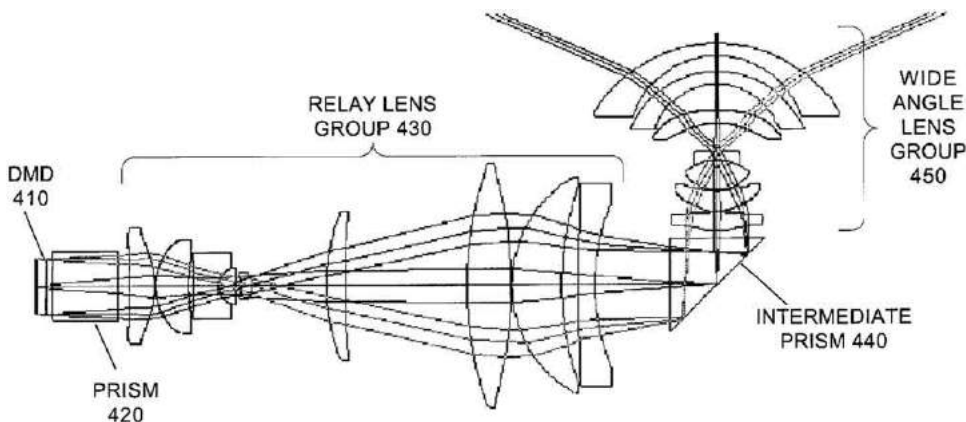


Fig. 12 US Patent 7,009,765B2, Assignee: Infocus.

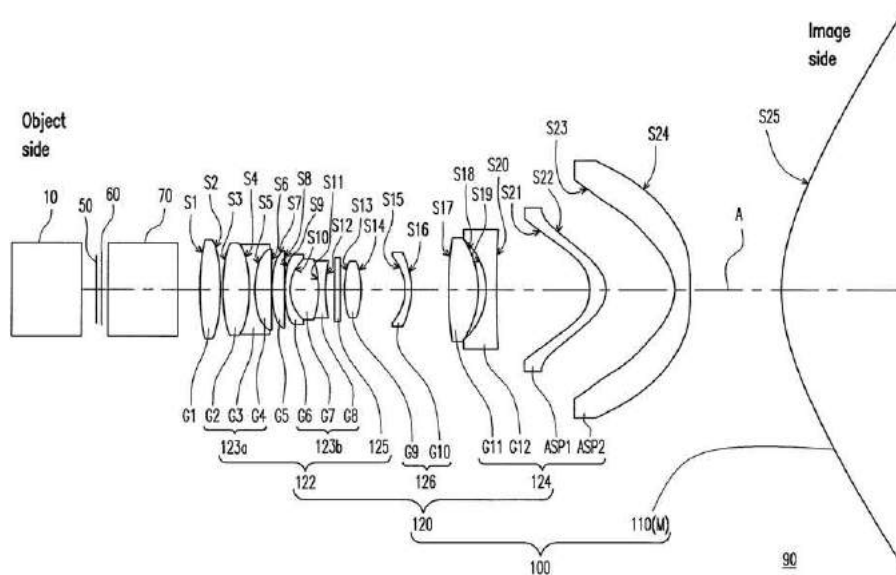


Fig. 13 US Patent 20130162957 A1, Assignee: Young Optics.

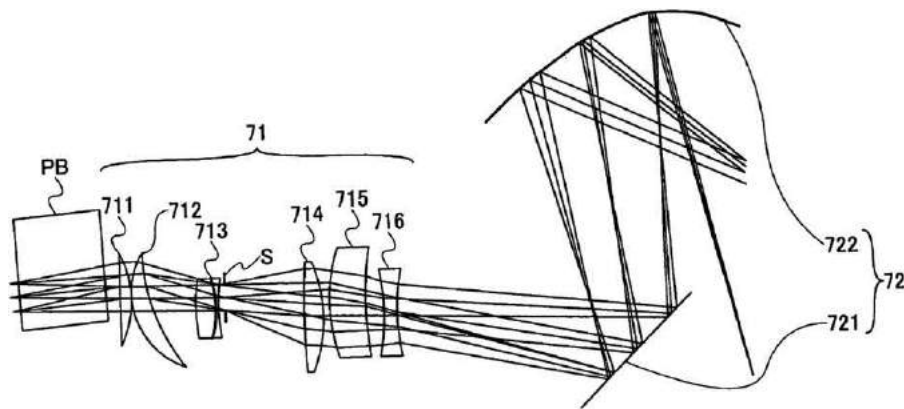


Fig. 14 US Patent 7441908, Assignee: Ricoh.

US Patent 7441908 (**Fig. 14**): One intermediate image is formed in front of concave free-form mirror.

US Patent 9,372,388 (**Fig. 15**): two intermediate image, one intermediate stop, and using concave mirror. This structure has the most compact reflection mirror and lens element.

Laser Phosphor Illumination Projector

Pure laser, also known as RGB laser, generates light directly from three individual red, green and blue lasers. The primary benefit of a RGB laser system is light output while also achieving higher performance in other standard image quality parameters such as color gamut, contrast ratio and dynamic range when compared to standard lamp-based systems. As such, RGB laser is ideal for largescale applications and giant screen cinema.

Laser phosphor illumination projector (**Fig. 16**) uses blue laser diodes as the light source instead of a high intensity discharge lamp to generate the three primary colors in a single DMD laser phosphor projector, the blue laser diode focuses laser light onto a phosphor wheel (Laser phosphor is a lampless projection illumination platform that uses blue laser diodes as the primary light source. To generate the three primary colors – red, blue, green – the blue light from the laser diodes shines onto a spinning wheel that is coated in a phosphor compound) to create red and green light, while blue laser light passes through an opening in the phosphor wheel. These red, green, and blue colors are then directed onto SLM chip, which directs the light through a lens and onto the projection screen. The basic concept is shown as **Fig. 16**. In a laser-phosphor light source, only uses the less expensive blue lasers, that can be made more cost-effective compared to pure laser projectors.

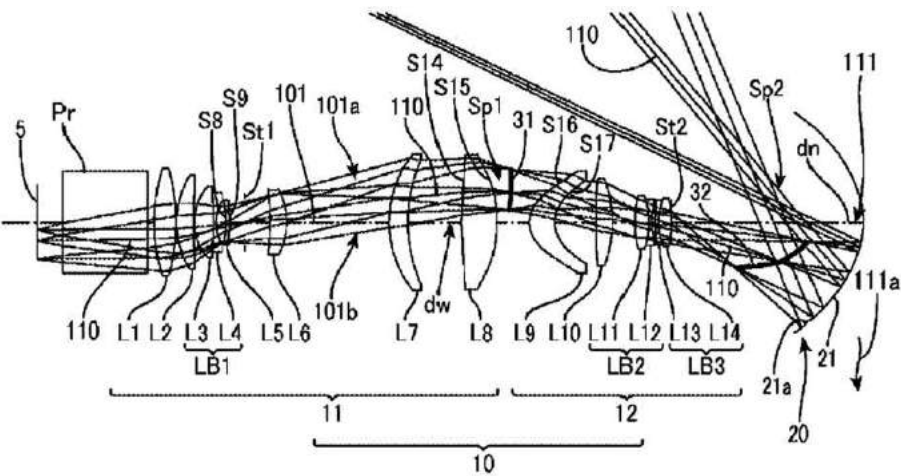


Fig. 15 US Patent 9,372,388 B2, Assignee: Nittoh K.K.

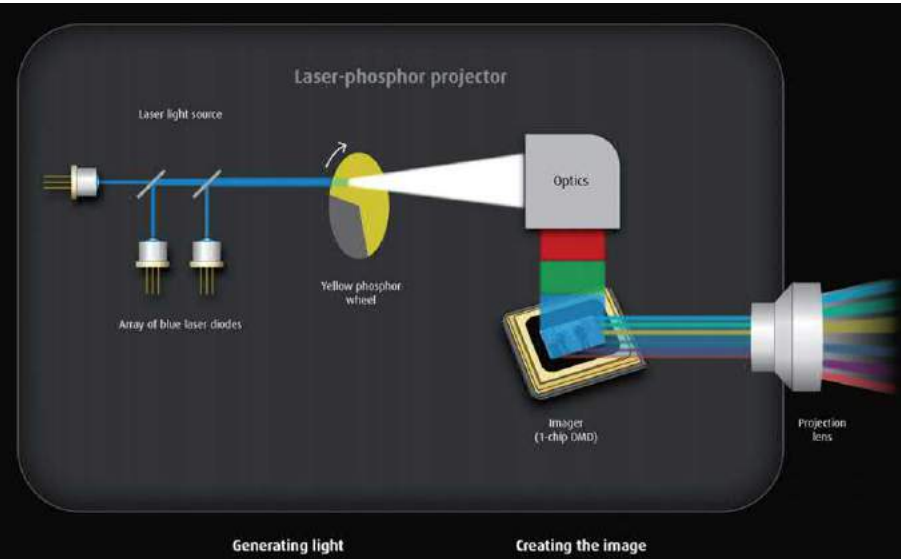


Fig. 16 Laser-phosphor projector (a single-chip projector) (Barco).

The primary advantage of a laser phosphor projector is the long life of the illumination system. As a lampless system, laser phosphor also eliminates the need for lamp and, in many designs, filter replacements, reducing the down-time, maintenance and costs associated with lamp-based projectors.

There are diversified optical structures in designing the laser phosphor projector. The phosphor wheel could be reflecting type (Fig. 17) or combined with color filter to make structure compact, like below patent application (Fig. 18).

Color Gamut and HDR

Displays cannot reproduce all the colors occurring in nature. HDR content preserves details in the darkest and brightest areas of a picture that are lost using current standards. It also allows for more natural, true-to-life colors that are closer to how we see them in real world. High Dynamic Range display provides viewers with an enhanced visual experience by providing images that have been produced to look correct on brighter display, that provide much brighter highlights, and provide improved detail in dark areas. Extension of the color gamut, necessary to realize more realistic color reproduction, is possible achieved by changing the chromaticity of the three primary colors. DCI-P3 describes a wider color gamut than Rec. 709 with three primary colors and is implemented for digital cinema based on xenon arc discharge lamps. Rec. 2020 reproduces images more realistically and includes the gamut of the major colorimetric standards but requires quasi-monochromatic light sources such as RGB lasers. The chromaticity diagram (CIE 1931) of Rec. 709, DCI-P3, and Rec. 2020 are shown in Fig. 19. The HDR regulation for HDR10

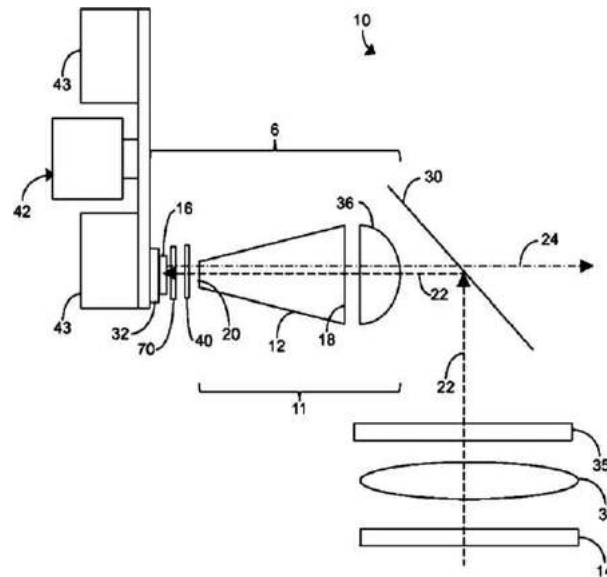


Fig. 17 US Patent 9,170,475: Light valve projector with a laser-phosphor light converter.

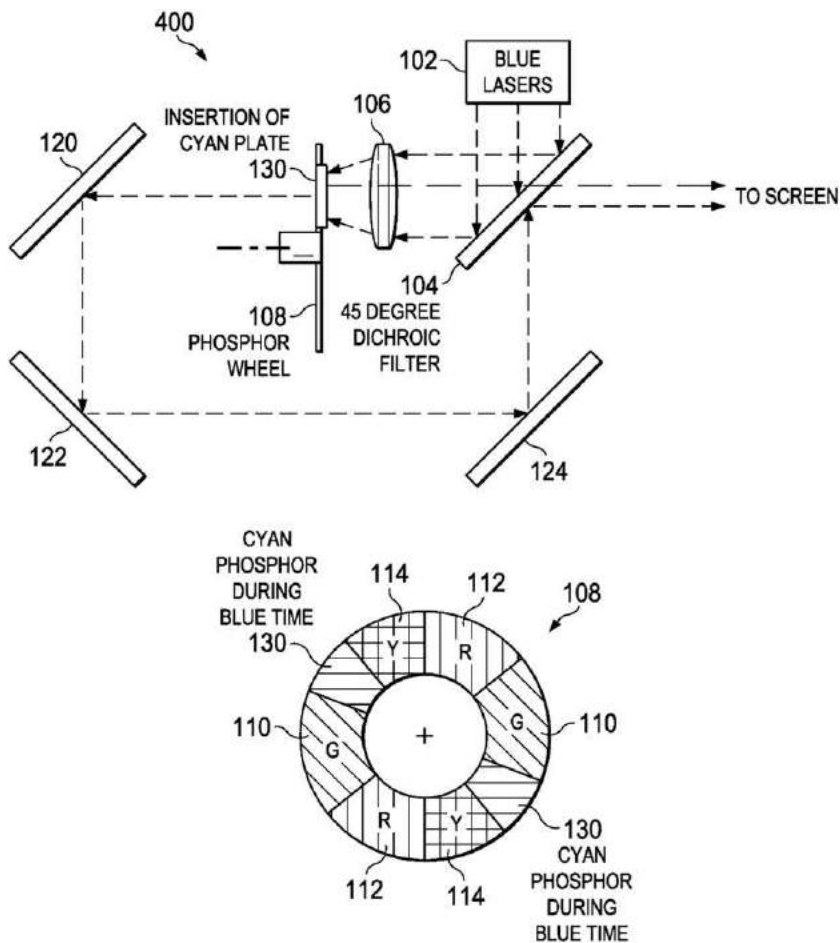


Fig. 18 US Patent 2014/0253882 A1: Hybrid laser excited phosphor illumination apparatus and method.

(see “Relevant Websites” section) and Dolby Vision (see “Relevant Websites” section) could be found in reference. Solid state light sources are gradually becoming the preferred light sources for projection displays instead of the traditional (broadband) short-arc xenon or ultra-high performance (UHP) discharge lamps.

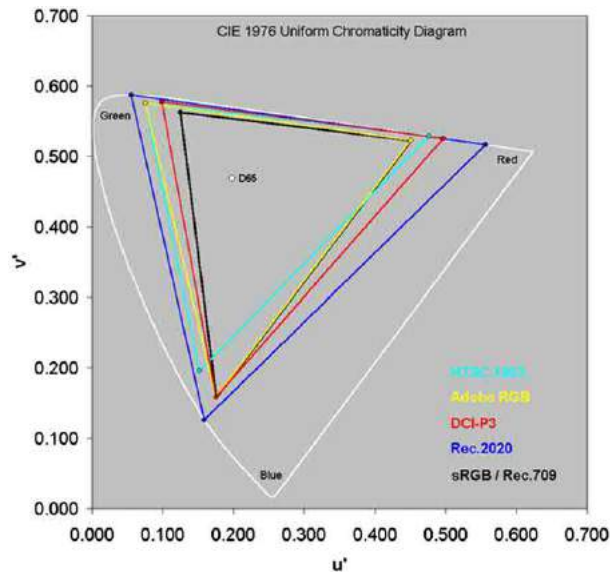


Fig. 19 Standard Color Gamut. Plotted on a CIE 1976 Uniform Chromaticity Diagram.

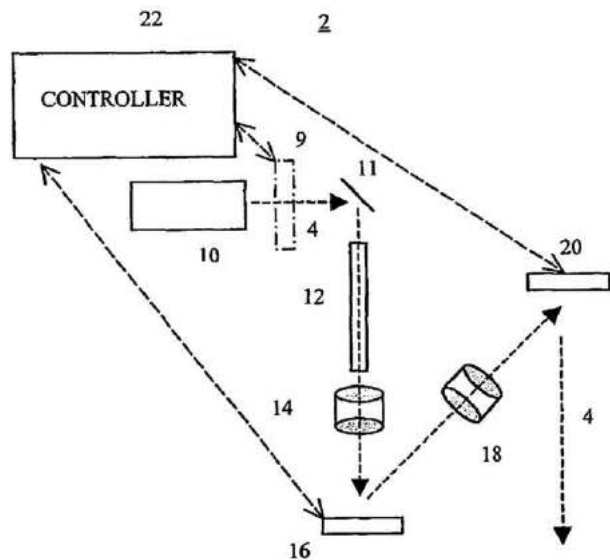


Fig. 20 US Patent 7224335 Assignee: IMAX DMD-based image display systems.

Dynamic irises (Laser phosphor is a lampless projection illumination platform that uses blue laser diodes as the primary light source. To generate the three primary colors – red, blue, green – the blue light from the laser diodes shines onto a spinning wheel that is coated in a phosphor compound) first came into general use with LCD projectors as a means for improving on this projector’s relatively low native on/off contrast ratio and elevated black levels. Dynamic irises can now be found in many home theater projectors using DLP, LCoS and 3LCD digital display technologies including those with a relatively good native contrast ratio and low black levels.

When SLM based projection systems are to be used to meet the HDR regulation, it is desirable that the system has a high spatial resolution and that the system possesses a wide dynamic range image quality. SLM based systems such as those based on DMDs are typically limited to contrast ranges of 1000: 1 or less. IMAX proposed invention dual SLM structure is directed towards a projection image display systems that employ so-called reflective spatial light modulators (SLMs) such as those in the form of digital micro-mirror devices (DMDs). It relates to techniques for using multiple SLM devices arranged serially. The invented system (Fig. 20) comprises a controller, a first reflective SLM and a second reflective SLM. The first reflective SLM is operatively coupled to the controller to receive first image data from the controller and is aligned to receive light from a light source and to reflect imaging light correlated to the first image the second reflective SLM is operatively coupled to the controller to receive second image data from the controller and is aligned to receive imaging light from the first reflective SLM and to reflect modified imaging light.

Assume DMD has 100% optical efficiency and supports n bit depth, if the first DMD maintains 2^k equally spaced brightness levels, a serial DMD device should have a dynamic range performance equivalent to $k + n$ bit depth, at least within the lowest level (US Patent 7224335). Hence, by applying dual-DMD, the contrast ratio could meet the HDR regulation by well design.

See also: Displays – Introduction

References

- Armitage, D., Underwood, I., Wu, S.-T., 2006a. Introduction to Microdisplays, p. 327.
 Armitage, D., Underwood, I., Wu, S.-T., 2006b. Introduction to Microdisplays, p. 328.
 Barco. Laser-phosphor illumination in projectors, White paper. Available at: <http://www.barco.com/en/Page/2016/Entertainment/Downloads/Laser-phosphor-illumination/Downloadpage>.
 Bone, M., 2001. 46.2: Method to improve contrast in an off-axis projection engine. SID Symposium Digest of Technical Papers, 32: 1180–1183. doi:10.1889/1.1831770.
 Brennesholtz, M.S., Stupp, E.H., 2008. Projection Displays, p119.
 Hutchison, D.C., Neal, H.W., 2008. The design and implementation of a stereoscopic microdisplay television. IEEE Trans. Consumer Electron. 54, 254–261.
 Im, H., Kim, W., Chung, J., 2010. A new transmissive type tilting actuator for a smooth projected television image. Proceedings IMechE 224 Part C: J. Mechanical Engineering Sciencepp. 1327–1333.
 Robinson, M.D., Sharp, G., Chen, J., 2005a. Polarization Engineering for LCD Projection, p. 187.
 Robinson, M.D., Sharp, G., Chen, J., 2005b. Polarization Engineering for LCD Projection, p. 192.
 Hutchison, D.A., 2007. The SmoothPicture™ Algorithm: An Overview, Texas Instruments, DLP™ TV.
 US Patent 6,183,091 B1, Color imaging systems and methods.
 US Patent 6,330,113, Color separation prism with adjustable path lengths.
 US Patent 7,198,373, B2, Display apparatus using LCD panel.
 US Patent 7,224,335, DMD-based image display systems.

Relevant Websites

- http://www.3lcd.com/explore/polarization_charger.aspx
 3LCD.
<http://www.businesswire.com/news/home/20160510005412/en/World%E2%80%99s-Smallest-Native-4K-Imaging-Device-Introduced>
 Business Wire.
<https://www.dolby.com/us/en/technologies/home/dolby-vision.html>
 Dolby.
<http://www.uhdalliance.org/uhd-alliance-press-releasejanuary-4-2016/#more-1227>
 UHD Alliance.

3D Displays, Stereoscopic/Autostereoscopic 3D

Yi-Pai Huang and Chun-Ho Chen, National Chiao Tung University, Hsinchu, Taiwan

© 2018 Elsevier Ltd. All rights reserved.

Introduction of 3D Display

In recent years, the requirement is emerging for more vivid and realistic 3D experience. The development of 3D display technology comes from the desire of human's nature vision, and results from the maturity of manufactory and display technology. The public usually record and display images for memorizing and sharing activities with others. Therefore, display is the closest electronics product to human. The progress of display development represents the efforts of pursuing more realistic visual experience. When we review the display developing history, human's color perception ability is a strong driving force to replace black-and-white TVs with full color ones. More than one trillion vision cells also encourage us to replace standard definition displays with full high definition ones. In addition to the progress of color and definition, a depth perception is a disruptive innovation of display technology. 3D displays utilize this depth perception to demonstrate the real 3D world (Holliman, 2005). Watching the primary 3D displays needs to wear eye glasses, which is called the stereoscopic type (Hill and Jacobs, 2006). The glasses usually block human's line of sight and do not be assimilated into our life. The scenarios could be limited in movie theaters (Srivastava *et al.*, 2010). Therefore, the glasses-free technology, or called the auto-stereoscopic type, obtains more attention to be applied on billboards, TVs, and mobile phones.

Principle

Our world is a three-dimension space but there is only two-dimension intensity distribution in our retina. However, the human's vision system can transfer this planar intensity distribution to the stereo concept by multi depth cues. Depth cues are classified by psychological and physiological depth cues (Pastoor and Wopking, 1997).

Psychological Depth Cue

Even the objects in the same distance, viewers perceive the closer objects when the objects are more colorful and brighter. Besides, human considers the relative small object is far than the big object in the same image. This is due to the perspective cue as shown in Fig. 1(a). The spacing of two parallel lines seems to decrease when the distance of lines away from the viewer increases. The stereo perception also comes from the variation of light and shadow as shown in Fig. 1(b). In addition, the occlusion cue between objects help determining near and far objects. The near object blocks the viewing of far objects as shown in Fig. 1(c). Moreover, the texture gradient in near objects contains more detail and surface feature on objects infer 3D shape and distance as shown in Fig. 1(d).

Physiological Depth Cue

The crystalline lens in the eye structure, exactly as the focus lens in the camera, can accommodate itself to image the object on the retina. And the viewing lines of two eyes converge on the object. Based on the different depth, this converging angle varies. Near objects have wide converging angles and far objects have narrow converging angles. The average distance between two eyes is 6.5 cm. This distance induces a little difference between the viewing angles of two eyes. Therefore, the image disparity of object on

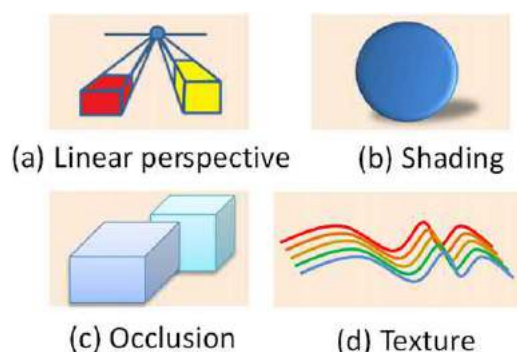


Fig. 1 Illustration of psychological depth cues such as (a) linear perspective: parallel lines converge at distance point on horizon, (b) shading: cast by an object gives a strong depth cue, (c) occlusion: invisible portion of objects behind an opaque object, and (d) texture: surface feature on objects infer 3D shape and distance.

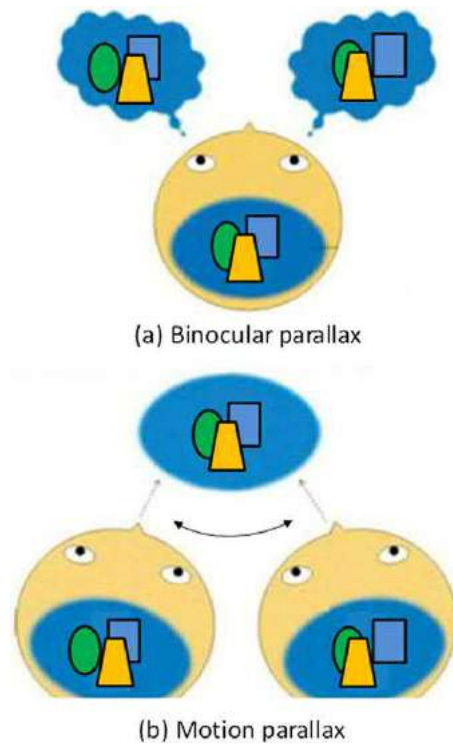


Fig. 2 (a) Binocular and (b) motion parallax provide the image disparity that our brain can merge to create 3D perception.

two retinas exists, called binocular parallax as shown in [Fig. 2\(a\)](#). According to this parallax, human can perceive the difference of object depth. Furthermore, when the viewer's eye moves, the image disparity of object also varies, called motion parallax as shown in [Fig. 2\(b\)](#). 3D displays apply the depth cues of parallax to realize stereo images. If the viewer wants to perceive a clear 3D image, two eyes can only receive individual images with the disparity.

Health Issue– Visual Fatigue

3D techniques use 2D left/right images and depth cue to yield 3D illusion in our brain. In the physiology, how the left/right images are received by the left/right eyes not only affects the 3D effectiveness but also the comfort level in the psychology. For examples, the interference between left and right image can degrade the 3D perception and may induce the visual fatigue ([Pastoor and Wopking, 1997](#)). One issue of 3D display is the accommodation–convergence conflict (A.C. conflict). While the human eye's convergence capability perceives the correct depth of 3D objects, the accommodation function makes the eyes focus on the display screen as shown in [Fig. 3\(b\)](#). Since there is a close interaction between convergence and accommodation, this A.C. conflict often causes visual fatigue in 3D viewers ([Pastoor and Wopking, 1997](#)). The relation between accumulated fatigue and eye disease is worth further exploring in the future. The 3D display revolution is still waiting for more research on physiological and psychological fields.

Stereoscopic and Auto-Stereoscopic Display

From the 19th to 20th century, 3D displays mainly belong to the stereoscopic type in need of special eye glasses. The glasses and other technical issues limit the popularization of 3D displays. By the end of the 20th century, more and more efforts have been made on the auto-stereoscopic display. These 3D displays are introduced as follows.

Polarizing Glass

The pixels of display are divided into odd rows and even rows. Micro-retarder film, arranged with corresponding rows, polarizes light into orthogonal polarization states ([Wu et al., 2008](#)). The viewer wears the glasses with polarization films for individual states. The binocular parallax can be achieved because another polarization will be blocked. The image interference could happen when view's head rotates with the linear polarization. Circular polarization is preferred to avoid this interference ([Jung et al., 2009](#)). The major issue is resolution decreasing, only a half of 2D images.

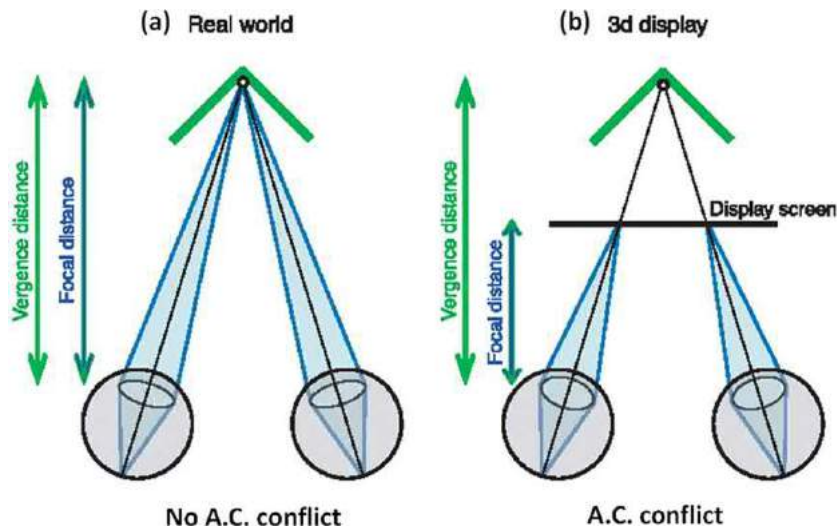


Fig. 3 In (a) real world, convergence and focal distance (accommodation) is the same; however, accommodation–convergence conflict (A.C. conflict) would happen on (b) 3D display. Reproduced from Hoffman, D.M., Girshick, A.R., Akeley, K., Banks, M.S., 2008. Vergence–accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of Vision* 8, 33. doi: 10.1167/8.3.33

Anaglyph

Anaglyph is also called red-green glasses as shown in Fig. 4. Complementary colors for individual eyes are presented in one image. When viewer wears the red-green glasses, the glasses can filter out another color light. The binocular parallax can be achieved. In anaglyph 3D, partial color is filtered out by color glasses, resulting in poor color performance. Anaglyph 3D has also the resolution issue compared to 2D images.

Shutter Glasses

The display alternatively plays right and left images with a double frame rate. The viewer wears the glasses with liquid crystal shutters, which is synchronized by the display. The shutter switches on and off sequentially and each eye can receive its corresponding image. Because of vision persistence, these left and right image sequences are fused together in our brain, achieving the stereo perception. This technology can overcome the resolution decreasing issue, but the drawback is higher cost than other passive glasses. The battery is also needed for the active glasses.

Head Mounted Display

This kind of 3D display provides left and right two displays just in front of our eyes. The head mounted display can solve the interference between left and right images due to independent display for each eye (Tomilin, 1999). Only one viewer can watch 3D at once. Usually, the device is expensive, heavy, and uncomfortable for long-time wearing (Keller *et al.*, 2008).

These stereoscopic displays effectively yield a pair of disparity images to produce our stereo perception. However, viewers must wear eyeglasses when they watch the stereoscopic display. Therefore, the auto-stereoscopic display, which is eyeglasses-free, was developed.

E-Holography

This auto-stereoscopic display consists of red/green/blue lasers as light sources and acoustic optical modulator (AOM) as a phase grating (Javidi and Tajahuerce, 2000). Through the vertical scanning mirror and polygonal mirror, the lasers scan in vertical and horizontal direction to form a phase hologram (Son *et al.*, 2010). The 3D image size is limited by the AOM size and the scan speed of polygonal mirror must be synchronized with lasers (Takaki and Okada, 2010). This system will meet great challenge in data processing and mirror scan speed if the system scales up (Takaki *et al.*, 2010).

Volumetric

This volumetric display is composed by the spiral surface and bottom laser light source (Soltan *et al.*, 1995). When the light is projected to the rotated spiral surface, the light is scattered as a point of image. The spiral surface will scan all points in the working space to form the whole image as shown in Fig. 4. This kind of 3D display presents a transparent image without depth cue of occlusion (Refai, 2009). Meanwhile, in the region close to rotating axial, the slower speed induces a partial blurring image (Gately *et al.*, 2011). These effects degrade the stereo perception.

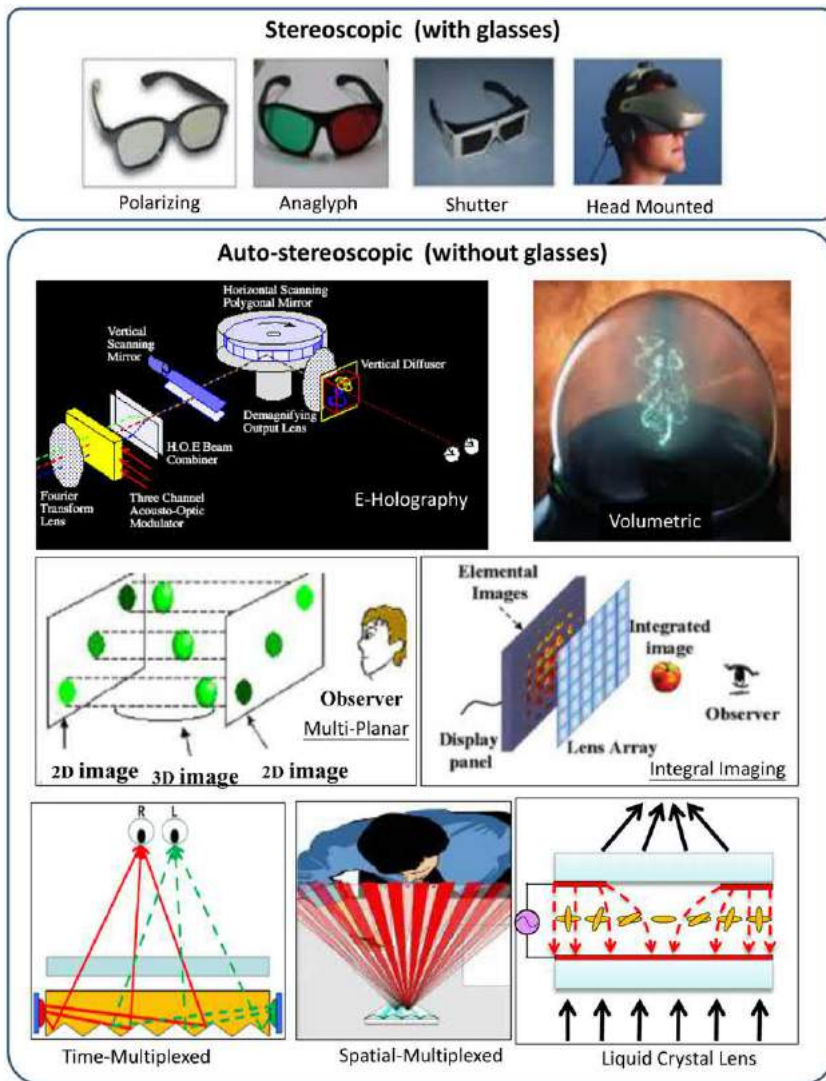


Fig. 4 3D displays classification based on with or without eyeglasses is divided into two groups, stereoscopic and auto-stereoscopic including various kinds of technologies (Jung *et al.*, 2009; Tomilin, 1999; Keller *et al.*, 2008; Javidi and Tajahuerce, 2000; Son *et al.*, 2010; Takaki and Okada, 2010; Takaki *et al.*, 2010, 2011; Soltan *et al.*, 1995; Refai, 2009; Gately *et al.*, 2011; Suyama *et al.*, 2004; Kuribayashi *et al.*, 2006; Hsu *et al.*, 2008; Berkel and Clarke, 1997; Young and Javidi, 2005; Chen *et al.*, 2009; Maphepo *et al.*, 2010; Tsuboi *et al.*, 2010; Takaki, 2010; Holliman *et al.*, 2011; Nakamura *et al.*, 2012; Hamagishi, 2009; Im *et al.*, 2008; Javidi and Okano, 2002; Son and Javidi, 2005; Sato, 1979; Woodgate, 2003, 2005; Willemsen and De Zwart, 2006; Krijn and de Zwart, 2008; Li and Wu, 2011; Ren and Wu, 2004; Naumov *et al.*, 1998; Ye and Sato, 2010; Chang *et al.*, 2014; Uehara *et al.*, 2013).

Multi-Planar

Two LCD panels are used in this multi-planar 3D display (Suyama *et al.*, 2004). These two LCD panels present the front and back image individually. The depth cues of various light and shadow in different viewing distance are applied to integrate front and back images. Precise alignment between the front and back panel could be a manufacturing issue (Kuribayashi *et al.*, 2006). In addition, only in the direct view direction, viewers can obtain the stereo perception. For other oblique viewing angle, the mismatch of two images cannot provide a 3D concept (Hsu *et al.*, 2008).

Time-Multiplexed

The LCD makers commonly apply the multiplexed type for 3D displays. One panel provides at least two views in time or spatial domain (Berkel and Clarke, 1997; Young and Javidi, 2005). In the time-multiplexed type, the left and right images are sequentially sent to left and right eye. Different from eyeglasses of stereoscopic display, the light guiding devices are implemented inside the LCD panel (Holliman *et al.*, 2011). When the switching speed is higher than that of our brain can distinguish,

the fusion of stereo images occurs. This 3D display preserves the resolution but requires a high frame rate LCD panel (Hamagishi, 2009).

Spatial-Multiplexed

Lenticular lens is attached on the LCD panel for guiding light from odd/even column pixels to left/right eyes (Im *et al.*, 2008). The cylindrical lens can focus and refract light from one image into more than two sub-images (Chen *et al.*, 2009; Takaki, 2010). The number of sub-images determines the number of views in stereo perception. More views can provide more continuous 3D viewing concept when viewers are moving their heads. In addition, parallax barrier is also popular as a light splitting device for disparity images (Maphopo *et al.*, 2010). The barrier is a black and transparent strip lines to restrict the light path from LCD panel to eyes. One eye of viewers only receives light from a specific LCD region (e.g., the odd or even column pixels); the other one eye receives light from the other region. Compared to the lenticular lens, the light intensity of barrier 3D display is lower due to the barrier's blocking light (Tsuboi *et al.*, 2010).

Integral Imaging Display

An integral imaging display uses the micro-lens array instead of the cylindrical lenses used by lenticular lens display (Javidi and Okano, 2002). Each micro lens directs light from the pixels and covers in different horizontal and vertical directions, thus producing a full parallax image. Therefore, the floating 3D image can be achieved by integral photography theory (Son and Javidi, 2005). The integral imaging display has an issue of image overlapping in oblique viewing angles. By modifying the pixel emitting into correct micro lens, this issue can be overcome (Javidi and Okano, 2002).

LC Lens for a 2D/3D Switchable Display

Spatial-multiplexed display degrades the image resolution, even just presents 2D contents. The advantage of LC lens-based switchable display is to preserve the high image resolution in 2D display mode (Sato, 1979). By switch-on the LC lens, display will show the 3D images (Ren and Wu, 2004; Naumov *et al.*, 1998; Ye and Sato, 2010). Three types of LC lens, polarization active micro-lens (Li and Wu, 2011), active LC lenticular lens (Krijn and de Zwart, 2008), and electric-field-driven (Willemsen and De Zwart, 2006), had been reported and successfully demonstrated. LC alignment direction (LCAD) (Woodgate, 2005) and electrode array direction (EAD) (Woodgate, 2003) was then proposed to have 3D rotational function for mobile display (Chang *et al.*, 2014; Uehara *et al.*, 2013). Additionally, the temporal scanning LC-lens (Huang and Chen, 2012), such as multi-electrode driven (MeD) Fresnel LC lens was proposed to improve the resolution of 3D images (Huang *et al.*, 2010).

Super-Multi-View Display

Due to the limitation of the number of multi-view, there is always discontinuity of view switching with respect to the viewing direction. This effect degrades the 3D perception. Super-multi-view (SMV) techniques use an extremely large number of views (e.g., 256 or 512 views) to generate more natural 3D display visual effects (Takaki *et al.*, 2011). The horizontal interval between views is smaller than the diameter of the human pupil. Light from slightly different viewpoints enters the pupil simultaneously. This is called the SMV display condition (Nakamura *et al.*, 2012). SMV provides smooth motion parallax and reduces the A.C. conflict since the brain unconsciously predicts the image change due to motion.

Augmented and Virtual Reality Display

Although wearing a head mounted display (HMD) is not very convenient for users, but Augmented Reality (AR) and Virtual Reality (VR) technologies can produce the virtual images to make people's brain perceived as a real 3D object. In addition, the hardware devices have been improved to be much lighter than that in the 20th century. AR/VR HMD becomes a popular system for many applications. The major difference between AR and VR is that VR make people thoroughly immersed with an ocular head mounted device, interactivity with sensory feedback. While AR only produces the virtual images covering part of observer's field of view and integrate the images with the environmental background.

There are several optical arrangements to render the images for AR and VR applications. For an occlusion type, mainly for VR application, the display can be placed closed to the eyes, then use optical lens to generate a virtual image with large field of view, as shown in Fig. 5(a). While in the case of see-through type, mainly for AR applications, the optical setup includes off-axis and partially reflective mode, light guided mode, or total internal reflection mode, to combine the virtual image with environmental scenes, as shown in Fig. 5(b).

The applications of AR/VR displays are not only limited on consumer electronics, such as video games, video entertainment, or live events. This technology can be further used on healthcare/medical, engineering, education, retail, and military field. Such as in the case of medical application, the VR display combined with a brain-wave detector can be used for detecting glaucoma damage

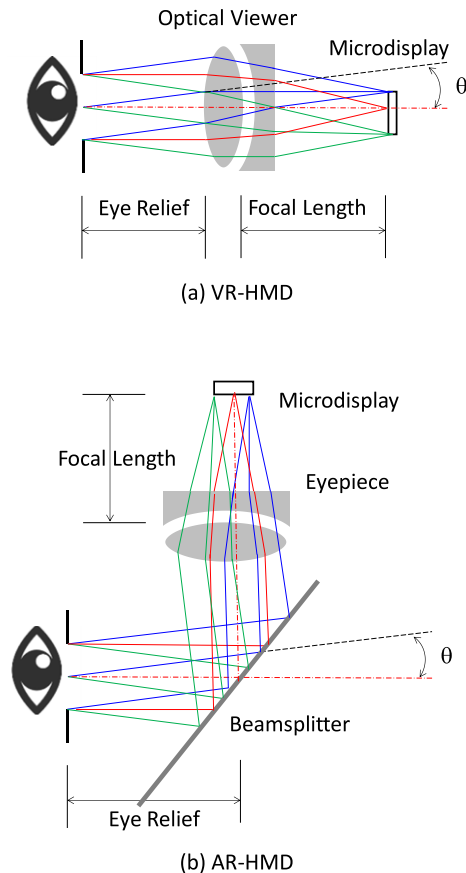


Fig. 5 Optical structure of (a) virtual reality head mounted display. The lens is placed in front of the eye and micro display located near eye is floated to the desired depth. (b) Augment reality head mounted display. The micro display is located at the side part of the observer and the displayed virtual information is bent to the eye by the optical combiner.

(Chien *et al.*, 2013; Lin *et al.*, 2015). It could even potentially allow users to routinely and continuously monitor the electrical brain activity associated with visual field in their living environments, representing a transformative way of diagnosing and monitoring disease progression in not only glaucoma but other functional visual losses.

See also: Displays – Introduction. Liquid Crystal Physics and Materials

References

- Berkel, C.V., Clarke, J.A., 1997. Characterization and optimization of 3D-LCD module design. *Proc. SPIE* 3012, 179–187.
- Chang, Y.C., Jen, T.H., Huang, Y.P., 2014. High-resistance liquid crystal lens array for rotatable 2D/3D auto-stereoscopic display. *Opt. Express* 22, 2714–2724.
- Chen, C.H., Huang, Y.P., Chuang, S.C., *et al.*, 2009. Liquid crystal panel for high efficiency barrier type autostereoscopic three-dimensional displays. *Appl. Opt.* 48, 3446–3454.
- Chien, Y.Y., Lin, F.C., Chou, C.C., *et al.*, 2013. Polychromatic high-frequency steady-state visual evoked potentials for brain-display interaction. In: *SID Symposium Digest of Technical Papers*, vol. 44, pp. 146–149.
- Gately, M., Zhai, Y., Yeary, M., Petrich, E., Sawalha, L., 2011. A three-dimensional swept volume display based on LED arrays. *J. Display Technol.* 7, 503–514.
- Hamagishi, G., 2009. Analysis and improvement of viewing conditions for two-view and multi-view display. In: *Proceeding of the SID Symposium Digest*, vol. 40, pp. 340–343.
- Hill, L., Jacobs, A., 2006. 3-D liquid crystal displays and their applications. *Proc. IEEE* 94, 575–590.
- Holliman, N., 2005. 3D display system. In: *Handbook of Optoelectronics*. Institute of Physics Press. ISBN: 0 7503 0646.
- Holliman, N.S., Dodgson, N.A., Favalora, G.E., Pockett, L., 2011. Three-dimensional displays: A review and applications analysis. *IEEE Trans. Broadcasting* 57, 362–371.
- Hsu, C.Y., Huang, Y.P., Chang, Y.C., 2008. Wide viewing angle 3D depth-fused display (DFD) system. In: *Proceeding of the SID Symposium Digest*, vol. 39, pp. 1655–1658.
- Huang, Y.P., Chen, C.W., 2012. Superzone Fresnel liquid crystal lens for temporal scanning auto-stereoscopic display. *IEEE/OSA J. Display Tech.* 8 (11), 650–655.
- Huang, Y.P., Liao, L.Y., Chen, C.W., 2010. 2D/3D switchable autostereoscopic display with multi-electrically driven liquid crystal (MeD-LC) lenses. *J. Soc. Inf. Display* 18 (9), 642–646.
- Im, H.J., Jung, S.M., Lee, B.J., Hong, H.K., Shin, H.H., 2008. Mobile 3-D display based on a LTPS 2.4-in. VGA LCD panel attached with lenticular lens sheets. In: *Proceeding of the SID Symposium Digest*, vol. 39, pp. 256–259.

- Javidi, B., Okano, F., 2002. Three-dimensional television, video, and display technologies. New York: Springer-Verlag.
- Javidi, B., Tajahuerce, E., 2000. Three dimensional object recognition using digital holography. *Opt. Lett.* 25, 610–612.
- Jung, S.M., Park, J.U., Lee, S.C., *et al.*, 2009. A novel polarizer glasses-type 3D displays with an active retarder. In: *Proceeding of the SID Symposium Digest*, vol. 40, pp. 348–351.
- Keller, K., State, A., Fuchs, H., 2008. Head mounted displays for medical use. *J. Display Technol.* 4, 468–472.
- Krijn, M.P.C.M., de Zwart, S.T., 2008. 2-D/3D displays based on switchable lenticulars. *J. Soc. Inf. Display* 16/8, 847–855.
- Kuriyayashi, H., Date, M., Suyama, S., Tatada, H., 2006. A method for reproducing apparent continuous depth in a stereoscopic display. *J. Soc. Info. Display* 14, 493–498.
- Lin, F.C., Chien, Y.Y., Zao, J.K., *et al.*, 2015. High-frequency polychromatic visual stimuli for new interactive display systems. *SPIE Newsroom*. doi:10.1117/2.1201504.005851.
- Li, Y., Wu, S.T., 2011. Polarization independent adaptive microlens with a blue-phase liquid crystal. *Opt. Express* 19, 8045–8050.
- Maphepo, W., Huang, Y.P., Shieh, H.P.D., 2010. Enhancing the brightness of parallax barrier based 3D flat panel mobile displays without compromising power consumption. *J. Display Technol.* 6, 60–62.
- Nakamura, J., Takahashi, T., Chen, C.W., Huang, Y.P., Takaki, Y., 2012. Analysis of longitudinal viewing freedom of reduced-view super multi-view display and increased longitudinal viewing freedom using eye-tracking technique. *J. Soc. Info. Display* 20, 228–234.
- Naumov, A.F., *et al.*, 1998. Liquid-crystal adaptive lenses with modal control. *Opt. Letters* 23 (13), 992–994.
- Pastoor, S., Wopking, M., 1997. 3-D displays: A review of current technologies. *Displays* 17, 100–110.
- Refai, H.H., 2009. Static volumetric three-dimensional display. *J. Display Technol.* 5, 391–397.
- Ren, H.W., Wu, S.T., 2004. Getting in tune: Electronically controlled liquid crystal yields tunable-focal-length lenses. *SPIE's Optics Express Magazine*. 24–27.
- Sato, S., 1979. Liquid-crystal lens-cells with variable focal length. *Jpn. J. Appl. Phys.* 18, 1679–1684.
- Soltan, P., Trias, J., Dajke, W., Lasher, M., McDonald, M., 1995. Laser based 3-D volumetric display system. *Naval Engineers J.* 107, 233–243.
- Son, J.Y., Javidi, B., 2005. Three-dimensional image methods based on multi-view images. *J. Display Technol.* 1, 125–140.
- Son, J.Y., Javidi, B., Yano, S., Choi, K.H., 2010. Recent developments in 3D image technologies. *J. Display Technol.* 6, 394–403.
- Srivastava, A.K., de, J.L., de la Tcnaye, B., Dupont, L., 2010. Liquid crystal active glasses for 3D cinema. *J. Display Technol.* 6, 522–530.
- Suyama, S., Ohtsuka, S., Takada, H., Uehira, K., Sakai, S., 2004. Apparent 3-D image perceived from luminance-modulated two 2-D images displayed at different depths. *Vis. Res.* 44, 785–793.
- Takaki, Y., 2010. Multi-view 3-D display employing a flat-panel display with slanted pixel arrangement. *J. Soc. Info. Display* 18, 476–482.
- Takaki, Y., Okada, N., 2010. Reduction of image blurring of horizontally scanning holographic display. *Opt. Express* 18, 11327–11334.
- Takaki, Y., Tanaka, K., Nakamura, J., 2011. Super multi-view display with a lower resolution flat-panel display. *Opt. Express* 19, 4129–4139.
- Takaki, Y., Yokouchi, M., Okada, N., 2010. Improvement of grayscale representation of the horizontally scanning holographic display. *Opt. Express* 18, 24926–24936.
- Tomilin, M.G., 1999. Head-mounted displays. *J. Optical Technol.* 66, 528–533.
- Tsuboi, M., Kimura, S., Takaki, Y., Horikoshi, T., 2010. Design conditions for attractive reality in mobile-type 3-D display. *J. Soc. Info. Display* 18, 698–703.
- Uehara, M.K.S., Takagi, A., Baba, M., 2013. LC GRIN lens mode with wide viewing angle for rotatable 2D/3D tablet. In: *Proceeding of the SID Symposium Digest*, pp. 154–157.
- Willemsen, O.H., De Zwart, S.T., 2006. 2-D/3D switchable displays. *J. Soc. Inf. Display* 14/8, 715–722.
- Woodgate, G.J., 2003. High efficiency reconfigurable 2D/3D autostereoscopic display. In: *2003 SID International Symposium Digest of Technical Papers*, pp. 394–397.
- Woodgate, G.J., 2005. A new architecture for high resolution autostereoscopic 2D/3D displays using free-standing liquid crystal microlenses. In: *2005 SID International Symposium Digest of Technical Papers*, pp. 378–381.
- Wu, Y.J., Jeng, Y.S., Yeh, P.C., Hu, C.J., Huang, W.M., 2008. Stereoscopic 3D display using pattern retarder. In: *Proceeding of the SID Symposium Digest*, vol. 39, pp. 260–263.
- Ye, M., Sato, S., 2010. Low-voltage-driving liquid crystal lens. *Jpn. J. Appl. Phys.* 49, 100204-1–100204-3.
- Young, S.J., Javidi, B., 2005. Three-dimensional image methods based on multi-view images. *J. Display Technol.* 1, 125–140.

Wearable Displays

Paul Yang, Sun Yat-Sen University, Guangzhou, China

© 2018 Elsevier Ltd. All rights reserved.

Spectacle-Type

Spectacle-type of wearable display, also known as head mounted display (HMD), includes two major categories, augmented reality (AR) HMD and virtual reality (VR) HMD. Both AR and VR technologies produce the virtual images for human eyes and make people's brain perceived as something really exists. The major difference between AR and VR is that VR make people thoroughly immersed with ocular device, while AR only produce the virtual images covering part of observer's field of view and integrate the images with the environmental facilities.

There are several optical arrangements to render the images for AR and VR applications. For occlusion type, mainly for VR application, the display can be placed very closed to the eyes, or use mirror to direct the light into the human pupil to render images, and maximize the field of view (FOV), as shown in Fig. 1(a).

While in the case of see-through type, mainly for AR applications, the optical setup includes off-axis and partially reflective mode, light guided mode, and TIR mode, to maximize the observation of environmental science, which thus decreased the FOV, as shown in Fig. 1(b). For different application segments, the optical designs may differ. For example, in the smart glasses application which can function like handless cellphone, the aesthetic appearance is more essential, while in the case of medical application, the resolution and contrast become much more important. As shown in Fig. 2(a) and (b), it is very helpful for a doctor to use

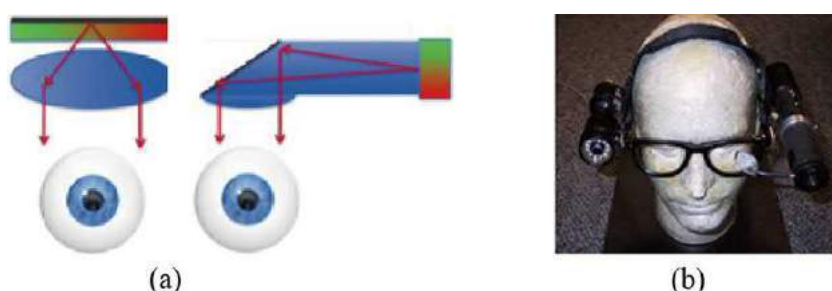


Fig. 1 The optical setup of AR and VR includes two basic arrangements: (a) totally occlusion with near-eye display and (b) optical combiner with light-guiding system. Reproduced from Kress, B., Saeedi, E., Brac-de-la-Perriere, V., 2014. The segmentation of the HMD market: Optics for smart glasses, smart eyewear, AR and VR headsets. In: Kazemi, A.A., Kress, B.C., Mendoza, E.A., *et al.* (Eds.), *Photonics Applications for Aviation, Aerospace, Commercial, and Harsh Environments V*, vol. 9202. Bellingham: Spie-Int Soc Optical Engineering; Bryant, R.C., Lee, C.M., Burstein, R. A., Seibel, E.J., 2004. 59.2: Engineering a low-cost wearable low vision aid based on retinal light scanning. *SID Digest* 35 (1), 1540–1543.

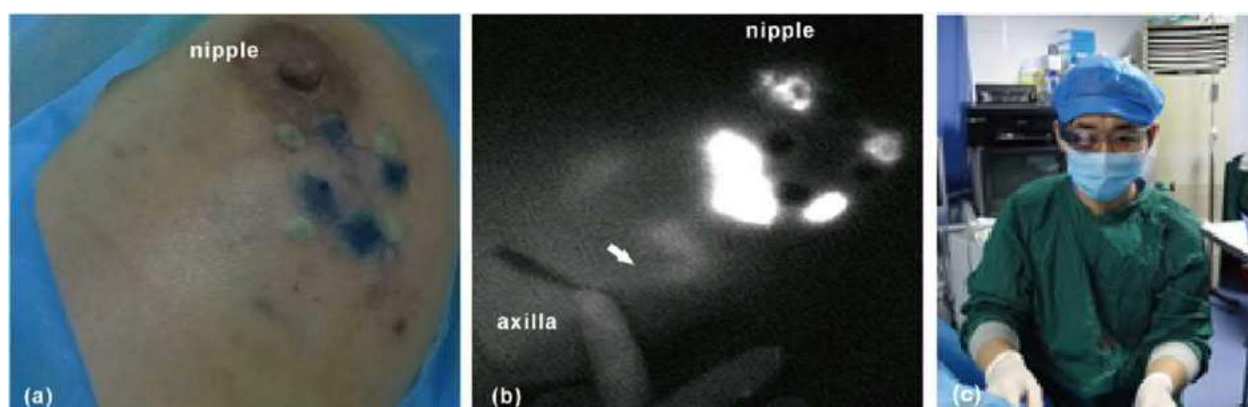


Fig. 2 The demonstration of using AR wearable glasses for medical operation: (a) the photograph of patient's breast, (b) the corresponding fluorescence image, and (c) the appearance of doctor wearing the AR glasses. Reproduced from Zhang, Z.S., Pei, J., Wang, D., *et al.*, 2016. A google glass navigation system for ultrasound and fluorescence dual-mode image-guided surgery. In: VoDinh, T., Mahadevan-Jansen, A., Grundfest, W.S. (Eds.), *Advanced Biomedical and Clinical Diagnostic and Surgical Guidance Systems XIV*, vol. 9698. Bellingham: Spie-Int Soc Optical Engineering.

fluorescent image to identify the ordinary cell and the tumor cell. In this application, the doctor can conveniently observe by wearing the smart glasses.

Clothes-Integratable-Type

With the upcoming trend of “internet of things, IoT,” people expect every device can be integrated with clothes or environments, which can make the users can “hands off.” To realize this fantasy, the devices need to be flexible and stretchable, and so do the displays. Making display flexible is of one route, while making it stretchable is the other. Here, we introduce essential technologies for achieving flexible and stretchable displays, and therefore, these technologies can be combined for realizing clothes integrated type of wearable display, as shown in Fig. 3 (Lee *et al.*, 2016).

Flexible Displays Integrating With Clothes

Some wearable displays integrated with clothes were demonstrated, which are expected to provide more functions for users beyond the current smart phone applications. However, the appearance is still not aesthetic enough for fashion design, because the existing technology cannot be stretchable and conformable for fashion designers to realize their design (Fig. 4).

Flexible Displays Integrating With Human Body

Along with the development of flexible electronics, more and more sensors were prepared that contributes to human’s convenience. For the further application, the integration with human body makes these devices more attractive.

Therefore, sensors are expected to be attached to the human body and people will need to know what the sensors detected. For example, the detected temperature sends the signal of your body condition that is needed to be known timely and a human-sensor interaction is therefore essential for practicability. In this regard, as an information carrier, a wearable display was also needed in these applications. The integration with human body makes this kind of flexible display different from the conventional. There are several design considerations should be included. Because the direct contact with human skin, the attached devices should not affect human’s health as well as the ability of other activities. On one hand, the flexibility should be fully realized that does not

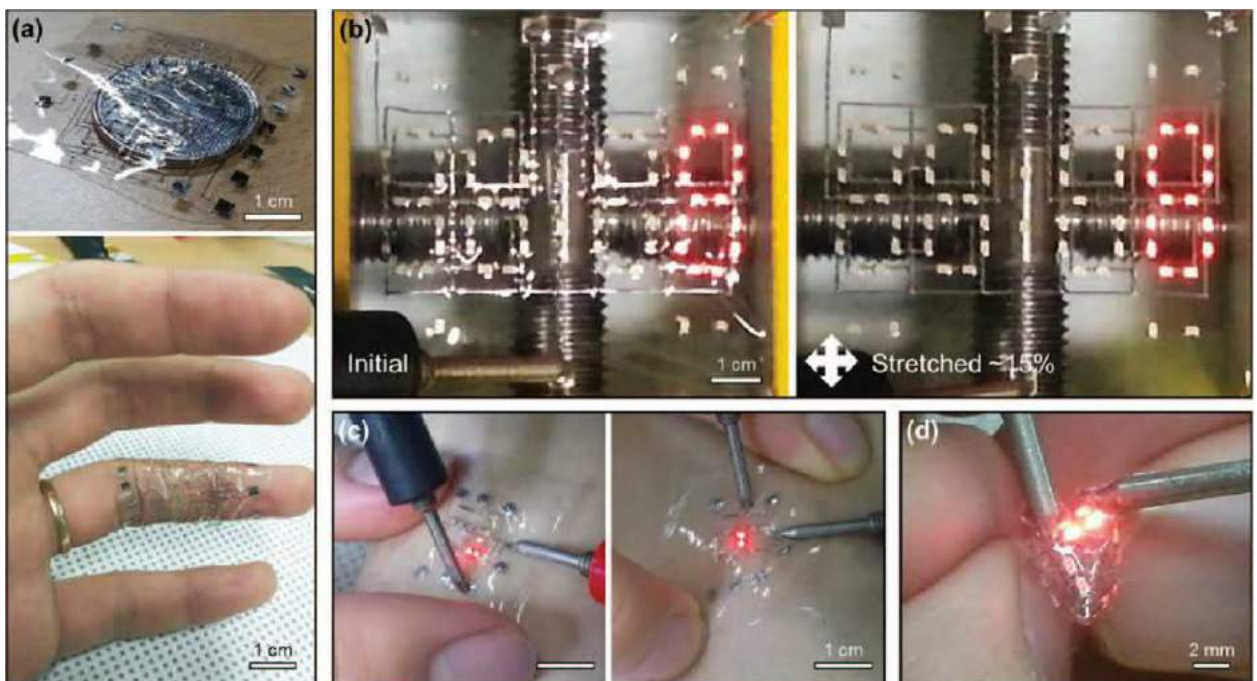


Fig. 3 (a) Photographs of freestanding stretchable 4-digit 7-segment on coin (upper) and skin (lower). (b–d) Photographs of operation of stretchable all-ink-jet-printed 4-digit 7-segment display with initial and stretched state (b), when attached to skin (c) and crumpled (d). Reproduced from Lee, B., Byun, J., Oh, E., *et al.*, 2016. All-ink-jet-printed wearable information display directly fabricated onto an elastomeric substrate. *SID Digest 47 (1)*, 672–675.



Fig. 4 The prototypes of the wearable-display wrist unit. Reproduced from Ma, R.Q., Rajan, K., Silvernail, J., *et al.*, 2010. Wearable 4-in. QVGA full-color-video flexible AMOLEDs for rugged applications. *Journal of the Society for Information Display* 18, 50–56; Tajima, R., Miwa, T., Oguni, T., *et al.*, 2014. Truly wearable display comprised of a flexible battery, flexible display panel, and flexible printed circuit. *Journal of the Society for Information Display* 22, 237–244.

obstruct human's activities. To satisfy this requirements, our display should be flexible and stretchable to adapt muscle movement. So an integration with several rigid substrates cannot be accepted. On the other hand, the materials of good biological compatibility and breathability should be chosen as the flexible substrates. And the complex operating environment of human body raise new demands of endurance. Taking these considerations into account, the wearable display can not only succeed the advantages of light weight and facile fabrication of flexible devices and also be applied on human body that makes human's life more convenient.

Watch-Type

Display makers tried very hard to make the watch-type of display looks like conventional, especially the appearance needs to be round. The pixel array used in display is usually in rectangular arrangement. By applying special algorithm to dim the lightness a bit around the boundary, it is possible to render round watch appearance with conventional rectangular pixel arrangement (**Fig. 5 (a)**). In addition, by folding the COF underneath the watch, the form factor can be as slim as the actual watch.

Flexible Substrates

Flexible displays are made with flexible components including substrates, thin film transistors (TFTs), and display medium. Usually, the devices need to be annealed to stabilize and optimize the performances. However, flexible for ordinary TFT device processing. Polyethylene naphthalate (PEN) has better thermal durability, however, ordinary TFT needs 350°C of post-annealing, it is still too low in this case. Polyimide (PI) can endure 350°C and its thermal expansion coefficient matches with substrate made of plastics are not endurable for high temperature processing, while flexible substrate made of steel are too heavy to be made for mobile devices. As shown in **Table 1**, polyethylene terephthalate (PET) is the cheapest one, however, its glass transition temperature (T_g) is too low for other layers of a flexible device, however the cost of PI is still too high for consumer products. As a result, scientists are paying much attention develop a substrate with simultaneously high T_g , high transparency (for aesthetic purpose), high flexibility, and low cost.

Flexible TFT

Flexible TFT is also a critical component of flexible display, taking charge of controlling the grey scale of each pixel. The desired features of a good flexible TFT array are high switching speed, low power consumption, high aperture ratio, good flexibility, and reliable electrical performances. The material characteristics for these requirements are high mobility of semiconducting layer, low resistance of connection metal line, small capacitance coupling between layers, smaller area of TFT blocking the transmissive light, high power efficiency by eliminating the unwanted polarizers or color filters, capability of active layers to be deformed, and the stable material characteristics after photo or electrical stressing.

Here is the summary of flexible TFTs, as shown in **Table 2**. Organic TFT has the best capability of being bent, however, its mobility is still too low for high end applications, such as smart phone. Amorphous Si TFT (*a*-TFT) is the most mature TFT, well adopted in existing displays, however, its flexibility is not good enough for future wearable displays. While oxide TFT seems to have better overall performances in terms of flexibility, mobility, cost, and compatibility with existing manufacturing facilities. However, the oxide TFT still needs to be improved for its long-operation reliability, which is induced by the stressing and resulting the electrical performance degradation.

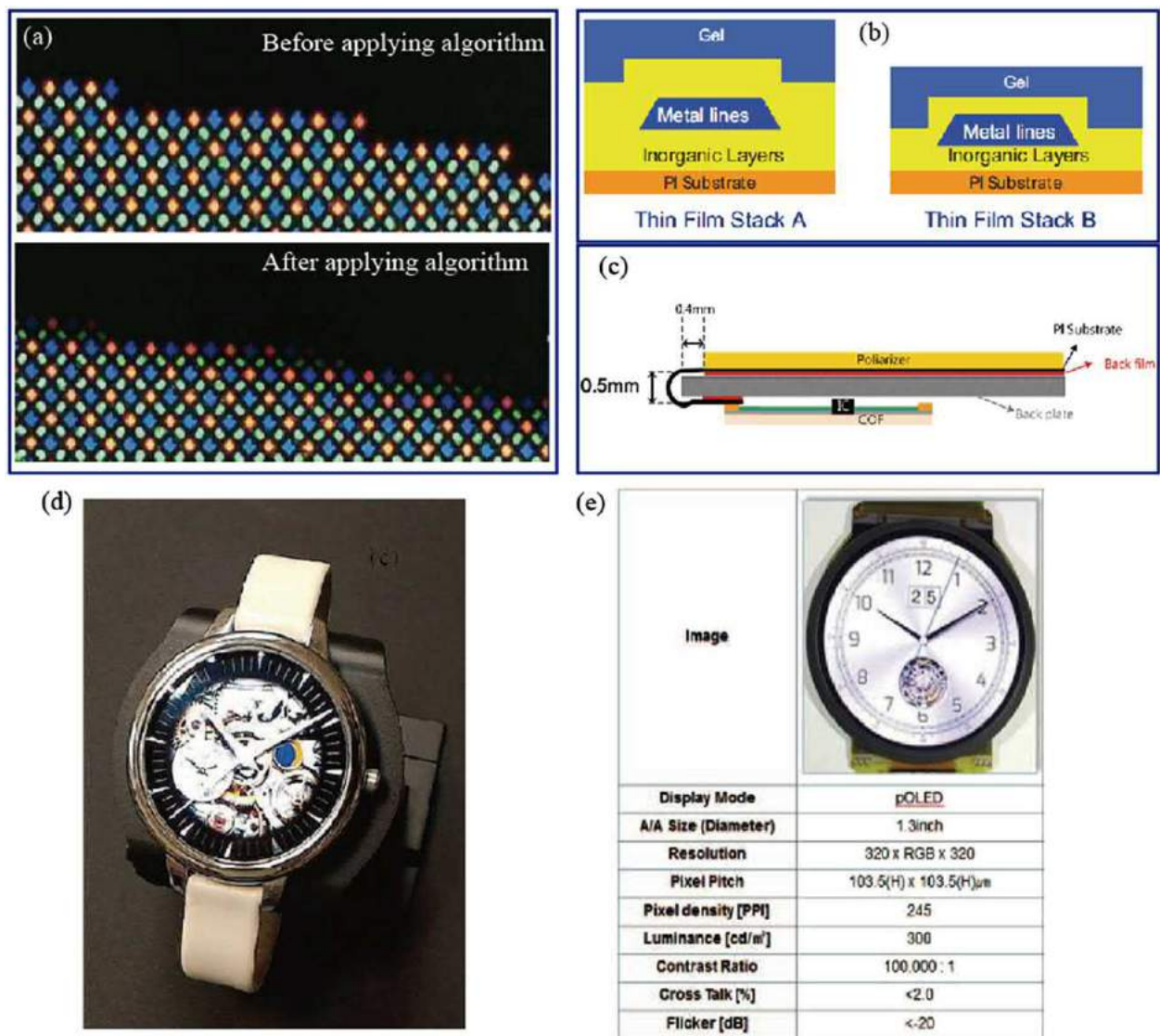


Fig. 5 (a) Image of before and after applying algorithm, (b) two thin film stacks with different total thickness of inorganic layers, (c) bending structure of the panel, (d) 1.26" circular plastic AMOLED panel, and (e) specification of circular plastic OLED display. Reproduced from Jung, D.S., Song, S.M., Kang, H.K., *et al.*, 2016. Panel design technology for circular OLED display. *SID Digest* 47 (1), 844–846; Li Fong, L., Chia Hsum, T., Ko Ruey, J., *et al.*, 2016. A circular flexible AMOLED display with a 1-mm slim border. *SID Digest* 47 (1), 847–850

Table 1 Basic properties of film used for base substrates

	PET	PEN	PC	PES	PAR	PCO	PI
CTE (–55 to 85°C) ppm/°C	1.5	13	60–70	54	53	74	17
Transmission at 400–700 nm (%)	>85	>85	>90	90	90	91.6	yellow
Water absorption	0.14	0.14	0.2–0.4	1.4	0.4	0.03	1.8
Young's modulus (GPa)	5.3	6.1	1.7	2.2	2.9	1.9	2.5
Tensile strength (MPa)	22.5	275	NA	83	100	50	231

To accommodate the exiting processes for flexible TFT array, the plastic substrate is usually bonded onto the rigid substrate with releasable adhesive. As shown in Fig. 6(a), the TFTs were formed onto a pre-laminated substrate with the structure of plastic/adhesive/carrier-plate by conventional photo-lithography, and be peeled off after the TFT array finished. Other modified approaches include using sacrificing layer as shown in Fig. 6(b) and using transfer process as shown in Fig. 6(c). They are all based on the concept of adopting existing TFT process, while forming the TFT arrays on the flexible substrates, as shown in Fig. 7.

Table 2 The comparison of flexible TFTs

		<i>Poly-Si</i>	<i>Amorphous-Si</i>	<i>Organic TFT (Pentacene)</i>
Performance	Mobility	Very Good	OK for PHOLEDs	OK
	Leakage	OK	Very good	OK
	Stability	Good	Issue	Issue
	Uniformity	Issue	OK	Issue
Manufacturability		Maturing	Excellent	N/A
Cost		> Medium	Medium	Low??
Plastic-substrate compatibility		Under development	OK	Excellent
Metal-foil-substrate compatibility		Good	Good	Good

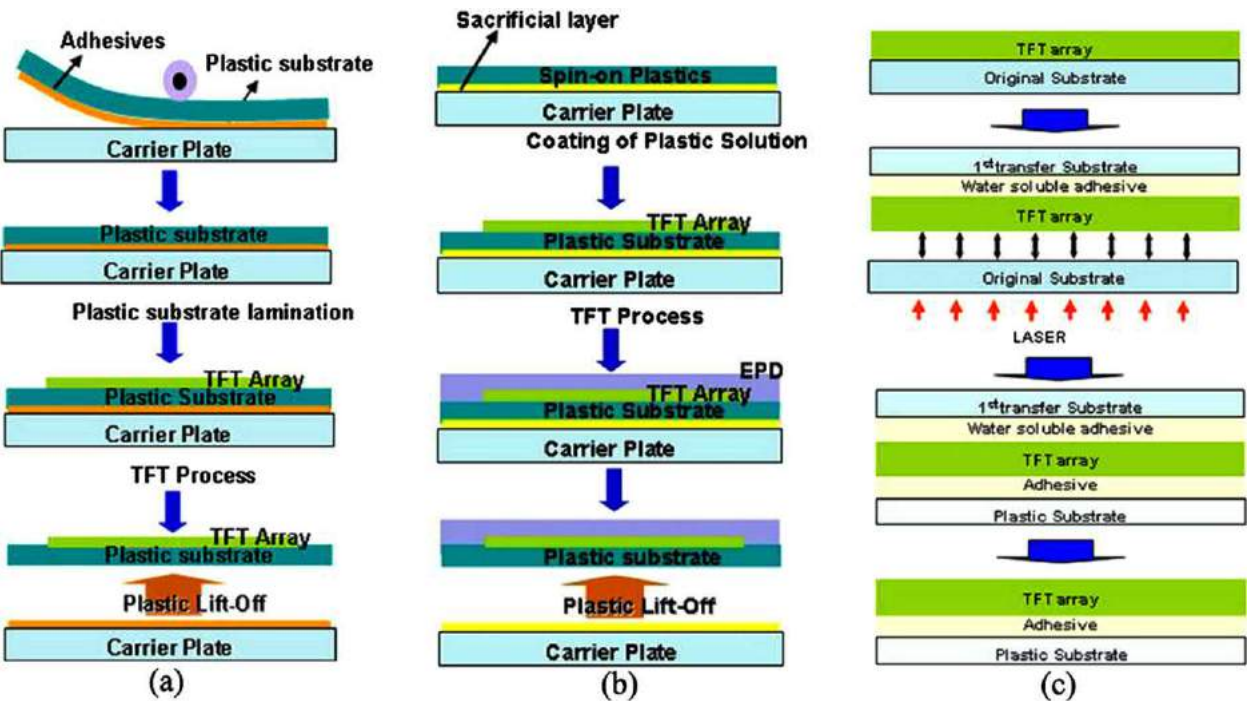


Fig. 6 The handling process for making flexible TFT array. Reproduced from Li, Z.M., Lv, Q.M., Huang, Y.J., 2010. Versatile soft backplane Technology. Journal of Industrial Materials.

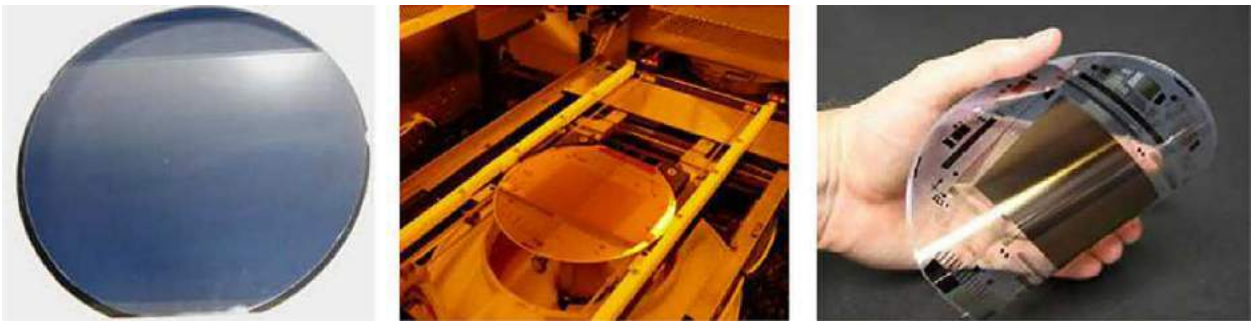


Fig. 7 A handling process for making flexible TFT array. Reproduced from Li, Z.M., Lv, Q.M., Huang, Y.J., 2010. Versatile soft backplane Technology. Journal of Industrial Materials.

Flexible Organic Light Emitting Diode

There are two major kinds of flexible display made of organics: small molecules, and polymer materials (usually called polymer light emitting diode, PLED, to differentiate with small molecular organic light emitting diode, OLED). The former one was firstly invented by Dr. Ching W. Tang in Kodak, and the latter was invented by Prof. Richard Friend's team at University of Cambridge. The lighting mechanism of OLED is to inject both electrons and holes from cathode and anode, respectively, where both carriers would transport in the organic layers to form excitons, and later on illuminate the light and go back to the energy ground state. To optimize the lighting efficiency, the rate of injection, transportation, and recombination of electrons and holes should be delicately designed.

Flexible OLED can perform the best image quality, however, it is very vulnerable to moisture and oxygen, which thus needs to combine with flexible passivation technology before it can be commercialized (the structure is shown in Fig. 8; Ma *et al.*, 2010). The mature technology used for passivation is the multi-stacking of organics and inorganics, where inorganics are used for blocking the oxygen or water molecules's diffusion, and the organics are used for balancing the inter-layer stress while being bent. The other popular method is called atomic layer deposition (ALD), which is made with certain reactive precursors, mixing with reacting gas and precisely forming an atomic layer for passivating the OLED. As the layer is of atomic level, the device is ultimately flexible.

Flexible E-Paper

Different from the lighting mechanism of flexible OLEDs, flexible electrophoretic displays (EPD) can reflect the ambient light to show the image. Because the image appearance is similar to actual paper, it is usually called E-paper. Not only EPD, but also many other kinds of reflective display technologies are promising for future E-paper applications, such as liquid powder, electro-wetting, electro-chromic, cholesteric liquid crystal, polymer dispersed liquid crystal displays, and so on. Among them, EPD is the first flexible display commercialized for e-reader applications, however its color image performance is not as good as flexible OLED, and it is not suitable for video application because of its slow updating speed. Even though, EPD has two unique advantages superior to flexible OLED, one is the good sunlight readability due to its particle scattering, and the other is the power-saving due to its bistability.

The fabrication processes of making E-paper includes two major categories: microcapsule type and micro-cup type, and both of them can be leveraged for future flexible display fabrication. Both type of E-paper includes the synthesis of particles, formulation with surfactant and additives, encapsulated with a small chamber (either capsule or cup), and then laminated onto the flexible substrate with roll-to-roll (R2R) processes.

To realize a full-color E-paper, many approaches have been proposed. The easiest way is to add on color-filter (Lai *et al.*, 2012), however, considering the low reflective efficiency, the image was not good enough. Usually users do not like this kind of vague color image. Until recently, E-ink demonstrated a structure utilizing multiple color particles within one microcup to render color images by selectively control certain particle to be on top of the surface (Telfer and McCreary, 2016). This approach has the best performance in terms of contrast, color saturation, and resolution. However, such method needs to mix many kinds of color particle in one small chamber, and thus the fabrication repeatability is hard to control and the driving scheme is quite complicated. To overcome such difficulty, transfer method to spatially deploy the particle color particles to sub-pixels, which only need two kinds of color particles within one chamber was demonstrated, and thus the process is much easier to control and the driving scheme is comparably simple.

Flexible Liquid Crystal Display

LCD is the most mature display technology in display thus far. It would be imperative if LCD can be flexible. However, the high birefringence of the plastic substrate is an issue since LCD is polarization dependent. As shown in Fig. 9(a) (Oka *et al.*, 2016), phase retardation caused by the high birefringence is relevant to thickness of the plastic substrate. So several approaches have been

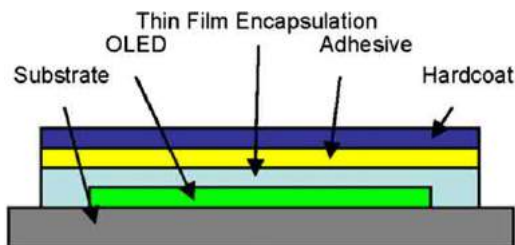


Fig. 8 A structure of thin film encapsulated OLED. Reproduced from Ma, R.Q., Rajan, K., Silvernail, J., *et al.*, 2010. Wearable 4-in. QVGA full-color-video flexible AMOLEDs for rugged applications. *Journal of the Society for Information Display* 18, 50–56.

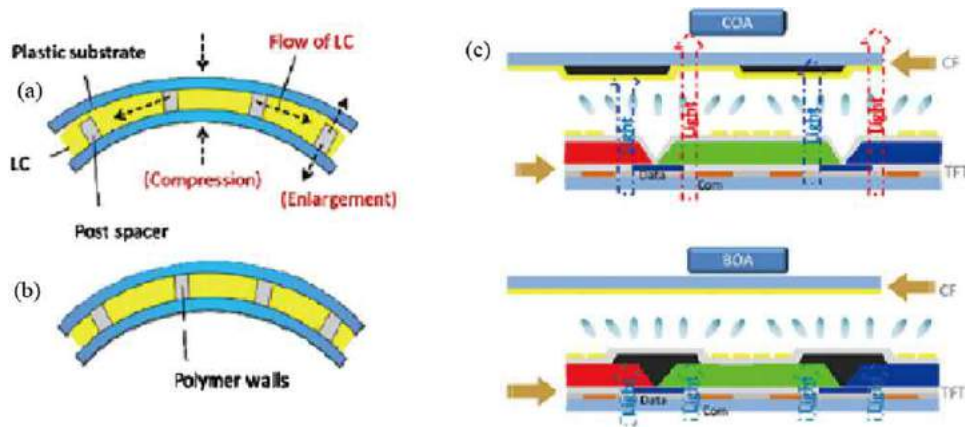


Fig. 9 (a) Deformation of plastic substrate during bending, (b) flexible LC device with polymer wall spacers, and (c) the principle of BOA to solve the BM shift issue in curved LCD panel. Reproduced from Ishinabe, T., Obonai, Y., Fujikake, H., 2016. A foldable ultra-thin LCD using a coat-debond polyimide substrate and polymer walls. *SID Digest 47* (1), 83–86; Ye, C.L., Fu, R.H., Qiu, J., *et al.*, 2016. The application of BOA on curved panel. *SID Digest 47* (1), 25–27.

taken: phase compensation film (Pin Hsiang *et al.*, 2016), in-cell polarizer (Fujiwara *et al.*, 2016), ultrathin substrate (Oka *et al.*, 2016; Pin Hsiang *et al.*, 2016) and low birefringence plastic substrate (Ishinabe *et al.*, 2016). Besides, thermal durability of plastic substrate is also important because of the high temperature during the LCD device fabrication process. Polyimide (PI) is widely applied as thermal durable plastic substrate, whose glass transition temperature ($>200^{\circ}\text{C}$) is compatible with LCD fabrication. However, PI needs further material design to be colorless since it is a little yellowish.

In addition to plastic substrate, curvature mismatching between different layers is another critical issue, which needs flexible spacers, black matrix alignment, etc. (as shown in Fig. 9(b)). Recently, Prof. T. Ishinabe from Tohoku University demonstrated with a so-called coat-debond method to fabricate the low birefringence polyimide (PI) and polymer wall spacer to make a flexible LCD demo, which has testified that flexible LCD is possible to be a candidate for future flexible display applications. Polymer walls used as the flexible spacer which could stick both substrate together are able to maintain the cell gap when compressing the device (Ishinabe *et al.*, 2016).

Flexible Transparent Conductive Film

The desired feature of conductors for wearable displays are: flexibility, stretchability, transparency, and conductivity. Conventional conductors like Indium Tin oxide (ITO) seem too fragile in flexible applications. Currently, the most popular substitute materials are graphene and its derivatives, metal nanowire, carbon nanotubes, and two-dimensional materials, which are all compatible with flexible substrate. However, carbon based materials synthesize still needs vacuum and even toxic process, which increases the cost. Two-dimensional materials are also challenging for large size preparation process. Metal nanowire such as silver nanowire forms conducting network. According to percolation theory, the ratio of conducting component which is non-transparent could be decreased to increase the transmittance with high conductivity, which is considered as one of the promising material.

Since metal nanowire is not transparent, it's essential to trade off the transmittance and conductivity. Several methods have been proposed including metal mesh, alignment coating, stacking structure, etc. (Yang *et al.*, 2016).

Patterning of flex transparent conducting film (Flex-TCF) is critical for device applications. The conventional way is lithography, but the etching recipe and conditions are totally different from ones in glass-based process. The plastic substrate is softer and less chemical resistive than glass, which means the Flex-TCF needs milder treatment to avoid negative effects to the substrate.

In order to be compatible with solution-based process and roll-to-roll fabrication, hydrophile/hydrophobic treatment is a more appropriate patterning method. Common hydrophile treatment includes UV/Ozone, plasma and chemical decoration. Defining the hydrophile/hydrophobic region first, then coat the dispersion of conducting material on the substrate and tends to assemble in the hydrophile region.

Practical Processes for Flexible Display

To make a mass-producible flexible device, the key processes of thin-film deposition, registration, patterning, surface treatment, and lamination should be developed. As Prof. Jun Souk analyzed, the ultimate approach is to use roll-to-roll (R2R) process (Souk and Roh, 2012) without vacuum ambience and photo-lithography. These patterning methods were mature in the newspaper or magazine printing industry.

See also: Displays – Introduction

References

- Fujiwara, D., Ishinabe, T., Koma, N., Fujikake, H., 2016. Thin flexible liquid crystal display using dye-type in-cell polarizer and pet substrates. *SID Digest* 47 (1), 18–22.
- Ishinabe, T., Obonai, Y., Fujikake, H., 2016. A foldable ultra-thin LCD using a coat-debond polyimide substrate and polymer walls. *SID Digest* 47 (1), 83–86.
- Lai, Y.H., Chan, C.C., Hwu, K.L., Huang, W.M., 2012. Photolithographic color filter for 14.1-inch flexible color electrophoretic displays. *SID Digest* 43 (1), 1365–1367.
- Lee, B., Byun, J., Oh, E., *et al.*, 2016. All-ink-jet-printed wearable information display directly fabricated onto an elastomeric substrate. *SID Digest* 47 (1), 672–675.
- Ma, R.Q., Rajan, K., Silvernail, J., *et al.*, 2010. Wearable 4-in. QVGA full-color-video flexible AMOLEDs for rugged applications. *Journal of the Society for Information Display* 18, 50–56.
- Oka, S., Sasaki, T., Tamaru, T., *et al.*, 2016. Optical compensation method for wide viewing angle IPS LCD using a plastic substrate. *SID Digest* 47 (1), 87–90.
- Pin Hsiang, C., Wen yuan, L., Zheng Han, C., *et al.*, 2016. Roll TFT-LCD with 20R curvature using optically compensated colorless-polyimide substrate. *SID Digest* 47 (1), 15–17.
- Souk, J.H., Roh, N.S., 2012. Processing flexible displays. *SID Digest* 40 (2), 622–624.
- Telfer, S.J., McCreary, M.D., 2016. A full-color electrophoretic display. *SID Digest* 43, 574–577.
- Yang, B.R., Liu, G.S., Han, S.J., *et al.*, 2016. Coating, patterning, and transferring processes of silver nanowire for flexible display and sensing applications. *Journal of the Society for Information Display* 24, 234–240.

Nonuniqueness in Imaging

Greg Gbur, University of North Carolina at Charlotte, Charlotte, NC, United States

© 2018 Elsevier Ltd. All rights reserved.

Glossary

Computed tomography Also known as computed axial tomography, a classic technique by which multiple X-ray absorption images of an object are taken and digitally reconstructed into a two- or three-dimensional image of the object's internal structure.

Condition number The ratio of the largest singular value of a matrix to its smallest. A large condition number (much greater than unity) implies that inversion techniques involving the matrix are highly susceptible to noise.

Ewald limiting sphere Sphere in Fourier space that characterizes the complete information that can be determined in a scattering experiment if data are taken for all directions of incidence and illumination.

Ewald sphere of reflection Sphere in Fourier space that characterizes the information that can be determined in a scattering experiment for a single direction of illumination.

Ill-conditioning Refers to a linear inversion problem with a high condition number, which is therefore highly unstable in the presence of noise.

Ill-posedness Refers to a linear inversion problem in which at least one of the following conditions is not satisfied: existence, uniqueness, and continuity.

Inverse problem A physical problem in which a "cause" is deduced from measurements of an "effect," in essence inverting the usual cause-effect type of problems usually studied in physics.

Inverse scattering problem Problem in which the structure of a scattering object is determined by measurements of the field scattered by the object.

Inverse source problem Problem in which the structure of a primary radiation source is determined by measurements of the radiated field. Known to be nonunique due to the existence of nonradiating sources.

Noncontinuity Refers to a linear inversion problem in which the reconstructed object is highly sensitive to even small changes in the image data.

Nonexistence Refers to a linear inversion problem in which the measured data may not correspond to any known object, due to the presence of noise.

Nonradiating source A primary radiation source which does not, in fact, produce any radiation.

Nonuniqueness Refers to a linear inversion problem in which multiple objects produce the same image.

Phase problem Imaging problem in which a source distribution is determined from measurements of the intensity of the image alone.

Prior knowledge Additional knowledge, usually dictated by the physics of an inverse problem, in order to make the solution unique.

Singular value decomposition Mathematical technique by which any matrix may be decomposed in such a way that the singular values (essentially diagonal elements of the decomposition) can be used to assess the stability of any inversion technique.

Overview

The techniques of optical imaging may be considered a subset of the broad class of so-called inverse problems, in which a general "cause" of a phenomenon is deduced from measurements of the "effect," namely, the phenomenon itself. This process of cause determination reverses the usual study of physics, in which the "effect" is deduced from knowledge of the "cause." In this context, the latter case is usually referred to as the forward problem. A familiar example of an inverse problem is computed tomography, the determination of the internal structure of a patient from measurements of the intensity of X-rays transmitted through the body. The cause is the human body itself, while the effect is the absorption of X-rays.

Because of this reversal of the familiar causal process of physics, inverse problems are often beset with difficulties. Chief among these is nonuniqueness, in which the measured data is insufficient, even in principle, to uniquely determine the properties of the object. In this article, we discuss the manifestation of nonuniqueness in a general imaging system, and the typical means by which it is overcome, if possible. Examples of nonuniqueness are given in the form of the inverse source problem and the phase retrieval problem.

Matrix Formulation of Imaging and Inverse Problems

We introduce the mathematics of nonuniqueness by considering an abstract imaging system, as illustrated in Fig. 1. We assume that the object data is characterized by the vector $|o\rangle$ and that the image data are characterized by the vector $|i\rangle$; the action of the imaging system is described by the matrix A . Assuming a linear relation between object and image, we may express the imaging process as

$$|i\rangle = A|o\rangle \quad (1)$$

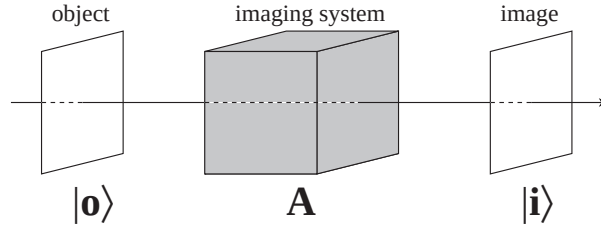


Fig. 1 Illustration of the notation for a general imaging system, in which an object vector $|o\rangle$ is converted to an image vector $|i\rangle$ through the action of an imaging system characterized by a matrix $\mathbf{A}$.

It is to be noted that we have drawn the geometry in [Fig. 1](#) as a traditional imaging system, in which a planar object is mapped to a planar image through the action of a “black box” imager, such as a system of lenses. However, the formalism to be used here applies to much more general systems. For example, in computed tomography, the object may represent a three-dimensional (3D) map of an object’s absorption coefficient, and the image data may represent X-ray projection data acquired through multiple measurements with different orientations of the tomographic device. In general, we are considering systems in which the total measured data of the image is $|i\rangle$, and the desired information about the object is represented by $|o\rangle$.

It is also to be noted that the physics of such an imaging problem is usually described by continuous functions and operators; that is, the object $o(\mathbf{r}')$ is mapped to an image $i(\mathbf{r})$ through an integration with an integral kernel $A(\mathbf{r}, \mathbf{r}')$ over the domain D of the object, such that

$$i(\mathbf{r}) = \int_D A(\mathbf{r}, \mathbf{r}') o(\mathbf{r}') d^2 r' \quad (2)$$

Since practical imaging problems inevitably require the discretization of such integrals, however, [Eq. \(1\)](#) is sufficient for our description.

In the presence of noise, we may modify [Eq. \(1\)](#) by adding a noise vector $|n\rangle$ to the image, with the result

$$|i\rangle = \mathbf{A}|o\rangle + |n\rangle \quad (3)$$

It is easiest to analyze the case for which $\mathbf{A}$ is a square $N \times N$ matrix, but more general $M \times N$ matrices may also be considered. When $M > N$, the image data collected is less than the amount of information available about the object, and it is expected that only a limited reconstruction of the object is possible. When $M < N$, more image data are collected than are in principle useful for reconstruction.

For a square Hermitian matrix, the properties of the imaging system may be characterized by the eigenvalues and eigenvectors of the matrix. However, as an imaging matrix is not necessarily Hermitian nor square, it is more appropriate to instead represent it using the singular value decomposition (SVD), given by

$$\mathbf{A} = \sum_{i=1}^r s_i |y_i\rangle \langle x_i| \quad (4)$$

where $r = \min(M, N)$, the vectors $|y_i\rangle$ are the eigenvectors of the Hermitian $m \times m$ matrix $\mathbf{Y} = \mathbf{A}\mathbf{A}^\dagger$, the vectors $|x_i\rangle$ are the eigenvectors of the Hermitian $n \times n$ matrix $\mathbf{X} = \mathbf{A}^\dagger \mathbf{A}$, and $s_i = (x_i)^{1/2}$, with x_i the eigenvalues of the matrix $\mathbf{X}$ (or $\mathbf{Y}$, if $N < M$). The vectors $|y_i\rangle$ are orthonormal, so that $\langle y_i | y_j \rangle = \delta_{ij}$, and also $\langle x_i | x_j \rangle = \delta_{ij}$. By [Eq. \(4\)](#), it can be readily seen then that

$$\mathbf{A}|x_i\rangle = s_i |y_i\rangle, \quad \mathbf{A}|y_i\rangle = s_i |x_i\rangle \quad (5)$$

The singular values are taken to be nonnegative by definition, which suggests that the signs of the eigenvectors must be taken appropriately. In matrix form, the SVD may be written as

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^\dagger \quad (6)$$

where $\mathbf{U}$ is an $m \times m$ unitary matrix made up of the eigenvectors $|y_i\rangle$ of $\mathbf{Y}$, $\mathbf{V}$ is an $n \times n$ unitary matrix made up of the eigenvectors $|x_i\rangle$ of $\mathbf{X}$, and $\mathbf{S}$ is an $m \times n$ “diagonal” matrix such that $S_{ij} = \delta_{ij} s_i$. More information about the SVD can be found in [Trefethen and Bau \(1997\)](#).

We may apply this decomposition to formally find a solution to the imaging equation ([Eq. \(1\)](#)). Assuming for the moment that all the singular values $s_i \neq 0$, a solution is given by

$$|o\rangle = \mathbf{V}\hat{\mathbf{S}}\mathbf{U}^\dagger |i\rangle \quad (7)$$

where $\hat{\mathbf{S}}$ is an $n \times m$ diagonal matrix with diagonal elements equal to $1/s_i$. The meaning of this solution is elaborated upon in the next section.

Nonuniqueness and Prior Knowledge

Let us now restrict ourselves to the ideal case in which $m=n$; the matrices $\mathbf{U}$, $\mathbf{V}$, and $\mathbf{S}$ are all square $m \times m$ matrices. The ability to solve our imaging problem then depends directly on the properties of the singular values s_i , and not on an overabundance or dearth of data.

An imaging problem is said to be well-posed if the solution exists, is unique, and is continuous. Any problem which fails to meet all of these conditions is said to be ill-posed, in a sense that was first introduced by Hadamard (1902). We next consider each of these conditions related to ill-posedness and how they are related, as well as basic strategies to ameliorate them.

The imaging system $\mathbf{A}$ may be viewed as a general mapping of the object data to the image data. The object data lies within an m -dimensional space called the object space, while the image data lies within another m -dimensional space called the image space. The set of all possible objects is referred to as the domain, while the set of all possible images is known as the range. The domain may be a subset of the object space, depending on how it is defined: for example, the object space might be considered as the set of all two variable functions, while the object domain might be defined as the set of all finite-valued two variable functions. The dimensionality of either the domain or range may be lower than that of the space in which it is embedded – or both. These regions, and the problems of nonuniqueness, noncontinuity, and nonexistence, are illustrated in Fig. 2.

Nonuniqueness is the case where two or more objects have exactly the same image. This arises when one or more of the singular values $s_i=0$. An object $|\mathbf{o}\rangle$ is said to lie in the null space of the imaging system if

$$\mathbf{A}|\mathbf{o}\rangle = 0 \quad (8)$$

Any object vector in the null space will contribute nothing to the image; therefore, if $\mathbf{o}_1$ represents an object, and $\mathbf{o}_2$ represents a null space vector, then $\mathbf{o}_1$ and $\mathbf{o}_1 + \mathbf{o}_2$ will have exactly the same image. Without additional information (to be discussed momentarily), there is no way to determine the contribution of the null space vector, or vectors, to the object. The values $s_i=0$ clearly make the inversion matrix $\hat{\mathbf{S}}$ singular, as it would contain terms of the form $1/s_i$.

As will be made explicit by example in the next section, the presence of nonuniqueness in an imaging problem may therefore be tied to the existence of invisible objects. An object which lies entirely in the null space will have no detectable image; it is invisible. Nonuniqueness in an imaging problem therefore implies the existence of invisible objects, and vice versa.

Nonexistence is closely related to nonuniqueness. The existence of null space vectors results in the range of the image having a smaller dimensionality than the domain of the object. There are, therefore, vectors in the image space that simply cannot be produced by an object. In a perfect imaging system, this would not be a problem; however, in the presence of additive noise $|\mathbf{n}\rangle$, it is possible for the image vector to be moved out of the range, in which case there exists no corresponding object.

Noncontinuity is a situation in which two very different objects produce very similar images. Strictly speaking, noncontinuity refers only to an integral imaging model, such as given by Eq. (2), in which the reconstructed object function $o(\mathbf{r})$ does not depend continuously on variations of the image function $i(\mathbf{r})$. In a discrete matrix model, the analogous problem is known as ill-conditioning, and arises when at least one of the singular values is extremely small compared to the others. To characterize this, we introduce the condition number κ , defined as the ratio of the largest singular value to the smallest,

$$\kappa \equiv \frac{s_{\max}}{s_{\min}} \quad (9)$$

A condition number which is not too much greater than unity (within two orders of magnitude) indicates a well-conditioned system, while an extremely large condition number represents ill-conditioning. Because the matrix $\hat{\mathbf{S}}$ depends on $1/s_i$, any vector corresponding to a very small singular value will be strongly amplified. In the presence of noise, this means that noise will also be highly amplified in an ill-conditioned system and will tend to drown out object information.

To resolve the issue of nonuniqueness, one must typically introduce prior knowledge based on known physics of the imaging system, as illustrated schematically in Fig. 2. Often, only a small class of mathematically achievable objects are physically reasonable, and restricting to this class will in some cases make the imaging problem unique; in other cases, additional prior knowledge may be needed. In phase retrieval problems, for instance, the size of the object is generally known and any

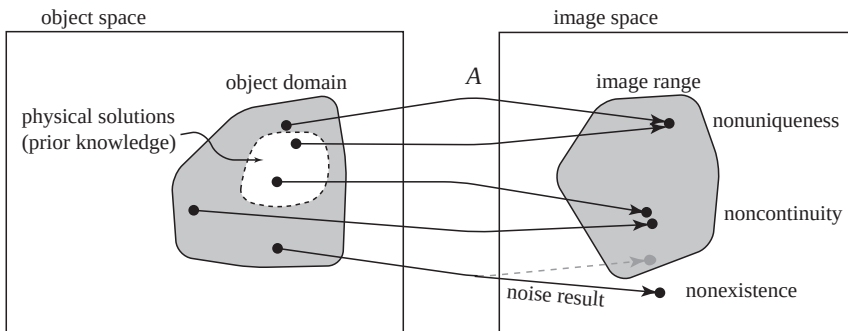


Fig. 2 Illustration of the different types of ill-posedness that can appear in an imaging system.

reconstruction which is much larger than this, or infinite, is rejected. Another commonly used condition used as prior knowledge is that the “physical” solution $|\mathbf{o}_i\rangle$ have minimum “energy,” i.e.,

$$\langle \mathbf{o}_i | \mathbf{o}_i \rangle \leq \langle \mathbf{o} | \mathbf{o} \rangle \text{ for any acceptable } \mathbf{o} \quad (10)$$

Prior knowledge needs to be chosen carefully based on the imaging problem being solved; it is possible to find a unique solution, which is nevertheless not useful or even closely related to the desired object.

In the context of the SVD, the best possible solution is often defined as the solution $|\hat{\mathbf{o}}\rangle$ which minimizes the quantity L , defined as

$$L \equiv |A|\hat{\mathbf{o}}\rangle - |\mathbf{i}\rangle| \quad (11)$$

It can be shown that this minimum solution is given by

$$|\mathbf{o}\rangle = V\hat{S}U^\dagger|\mathbf{i}\rangle \quad (12)$$

where $\hat{S}$ is the $n \times m$ diagonal matrix with diagonal elements equal to $1/s_i$ for $s_i \neq 0$ and 0 for $s_i = 0$; this is a generalization of Eq. (7). In essence, this solution calculates only those parts of the object that do not lie in the null space and ignores the rest. The solution can be further conditioned by setting the diagonal elements equal to zero for $s_i < S$, where S is a chosen critical value depending on practical properties of the system. In principle, throwing away those components of the solution which have small singular values results in a loss of information; however, as we have noted, those components would typically be corrupted by noise and overall be detrimental to the solution in any case.

The minimum solution of Eq. (12) in practice acts as an application of prior knowledge, though it may said to be philosophically different. The minimum solution is unique because it simply ignores the null space of the operator A ; ideally, prior knowledge is used to deduce the part of the solution which lies in the null space.

The Inverse Source Problem

One of the simplest, and perhaps most infamous, examples of nonuniqueness in an imaging problem is found in the inverse source problem. The geometry and related notation is illustrated in Fig. 3. We consider a 3D monochromatic primary source of light, confined to a domain D , defined by the function $q(\mathbf{r})$ and which produces a monochromatic field $u(\mathbf{r})$. The field and source are related by the 3D Helmholtz equation,

$$[\nabla^2 + k^2]u(\mathbf{r}) = -4\pi q(\mathbf{r}) \quad (13)$$

where $k = \omega/c$, ω is the frequency and c is the vacuum speed of light. We consider the scalar wave equation for simplicity, but similar arguments can be applied to a full electromagnetic radiation problem.

The solution for the field everywhere inside and outside the source domain is given by

$$u(\mathbf{r}) = \int_D \frac{e^{ikR}}{R} q(\mathbf{r}') d^3 r' \quad (14)$$

with $R = |\mathbf{r} - \mathbf{r}'|$. At a distance far away from the source, we may make the far zone approximation,

$$\frac{e^{ikR}}{R} \approx \frac{e^{ikr}}{r} e^{-iks \cdot \mathbf{r}'} \quad (15)$$

where $\mathbf{r} = r\mathbf{s}$, with $\mathbf{s}$ a unit vector pointing in the direction observation. In the far zone, then, the field may be written as

$$u(r\mathbf{s}) \approx \frac{e^{ikr}}{r} \int_D q(\mathbf{r}') e^{-iks \cdot \mathbf{r}'} d^3 r' \quad (16)$$

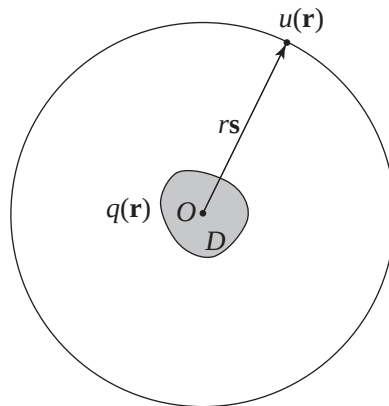


Fig. 3 Illustration of the notation and geometry of the inverse source problem.

The integral is now in the form of a 3D spatial Fourier transform, i.e.,

$$\tilde{q}(\mathbf{K}) = \frac{1}{(2\pi)^3} \int q(\mathbf{r}') e^{-i\mathbf{K} \cdot \mathbf{r}'} d^3 r' \quad (17)$$

In terms of this Fourier transform, we may write the far zone field as

$$u(r\mathbf{s}) \approx (2\pi)^3 \frac{e^{ikr}}{r} \tilde{q}(k\mathbf{s}) \quad (18)$$

It should be clear from this expression that any information about the source is contained in the Fourier transform, which depends only on the angular direction $\mathbf{s}$.

The inverse source problem may then be posed as an imaging problem as follows: given measurements of the field $u(r\mathbf{s})$ for all directions $\mathbf{s}$, determine the 3D structure $q(\mathbf{r})$ of the source. This problem has been shown to be severely nonunique due to the theoretical existence of invisible objects, known in this context as nonradiating sources. From Eq. (18), it can be seen that a nonradiating source is one for which

$$\tilde{q}(k\mathbf{s}) = 0 \text{ for all } \mathbf{s} \quad (19)$$

An example of such a nonradiating source is not difficult to come by (Kim and Wolf, 1986). Let us imagine a spherical source of radius a and uniform charge density q_0 , such that

$$q(\mathbf{r}) = \begin{cases} q_0, & |\mathbf{r}| \leq a \\ 0, & |\mathbf{r}| > a \end{cases} \quad (20)$$

The far zone field radiated by this source can be readily shown to be of the form,

$$u(r\mathbf{s}) = 2\pi \frac{e^{ikr}}{r} j_1(ka) \quad (21)$$

where $j_1(x)$ is the spherical Bessel function of order 1. The far zone field will be identically zero, i.e., the source will be nonradiating, if ka is equal to any of the zeros of $j_1(ka)$. The first zero of this Bessel function is at $ka = 4.49$, implying that the smallest nonradiating source of this structure has a radius $a = 4.49/k$.

Though radiationless sources have a long history in physics (Gbur, 2013), the influence of such sources on the inverse source problem was not clear until the 1980s, when a heated scientific argument clarified the matter (see Devaney and Sherman (1982) and the comments which follow).

It may be readily seen that the inverse source problem is what we might call fatally nonunique: the information present in the radiated field provides almost no information about the object structure. This can be seen by introducing a new nonradiating condition equivalent to Eq. (19). We multiply Eq. (16) by $\exp[i\mathbf{k}\mathbf{s} \cdot \mathbf{r}]$, and then integrate overall values of the unit vector $\mathbf{s}$; the result is a necessary and sufficient condition for a source to be nonradiating, of the form

$$\int_D q(\mathbf{r}') j_0(k|\mathbf{r} - \mathbf{r}'|) d^3 r' = 0 \text{ for all } \mathbf{r} \quad (22)$$

The spherical Bessel function $j_0(k|\mathbf{r} - \mathbf{r}'|)$ has a well-known modal expansion, namely,

$$j_0(k|\mathbf{r} - \mathbf{r}'|) = 4\pi \sum_{n=0}^{\infty} \sum_{m=-n}^n j_n(kr) j_n(kr') Y_n^{m*}(\theta, \phi) Y_n^m(\theta', \phi') \quad (23)$$

where $Y_n^m(\theta, \phi)$ are the spherical harmonics.

Let us now consider a general spherical source $q(\mathbf{r})$ of radius a . Integrated over such a spherical domain, the spherical harmonics are orthogonal. The condition given by Eq. (22) is therefore equivalent to the requirement that the nonradiating source $q(\mathbf{r}')$ be orthogonal to the infinite set of modes given by $j_n(kr') Y_n^m(\theta', \phi')$. However, for a fixed value of n and m , i.e., a fixed angular dependence, it is clear that there are an infinite number of orthogonal modes with the same angular dependence but different radial dependence (see examples given in Gbur and Wolf (1997)). It may be loosely stated that, for each mode that produces a measurable radiation pattern, there are an infinite number of modes with the same angular dependence that do not radiate.

Because of the fatal nonuniqueness of the inverse source problem, it is not generally possible to deduce anything significant about the source structure from the radiation pattern. The problem can be made unique by assuming (Porter and Devaney, 1982) that the source has minimum energy E , i.e., by minimizing the functional

$$E = \int_D |q(\mathbf{r}')|^2 d^3 r' \quad (24)$$

but this condition gives the same result as the SVD problem of Eq. (11), in that all the nonradiating source components are identically zero. This minimum energy condition does not seem to hold in realistic radiation problems.

Often, though not always, the uniqueness or nonuniqueness of an imaging problem can be determined by a simple dimensional analysis. From Eq. (18), it is to be noted that a measurement of the complete radiation pattern of a 3D source $q(\mathbf{r})$ determines the Fourier components $\tilde{q}(k\mathbf{s})$ on a spherical shell in 3D Fourier space. Because the function $\tilde{q}(\mathbf{K})$ is the Fourier transform of a function of finite extent, it follows that it is the boundary value of an analytic function in three complex variables,

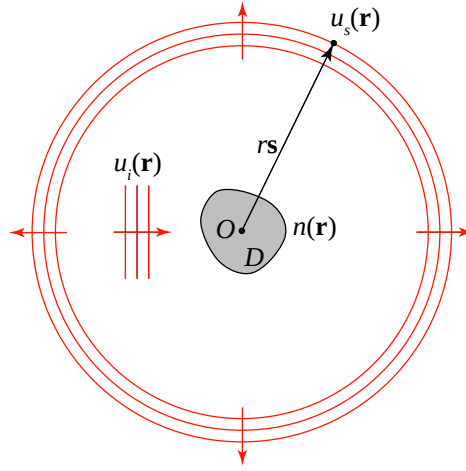


Fig. 4 Illustration of the notation and geometry of the inverse scattering problem.

K_x , K_y , and K_z . However, such a function can only be analytically continued if information is known in a 3D volume of space: the information on the two-dimensional (2D) shell is not sufficient to analytically continue the existing Fourier data to a volume.

Nonuniqueness can sometimes be the result of an insufficiently diverse dataset. This can be illustrated by the inverse scattering problem under weak scattering conditions, as illustrated in **Fig. 4**. An incident field $u_i(\mathbf{r})$ impinges on a localized scattering object of refractive index $n(\mathbf{r})$, producing a scattered field $u_s(\mathbf{r})$. The total field $u(\mathbf{r}) = u_i(\mathbf{r}) + u_s(\mathbf{r})$, and the incident field is defined as the field that would exist in the complete absence of the scatterer.

The total field satisfies the Helmholtz equation with inhomogeneous wave number,

$$[\nabla^2 + n^2(\mathbf{r})k^2]u(\mathbf{r}) = 0 \quad (25)$$

As is shown in [Born and Wolf \(1999, Chapter 13\)](#), this equation can be written in the form

$$[\nabla^2 + k^2]u_s(\mathbf{r}) = -4\pi F(\mathbf{r})u(\mathbf{r}) \quad (26)$$

where $F(\mathbf{r})$ is the scattering potential, given by

$$F(\mathbf{r}) = \frac{k^2}{4\pi} [n^2(\mathbf{r}) - 1] \quad (27)$$

Eq. (26) is formally the same as **Eq. (13)** for the source problem, and the far zone solution may be found in the same way to be

$$u_s(rs) \approx \frac{e^{ikr}}{r} \int_D F(\mathbf{r}') u(\mathbf{r}') e^{-iks \cdot \mathbf{r}'} d^3 r' \quad (28)$$

The difficulty with this solution is that the scattered field appears on both sides of the equation, implicitly in $u(\mathbf{r})$, making **Eq. (28)** an integral equation that cannot in general be solved. If one assumes that the scattered field is extremely weak, however, then $u(\mathbf{r}') \approx u_i(\mathbf{r}')$, and we have

$$u_s(rs) \approx \frac{e^{ikr}}{r} \int_D F(\mathbf{r}') u_i(\mathbf{r}') e^{-iks \cdot \mathbf{r}'} d^3 r' \quad (29)$$

Let us now further assume that the incident field is a plane wave propagating in a direction specified by the unit vector $\mathbf{s}_0$, of the form $u_i(\mathbf{r}) = \exp[ik\mathbf{s}_0 \cdot \mathbf{r}]$. Then the scattered field is of the form,

$$u_s(rs) \approx \frac{e^{ikr}}{r} \int_D F(\mathbf{r}') u_i(\mathbf{r}') e^{-ik(\mathbf{s}-\mathbf{s}_0) \cdot \mathbf{r}'} d^3 r' = (2\pi)^3 \frac{e^{ikr}}{r} \tilde{F}[k(\mathbf{s} - \mathbf{s}_0)] \quad (30)$$

The scattered field measured in a direction $\mathbf{s}$ therefore gives information about the Fourier components of the field at position $k(\mathbf{s} - \mathbf{s}_0)$; if measurements are made for all directions $\mathbf{s}$, one gets Fourier information on a sphere of radius k centered on $k\mathbf{s}_0$, known as an Ewald sphere of reflection. However, as in the inverse source case, the information on this Ewald sphere is not sufficient to reconstruct the scattering potential $F(\mathbf{r})$. If the scattered field is measured for all directions of incidence $\mathbf{s}_0$ and scattering $\mathbf{s}$, essentially over a continuum of Ewald spheres, one acquires information within a sphere of radius $2k$, known as the Ewald limiting sphere Σ_L . This geometry is illustrated in **Fig. 5**.

With the Fourier components determined within a 3D volume, it is possible to perform an inverse Fourier transform to determine a bandlimited reconstruction of the scattering object.

As the information on any finite number of Ewald spheres of reflection is insufficient to uniquely reconstruct the scattering object, one would expect that it is possible to make objects which are invisible, or nonscattering, for any finite number of directions of illumination, and this was in fact shown to be the case ([Devaney, 1978](#)). Evidently, as more data is taken, these nonscattering solutions play a correspondingly smaller role in the reconstruction process.

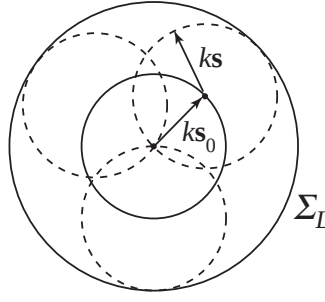


Fig. 5 Illustration of several Ewald spheres of reflection and the Ewald limiting sphere Σ_L .

Phase Retrieval

The discussion so far has been concerned with linear imaging systems, in which the image data has a linear relation to the object data. However, ill-posedness, in particular nonuniqueness, can also appear in nonlinear imaging, and the analysis of such systems is much less straightforward and must be dealt with on a case-by-case basis.

We consider here the so-called phase problem, or the phase retrieval problem, in which an object must be reconstructed from measurements of the image intensity alone. We assume that the object $f(x,y)$ and image $F(u,v)$ are related by a simple Fourier relation, i.e.,

$$F(u, v) = \int_S f(x, y) e^{-i2\pi(ux+vy)} dx dy \quad (31)$$

where S represents the object plane. The intensity of the image is given by $I(u, v) = |F(u, v)|^2$, and we are concerned with determining $f(x, y)$ from this data alone, if possible.

The nonlinear nature of this problem, due to the implicit reconstruction of the phase, results in some unexpected results. One of the most notable and famous is the observation that the problem is nonunique for one-dimensional (1D) systems but typically unique in 2D systems; we now examine how this is possible.

In considering the uniqueness of this imaging problem, we follow the discussion and notation of [Bruck and Sodin \(1979\)](#). The problem is almost certainly nonunique in general, and certain prior knowledge needs to be applied immediately. It is typically assumed that the source distribution is of finite extent and takes on only nonnegative and finite values.

The source is taken to consist of discrete points, represented by Dirac delta functions,

$$f(x, y) = \sum_{j=1}^m \sum_{k=1}^n f_{jk} \delta(x - x_j) \delta(y - y_k) \quad (32)$$

where $x_j = \Delta j$ and $y_k = \Delta k$, with Δ the spacing between points. The image data are then in the form of a discrete Fourier transform,

$$F(u, v) = \sum_{j,k} f_{jk} e^{-i2\pi\Delta(ju+kv)} \quad (33)$$

where for convenience we have used the Fourier transform in the form,

$$F(u, v) = \int f(x, y) e^{-i2\pi(ux+vy)} dx dy \quad (34)$$

The expression (33) may be simplified by writing the complex exponentials as complex variables, namely,

$$z \equiv e^{-i2\pi\Delta u}, \quad w \equiv e^{-i2\pi\Delta v} \quad (35)$$

where z and w lie on the unit circle in the complex plane. In this way, the image data may be represented as a polynomial in z and w ,

$$F(u, v) \equiv p(z, w) = \sum_{j,k} f_{jk} z^j w^k \quad (36)$$

The measured intensity of the image is given by the squared modulus of $p(z, w)$; since $|z| = |w| = 1$, $z^* = z^{-1}$ and $w^* = w^{-1}$, which means that we may write

$$p(z, w) p^*(z, w) = p(z, w) p(z^{-1}, w^{-1}) \quad (37)$$

From basic Fourier theory, we know that any spatial shift in the overall object position by $s\Delta$ along x and $t\Delta$ along y will result in the polynomial $p(z, w) \rightarrow p(z, w) z^s w^t$. The same argument applies to the measured intensity. We define the autocorrelation polynomial of the image as

$$Q(z, w) = p(z, w) p(z^{-1}, w^{-1}) z^m w^n \quad (38)$$

The additional powers $z^m w^n$ make this a true polynomial, unlike the product of Eq. (37), which will possess positive and negative powers of z and w . As noted, however, the additional powers only modify the object by a trivial spatial shift.

Let us now consider the 1D phase problem, in which the object is $f(x)$ and the image polynomial is of the form

$$p(z) = \sum_{j=0}^m f_j z^j \quad (39)$$

with $f_j \geq 0$. By the fundamental theorem of algebra, any such polynomial may be factored into a product of m terms,

$$p(z) = \prod_{j=1}^m (z - z_j) \quad (40)$$

With the constraint that the $f_j \geq 0$, the zeros z_j of this polynomial must either satisfy $\text{Re}\{z_j\} \leq 0$ and $\text{Im}\{z_j\} = 0$, or must come in a complex conjugate pair, i.e., if z_j is a zero, then so is z_j^* . This must be the case because the coefficients f_j will feature every combination of products of the z_j 's.

With this in mind, the autocorrelation function may be written as

$$Q(z) = \prod_{j=1}^m (z - z_j)(z - z_j^{-1}) \quad (41)$$

and we note that the phase problem may be stated as determining the coefficients f_j from the values of this autocorrelation function, or equivalently determining the z_j 's. The function $Q(z)$ is found by direct measurement, but it features both z_j and z_j^{-1} as its zeros. If all $|z_j| = 1$, then both z_j and z_j^* must be zeros of the polynomial $p(z)$; the solution is uniquely determined. If all $|z_j| \neq 1$, then it is not possible from the autocorrelation function to determine if a measured z_j or its inverse is the true zero of the polynomial $p(z)$. Since there are two choices for each of the zeros, there are 2^m possible solutions for $p(z)$ for the given $Q(z)$ —the solution is nonunique. If there are r zeros with $|z_j| = 1$, then there will be 2^{m-r} possible solutions, and the solution is still nonunique. In general it can be seen that the phase problem is nonunique in 1D.

For the 2D case, however, the situation is quite different. This arises from the fact that the polynomial $p(z, w)$ cannot be represented in a factorized form,

$$p(z, w) \neq \prod_{j,k} (z - z_j)(w - w_k) \quad (42)$$

If it could, we could make arguments similar to the 1D case to prove nonuniqueness. It is very easy, though, to find examples which do not satisfy Eq. (42). For instance, the polynomial

$$p(z, w) = z^2 + w^3 \quad (43)$$

cannot be written in the form of Eq. (42). Without factorization, the earlier nonuniqueness argument breaks down. This alone does not demonstrate that the problem becomes unique, but experimental and computational studies have shown that it is generally so. Of course, in the rare case that the image polynomials can be factorized, uniqueness will fail.

Reconstruction of the phase can be done in many cases by the venerable algorithm of [Gerchberg and Saxton \(1972\)](#), though this method also runs into issues of nonuniqueness when the data possesses certain symmetries.

References

- Born, M., Wolf, E., 1999. Principles of Optics, seventh ed. Cambridge: Cambridge University Press.
- Bruck, Y., Sodin, L., 1979. On the ambiguity of the image reconstruction problem. *Optics Communications* 30, 304–308.
- Devaney, A., 1978. Nonuniqueness in the inverse scattering problem. *Journal of Mathematical Physics* 19, 1526–1531.
- Devaney, A., Sherman, G., 1982. Nonuniqueness in inverse source and scattering problems. *IEEE Transactions on Antennas and Propagation* 30, 1034–1037.
- Gbur, G., 2013. Invisibility physics: Past, present, and future. In: Wolf, E. (Ed.), *Progress in Optics*, vol. 58. Amsterdam: Elsevier, pp. 65–114.
- Gbur, G., Wolf, E., 1997. Sources of arbitrary states of coherence that generate completely coherent fields outside the source. *Optics Letters* 22, 943–945.
- Gerchberg, R., Saxton, W., 1972. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik* 35, 1–6.
- Hadamard, J., 1902. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin* 13, 49.
- Kim, K., Wolf, E., 1986. Non-radiating monochromatic sources and their fields. *Optics Communication* 59, 1–6.
- Porter, R., Devaney, A., 1982. Holography and the inverse source problem. *Journal of the Optical Society of America* 72, 327–330.
- Trefethen, L., Bau, D., 1997. *Numerical Linear Algebra*. Philadelphia, PA: SIAM.

View Angles of Liquid Crystal Displays

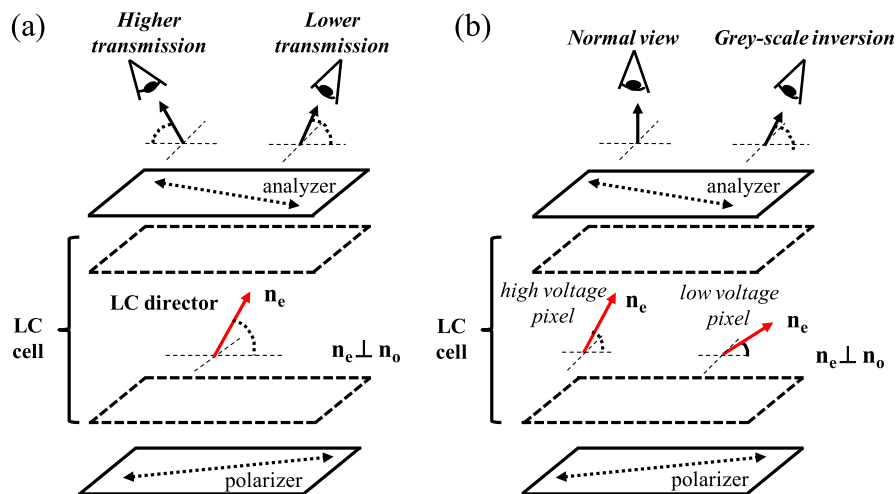
Huang-Ming Philip Chen, National Chiao Tung University, Hsinchu, Taiwan

© 2018 Elsevier Ltd. All rights reserved.

The viewing angle of liquid crystal displays is defined as the maximum angle at which a display panel can be viewed within acceptable visual performance when a contrast ratio exceeds the minimum ratio of 10:1. The viewing angle is measured from one position to the opposite direction, which gives a maximum of 180° in one axis. Four viewing angles need to be taken into account: the horizontal (0°), vertical (90°), and $\pm 45^\circ$ axes. The detail measurements can be performed either by hand or by a conoscopy at smaller increment degrees. The contour data mapping exhibits the viewing angle and viewing cone in displays. The narrow viewing angle of twist nematic liquid crystal displays (TN LCDs) is an intrinsic issue of LCDs due to the various retardation changes that are associated with LC molecular motion under electric field, also known as birefringent mode. The contrast ratio is reduced by the light leakage at oblique angle caused by elliptical polarization states. Two major approaches have been applied side by side to solve the issue. One approach is working on optical compensation films to compensate polarization loss in LC cells (Mori *et al.*, 1997). The other is dedicating in LC driving modes and pixel designs to achieve spatial effect in different viewing angles, such as In-Plane-Switching (IPS), or Multi-Domain Vertical Alignment (MVA) modes. Due to the manufacture limitation, both approaches need to work closely to achieve desired wide viewing angle LCDs. In addition to the enhancement of viewing angle in LCDs, many new optical designing concepts have been proposed and realized in the latest researches beyond the aforementioned framework to control the viewing angles that fit the specific needs of display applications.

Optical Compensation Film

The key of designing optical compensation film is based on the LC device in a particular driven state, which is associated with the birefringent mode in LC optics. When the linear polarized light passes through the LC medium, it also loses its linear polarization state and turns into elliptically polarized light. The long axis of elliptically polarized light does not closely follow the local LC director. The light leakage occurs when elliptically polarized light passes through the second polarizer (or called analyzer) at oblique angle. The polarization states of oblique angle can be calculated and mapped on the surface of a Poincaré Sphere. The calculation of correction leading to the linear polarization point can be applied as an optical compensation (Zhu *et al.*, 2006). The directors' profile is taken into consideration when designing the optical compensation film accordingly. It is important to note that this compensation method cannot resolve all viewing angle issues, especially in the gray-scale inversion, which requires divided pixels in domains with different orientations. A pixel divided into two domains may provide a significant reduction in the gray-scale inversion. A pixel that splits into four domains can give relatively symmetrical appearances to improve both gray level and contrast in a panel.



An optical compensation film (or so called Wide View film) is characterized by its refractive indices n_x , n_y , and n_z , in which n_z is the refractive index in the direction perpendicular to the film; n_x and n_y are the in-plane at X and Y axis. In order to compensate an on-state of TN LCD, an axial symmetric compensator with negative birefringence ($n_x = n_y > n_z$) can be applied. In general, the compensation films can be classified by its optical characteristic, such as A, C, O-plate, and biaxial films. Based on the specific manufacture processes, these films can also be described, as stretching films, or LC coated films.

	<i>A plate</i>	<i>C plate</i>
Positive (+)	$n_x > n_y = n_z$	$n_z > n_x = n_y$
Negative (-)	$n_x = n_z > n_y$	$n_x = n_y > n_z$

Many references have indicated the viewing angle issues can be solved by various optical compensation films for different LCD modes. In general, the viewing angle of TN LCD mode can be compensated by bend-splay hybrid orientation discotic material coated films. The viewing angle of MVA mode can be compensated by either biaxial film itself or positive A (A^+) plate in conjunction with negative C (C^-) plate. The IPS mode, on the other hand, requires either negative A (A^-) plate or positive C (C^+) plate to enhance its viewing angle. The complex structures of compensation films (i.e., multiple optical composite layers) were proposed in some reports to solve the finite polarization deviation in the LC cells. These calculated solutions may be unrealistic when facing limited manufacturing ability or rising costs. The Fuji film Corporation (Ito *et al.*, 2013) has revealed their significant Wide View film (WV film) development in 2013 (Ito *et al.*, 2013). The hybrid alignment of polymerized discotic materials is the main component of compensation film for TN LCDs. The grayscale inversion can be significantly minimized because the LCs' director in the middle of the cell is parallel to the axis of the polarizer. The retardation of the LC has least transmittance effect when the darker gray level moving to the brighter gray level. As a result, controlling the orientation of polymerized discotic materials is the key to formulate new WV films in improving the TN-LCDs' gray scale image qualities. In addition, they also developed a precise patterning chemical method by using one shot of non-polarized UV light to control and create patterned discotic alignments. The resulting film is called film patterned retarder (FPR) with out-of-plane retardation (R_{th}) close to 0 nm. The advantage of FRP is to reduce the oblique cross-talk in 3D display applications. Furthermore, the achromatic quarter wave films can be prepared by coating the rod-like LC layer on top of the polymerized discotic materials. Their achromatic quarter wave film measurement outcome suggests that the retardation is similar to the ideal curve below 550 nm. The deviation is much smaller than one-film type achromatic quarter wave film in the range of 550–650 nm. This new achromatic quarter wave films are great assets to LCD as well as to OLED for the true black appearance.

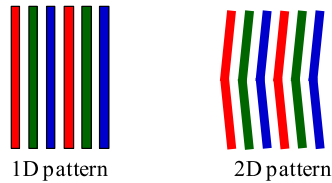
The optical isotropic Blue Phase LCD is assumed to have a good dark state and high contrast ratio in theory. The undesired light leakage occurs in the voltage off-state. Wu *et al.* explains that the causes of the dark-state light leakage are coming from the edge of protrusive electrodes, polarization rotation effect of BPLC material, and scattering from polymer network (Liu *et al.*, 2014). In this article, it states that the major source of light leakage comes from the polarization rotation effect in the dark state. Fortunately, this light leakage can be compensated by the Poincaré Sphere method. There are three major solutions using compensation films to suppress light leakage for the polarization rotation effect: (1) Small angle rotation of analyzer with biaxial film compensates the light leakage. This method can only reach the contrast ratio as high as 3700:1 in the normal direction. This simple approach faces more challenges when one wishes to further improve the contrast ratio through compensation. (2) Utilizing broadband and wide view circular polarizers to compensate the light leakage is proposed by Hong. The contrast ratio at normal direction can be as high as 12,000:1; and the viewing angle can be as wide as the entire viewing zone (CR 100:1). This approach requires that the two layers of biaxial films to form one circular polarizers for one side. Consequently, there are four layers of biaxial films that need to be aligned in order to obtain good results. The complicated film alignments may hinder one to apply this method for compensating the light leakage. (3) Using dispersive positive A (A^+) film is another method to compensate the light leakage. The birefringence of dispersive positive A film is designed to vary from 450 to 650 nm, for example: 0.008 for 450 nm, 0.005 for 550 nm, and 0.004 for 650 nm. The last method which requires not only dispersive positive A film, but also biaxial films is the same method as (1) and (2). The result of BPLCD can reach contrast ratio at normal direction as high as 13,500:1 with relatively wide viewing zone.

The flexible LCDs research becomes much more active recently in competing with the flexible OLED devices. The capability of flexible LCDs remains challenged. Many research groups have turned their attentions toward the enhancement of the optical performance that may exclusively occur in the flexible LCD. Ishinabe *et al.* (2014) published the article indicating that the contrast ratio of 650:1 and a wide viewing angle of 150° based on polycarbonate substrates in 2014. Jeong *et al.* (2016) attempted to eliminate the bending-induced retardation of a plastic film by coating polystyrene (PS) or poly (methyl methacrylate) (PMMA) on the polycarbonate substrates. The negative birefringence of the coated PS or PMMA layer is formed when bending the polycarbonate substrates. As a result, the bending-induced retardation from polycarbonate substrate can be compensated. Oka *et al.* (2016) (Japan Display) revealed their design using the polyimide (PI) and acrylic resin layers as C^+ -plate to compensate their flexible FFS LCD in 2016. The retardation can be controlled by the thickness of PI layer. The positive C (C^+)-plate ($R_{th} = -450$ nm) can be reached when PI is at 3.4 μm in thickness. A 2.95-in. prototype sheet LCD with contrast ratio of 860:1 in 400 μm panel thickness was demonstrated.

Improvement From LC Modes

Various nematic LC modes have been proposed and realized for the wide view liquid crystal displays. Hitachi manufactured the first super-TFT liquid crystal display (was later called In-Plane-Switching (IPS) mode) to enhance the viewing angle back in 1996. It was evolved from IPS mode to super IPS (S-IPS) mode to solve the color shifting issue in 1998, followed by the advance super IPS (AS-IPS) to enhance the transmittance in 2002. In order to solve the aspect ratio, transmittance, and contrast issues, the concept of Fringe Field Switching (FFS) was proposed by Lee in 1998 and was put into mass production of the IPS-Pro LCD TV

(as the abbreviation of IPS-Provectus) at the end of 2004. The FFS mode has over 3 times of higher contrast and 1.5 times of higher transmission than original IPS mode. Besides the electrode design, the optical compensation films still require for IPS or FFS modes to further enhance their viewing angle at $\pm 45^\circ$ and $\pm 135^\circ$. Due to the high demand of portable LC devices, p-FFS and n-FFS were proposed by using positive (p) and negative (n) dielectric LC materials. The pros and cons between p-FFS and n-FFS were mentioned by Wu (Chen *et al.*, 2015). In short, p-FFS gains the response time, yet n-FFS achieves the higher transmittance. Both have excellent viewing angle and minor color shift. The n-FFS has smaller image flicker, but it needs a 2D electrode patterned design to suppress its grayscale inversion.



Multi-Domain Vertical Alignment (MVA) liquid crystal display using negative dielectric LC to achieve wide view display was first developed by Fujitsu in 1997. The protrusion electrode divided one pixel into 4 domains. Single sided protrusion and the open slit on the counter side MVA structure was announced to achieve similar effect in 1999. AU Optronics revealed their AMVA with fast pixel response times yet small brightness changes. Their second generation of AMVA (AMVA II) was presented with the improved contrast ratio to 2500:1 in 2006. In 2009, the third generation Advanced MVA III (AMVA III) was developed to fulfill the high quality image and fast response time. The AMVA III adopted the new cell process based on polymer sustained alignment (PSA) process and new pixel electrode pattern design without protrusion and ITO main slit. Moreover, 8-domain pixel electrode pattern with new ART (additional refreshing transistor) architecture was implemented to gain the precise control in the lower levels of gray scale. As a result, a 46-in. AMVA III LCD can obtain contrast ratio greater than 8000:1 with 180° viewing angle (Chen *et al.*, 2009). It is notable that Samsung Electronics Co. Ltd. received a Gold Award for their 40-in. High Contrast, Wide Color Gamut, LED-Backlit LCD TV at the 45th annual Society for Display (SID) 2007 exhibition. This prototype LED TV featured a dynamic contrast ratio of 10,000:1 and claimed its viewing angle at 180° in all direction. MVA has superior viewing angle, good dark state, color reproduction, and fast response time. Comparing to IPS mode, the optical compensation film of MVA mode is much easier to reach its high contrast ratio and wide viewing angles. As a result, MVA mode dominates the TV market lately.

Another multi-domain VA LCD approach is called Axially Symmetric Vertical Alignment (ASVA) mode with circularly symmetric at 360° viewing angle. It was first proposed as the Axis Symmetric Micro-Cell (ASM) method by Yamada *et al.* (1995). Axis symmetry micro-cell (ASM) is fabricated by phase separation of liquid crystal and polymer wall, in which the LC molecules are surrounded by polymer walls under UV irradiation. The chiral compound is added to ensure nematic liquid crystal with a 90° twisted orientation. The ASM method has the advantage not only wide viewing angle due to the symmetry orientation axis, but also disclination-free due to the mono-domain orientation. They continue to develop a negative dielectric liquid crystal in the ASM (N-type ASM) mode in 1998. The N-type ASM mode combined with compensation film is able to reach 300:1 contrast ratio and over $\pm 80^\circ$ viewing angle. The difference between P-type and N-type is that axially symmetrical direction and tilted angle are set prior to carrying out the final curing of UV irradiation. The only drawback is the response time falling around 30 ms. They proposed the ASVA mode with multiple protrusion structures to define the plurality of liquid crystal regions in 2001. Liquid crystal molecules in the plurality of liquid crystal regions are aligned axially and symmetrically around an axis perpendicular to the surface. Each region is approximately a square of 70 μm .

In addition to IPS and MVA modes, researchers have continually improved the image quality by all means. Optically compensated bend (OCB) mode is known to possess the advantages of wide viewing angles and fast response time over nematic liquid crystal displays. It was modified from the original pi-cell (proposed by Bos and Koehler/Beran, 1984) to a much thinner cell gap with optical compensation films by Yamaguchi *et al.* (1993). Bos *et al.* proposed using a negative birefringence film with the optic axis in the normal direction (negative C-plate) to compensate the pi cell in 1993. Yamaguchi *et al.* (1993) suggested the use of a biaxial film for optical compensation of the OCB cell. The viewing angle performance, however, was not satisfied by these compensation methods. Mori offered the OCB-WV film consists of a hybrid polymerized discotic material layer alignment (as in TN WV film) with biaxial film to suppress the light leakage in the on-state of bend mode (Mori, 2005). Chen *et al.* proposed using reactive monomer to suppress the splay-to-bend transition (i.e., bias-free OCB mode) in 2009. The thin splay orientation of LC polymer layer not only provides the surface pre-tilted angle to suppress the splay-to-bend transition, but simultaneously becomes an inner-retarder to minimize the light leakage. The contrast ratio is improved up to a factor of 11 without using compensation films (Chen *et al.*, 2009).

The Latest Approaches of Wide View Angle LCDs

The latest approaches are focused in the optical structure design to enhance LCD's viewing angle after 2010. The design implements a diffuse optics on top of LCD. Kälantär reported the concept of shaping the luminance cone and making spatially uniform

luminance by a light scattering film in SID Symposium 2011. A directional backlight with a narrow angular luminance distribution was applied in the backlight unit. The result suggested that the normal and oblique contrasts were enhanced by a factor of 1.1 and 6 on the VA panel; and 2.5 and 2.7 on the TN panel, respectively. The imaging blur and low intensity were the main drawback of using conventional diffusing film. Yamamoto *et al.* (2014) recommended microstructure film to control passing light on top of LCD in 2014. The microstructure film consists of plate-like transparent polymer and the black dots were randomly arranged on the base film. The cone shape air cavities were placed beneath the black dots to control the passing light. Their result suggested that the microstructure film has over five times higher contrast ratio than the conventional diffusing film. In a similar way, Yang and Chen proposed photo-luminescent LCD with R, G, B-quantum dots (QDs) on top of the conventional TN LC cell to improve the LCD viewing angle in 2016. In particular, the single wavelength excitation at 410 nm gives the great advantage for LCDs without considering the retardation dispersion. Thus, the higher contrast ratio at normally black mode achieved $\pm 70^\circ$ viewing angle (CR > 10:1) and 3 ms under 4 V driving voltage in the QD array over TN-LC (Yang *et al.*, 2016).

Controlled Viewing Angle of LCDs

The criticism of the viewing angle has been raged over the years when TN LCD first appeared in the market. The issues of viewing angle are mostly attributed to the unique birefringent mode controlling the light intensity between crossed-polarizers. Wide viewing angle (WVA) LCDs dominate the display market. However, the narrow viewing angle (NVA) or dual-view LCD is getting its popularity for security, personal privacy, or safety reasons. The birefringent mode is more prominent for designing the directional viewing control system. A number of approaches have been realized in the controllable viewing angle LCDs. The double-cell architecture is the first proposed method, in which one additional liquid crystal panel is placed on top of the main panel to control the viewing angle (Jeong *et al.*, 2007). It is the simplest approach, but suffers from the panel thickness and greater costs. The second design is subpixel divided method, in which each pixel is divided into two domains: the primary pixel domain is for displaying images whereas the sub-pixel domain is to control the viewing angle (Lim *et al.*, 2010). The third method is to adopt dual light sources in conjunction with two light-guided films in the backlight system without changing the design of LC panel. The resulting controllable LCD is able to obtain the viewing angular ranges from 140° to 60° (Chien *et al.*, 2006). The fourth is based on a HAN (hybrid aligned nematic) cell, driven by fringe- and vertical-electric fields (Jeong *et al.*, 2008). The viewing angle in the horizontal direction can be controlled from 120° to about 20° . The last method, but not the least, is utilizing a patterned vertical alignment LCD (PVA-LCD) with the circular polarizers. The WVA mode is turned on when no bias voltage on the common electrode. A bias voltage (2.42 V) is applied on common electrode to turn on the NVA mode. The contrast ratio of WVA mode is more than 100:1 in all viewing angles. The viewing angle can be reduced to 20° when the NVA mode is turned on (Yu *et al.*, 2014).

Dual-view liquid crystal display is able to simultaneously deliver different information in the right and left viewing directions, i.e., two distinct viewing cones for right and left directions. After SHARP launched "Dual View LCD" technology in 2005, the dual view display technology was soon adopted by the automobile makers, in their concept cars. There are several approaches have been realized in the dual-view LCD. One of the approaches is to create spatial-multiplexing type, in which the main pixel of the DV LCD contains both a right and left sub-pixel. Parallax barriers and lenticular lenses are commonly applied to direct light from different pixels to achieve the intended field of view. Controlling the viewing angles of the right sub-pixel (RSP) and the left sub-pixel (LSP) can be fulfilled by two different tilted surfaces in one pixel, or by patterned electrodes, in which it provides the inclined electric field to control the LC direction. The other approach is based on the time-multiplexed type, which uses the directional backlight module with complex reflectors and fast response LC cell, such as OCB, Blue Phase modes. The negative aspect is that the aperture ratio is decreased as a result of the dead zones formed nearby the electrodes. In 2017, Chen *et al.* proposed a two-domain twisted nematic liquid crystal with a patterned E-type polarizer which is capable of suppressing the crosstalk between right and left views. The crosstalk is successfully lowered to less than 0.07% for all gray levels, while keeping the viewing angle around 50° for both left and right views (Chen *et al.*, 2017).

Summary

Viewing angle is an intrinsic matter of LCD associated with its birefringent mode optics. It has been intensively studied to achieve wide viewing angle for monitor and TV applications. The methods such as optical compensation films, pixel designs, and driving modes are available to resolve the negative viewing angle issues. Since Samsung succeeded in delivering the widest viewing angle in 2007, the viewing angle is no longer a critical issue in LCDs. It is possible to extend the viewing angle in all directions based on fulfilling the minimum contrast ratio 10:1 requirement. On the contrary, the small viewing angle of LCDs becomes a benefit when one needs to control and manipulate as needed, especially, for privacy purposes. Thus, the latest developments of alternating maximum and minimum viewing angles are valuable traits in LCDs.

See also: Displays – Introduction. Liquid Crystal Physics and Materials

References

- Bos, P.J., Koehler/Beran, K.R., 1984. The pi-cell: A fast liquid-crystal optical-switching device. *Molecular Crystals and Liquid Crystals* 113, 329–339.
- Chen, C.P., Wu, Y., Zhou, L., *et al.*, 2017. Crosstalk-free dual-view liquid crystal display using patterned E-type polarizer. *Applied Optics* 56, 380–384.
- Chen, H., Gao, Y., Wu, S.-T., 2015. n-FFS vs. p-FFS: Who wins? *SID Symposium Digest* 46, 735–738.
- Chen, S.-F.F., Chen, H.-M.P., Chow, L., Chang, Y.-Y., Shieh, H.-P.D., 2009. Splay-to-bend transition-free reactive monomer modified Pi-cell. *IEEE Photonics Technology Letters* 21, 636–638.
- Chen, T.-S., Huang, C.-W., Hsieh, C.-C., *et al.*, 2009. Advanced MVA III technology for high-quality LCD TVs. *SID Symposium Digest* 41, 776–779.
- Chien, K.-W., Hsu, Y.-J., Chen, H.-M., 2006. Dual light source for backlight systems for smart viewing-adjustable LCDs. *SID Symposium Digest* 37, 1425–1427.
- Ishinabe, T., Sato, A., Fujikake, H., 2014. Wide-viewing-angle flexible liquid crystal displays with optical compensation of polycarbonate substrates. *Applied Physics Express* 7, 111701.
- Ito, Y., Watanabe, J., Saitoh, Y., *et al.*, 2013. Innovation of optical films using polymerized discotic materials: Past, present and future. *SID Symposium Digest* 44, 526–529.
- Jeong, E., Lim, Y.J., Chin, M.H., *et al.*, 2008. Viewing-angle controllable liquid crystal display using a fringe- and vertical-field driven hybrid aligned nematic liquid crystal. *Applied Physics Letters* 92, 261102.
- Jeong, E., Lim, Y.J., Rhee, J.M., Lee, S.H., 2007. Viewing angle switching of vertical alignment liquid crystal displays by controlling birefringence of homogeneously aligned liquid crystal layer. *Applied Physics Letters* 90, 051116.
- Jeong, J., Park, D., Lee, J.-H., 2016. Simultaneous retardation compensation during bending of plastic film coated with a polymer layer of opposite birefringence. *Applied Optics* 55, 8738–8742.
- Lim, Y.J., Kim, J.H., Her, J.H., *et al.*, 2010. Viewing angle controllable liquid crystal display with high transmittance. *Optics Express* 18, 6824–6830.
- Liu, Y., Lan, Y.-f., Hong, Q., Wu, S.-T., 2014. Compensation film designs for high contrast wide-view blue phase liquid crystal displays. *Journal of Display Technology* 10, 3–4.
- Mori, H., 2005. The wide view (WV) film for enhancing the field of view of LCDs. *Journal of Display Technology* 1, 179–186.
- Mori, H., Itoh, Y., Nishiura, Y., Nakamura, T., Shinagawa, Y., 1997. Performance of a novel optical compensation film based on negative birefringence of discotic compound for wide-viewing-angle twisted-nematic liquid-crystal displays. *Japanese Journal of Applied Physics* 36, 143–147.
- Oka, S., Sasaki, T., Tamaru, T., *et al.*, 2016. Optical compensation method for wide viewing angle IPS LCD using a plastic substrate. *SID Symposium Digest* 47, 87–90.
- Yamada, N., Kohzaki, S., Funada, F., Awane, K., 1995. Axially symmetric aligned microcell (ASM) mode: Electro-optic characteristics of new display mode with excellent wide viewing angle. *SID Symposium Digest* 26, 575–578.
- Yamaguchi, Y., Miyashita, T., Uchida, T., 1993. Wide-viewing-angle display mode for the active-matrix LCD using bend-alignment liquid-crystal cell. *SID Symposium Digest* 24, 277–280.
- Yamamoto, E., Yui, H., Katsuta, S., *et al.*, 2014. Wide viewing LCDs using novel microstructure film. *SID Symposium Digest* 45, 385–388.
- Yang, J.-P., Hsiang, E.-L., Chen, H.-M.P., 2016. Wide viewing angle TN LCD enhanced by printed quantum-dots film. *SID Symposium Digest* 47, 21–24.
- Yu, Y., Dou, H., Ma, H., Sun, Y., 2014. Continuous viewing angle controllable patterned vertical alignment liquid crystal display. *Liquid Crystals* 41, 1595–1599.
- Zhu, X., Ge, Z., Wu, S.T., 2006. Analytical solutions for ni axial-film-compensated wide-view liquid crystal displays. *Journal of Display Technology* 2, 2–20.

Organic Light-Emitting Diode (OLED)

Chin H (Fred) Chen and Wen-Shi Wen, Shine Materials Technology Co., Ltd., Kaohsiung, Taiwan, R.O.C.
Chin-Ti Chen, Chemistry Institute, Academia Sinica, Taipei, Taiwan, R.O.C.

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Organic light-emitting diode (OLED) with its superb attributes in thinness, light-weight, brightness, color, dynamic contrast, viewing angle, response time, power efficiency and form factor has been declared as the “ultimate display” of our time. It is an emissive device with a very simple structure which comprises a stack of extremely thin layers ($<0.2\ \mu\text{m}$) of organic materials sandwiched between two electrodes. The anode which is usually made of a transparent semi-conductor of high work function (e.g., ITO) is capable of injecting holes into the organics while the cathode often selected from a highly reflective metal of low work function (e.g., Li, Ca, Mg, LiF/Al) is capable of injecting electrons. If the corresponding adjacent organic layers are chosen properly with a set of matching highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) and applied a positive bias voltage (DC driven), the injected electron and hole are transported to recombine near the heterojunction of the corresponding electron/hole transporting layers (ETL/HTL) to produce exciton, from which light emission will occur and emit through the ITO/glass. The color of the emitted photon will roughly correspond to the bandgap energy of the organic luminescent material with an internal quantum efficiency (η_{INT}) governed by the equation of ($\eta_{\text{INT}} = \gamma \cdot \eta_r \cdot \psi_{\text{pl}}$), where γ is the probability of carrier recombination (e/h ratio), η_r is the efficiency of exciton production and ψ_{pl} is the photoluminescence quantum efficiency. Electro-generated excitons statistically have a ratio of singlet/triplet ratio about 1/3. Abided by *Pauli* exclusion principle, only singlet is spin-allowed to efficiently transition radiatively to its ground state for most fluorescent materials while the spin-forbidden decay of triplet (phosphorescence) is usually too slow nor efficient to be useful. It follows that the theoretical limit of η_{INT} is limited to only about 25% without the capability of harvesting the triplet exciton. Moreover, as a Lambertian source, not all the light generated within OLED can escape out of the planar cell through various layers of media with different refractive index (n), for there is a further loss which can be estimated by the Fresnel equation of $\alpha = 1/(2n^2)$. Therefore the external quantum efficiency (η_{EXT}) can be further calculated by the equation of ($\eta_{\text{EXT}} = \alpha \cdot \eta_{\text{INT}}$) where α is designated as the out-coupling efficiency which without means of any light enhancement is about 20% for most organic materials with an average $n \approx 1.5 - 1.9$. Lastly, the power efficiency which is inversely proportional to the driving voltage must also be considered in order to meet the ever increasing demand of high resolution mobile displays. To drive down power consumption by minimizing carrier injection barriers and increasing the probability of recombination near the interface, a typical bottom-emission OLED architecture of today is shown in Fig. 1 (top) with the addition of several key functional organic layers such as electron injection layer (EIL), hole injection layer (HIL) and a R,G,B doped emitter to control color of emission.

Over the last 15 years, tremendous progress has been made in OLED materials overcoming the limitation of η_{INT} of fluorescence which has been enhanced to nearly 100% by exploiting phosphorescence as well as delayed fluorescence. Power efficiency can also be boosted by the discovery of bipolar host material and the introduction of extremely electron deficient (p -) and donating (n -) dopants into the respective HTL and ETL to increase its conductivity as illustrated in Fig. 1 (bottom) with a typical p - i - n OLED energy diagram with its LUMO and HOMO bending after applying a positive bias voltage. Tandem OLED to increase luminance without increasing drive current density nor jeopardizing OLED lifetime and light enhancement technology have also been developed to extract more light out of OLED with the understanding of its optics. As a result of these discoveries, the most eco-friendly, sun-like area luminaire of indoor OLED lighting has now become possible.

Materials

Early OLED materials development has all been focused on the design and synthesis of fluorescent molecules which can only harvest singlet excitons. In the first decade of 21st century, unusually high EL efficiency and lifetime have been reported for a number of blue fluorescence OLEDs, many of which are anthracene-derived fluorophores, such as non-dopant deep blue emitter TAT with $\lambda_{\text{max}}^{\text{EL}}$ 444 nm, CIE_{x,y} (0.16, 0.09), (η_{EXT} 7.18%) and doped sky-blue emitter of arylamine-substituted distyrylarylene DSA-Ph with $\lambda_{\text{max}}^{\text{EL}}$ 465 nm, CIE_{x,y} (0.15, 0.29) in MADN host (η_{EXT} 8.7%). In depth study of TAT has revealed that both triplet-triplet annihilation (TTA) and the outcoupling enhancement by dipole orientation of emitting molecules contribute to the production of singlet exciton from which exceedingly high η_{EXT} is thus achievable.

Major breakthrough of EL efficiency of OLED started to emerge before the end of last century. One of the early examples was a green phosphorescence material known as $\text{Ir}(\text{ppy})_3$ which is an organometallic-based coordination complex containing heavy metal Iridium. Utilizing the heavy atom effect that promotes spin-orbital coupling, these 2nd generation luminescent materials all possess a hybrid emissive excited state of singlet and triplet excitons with a rather short excited-state lifetime (in μs) thus easing up the forbidden phosphorescence at room temperature. Based on this emissive mechanism, numerous heavy metal containing

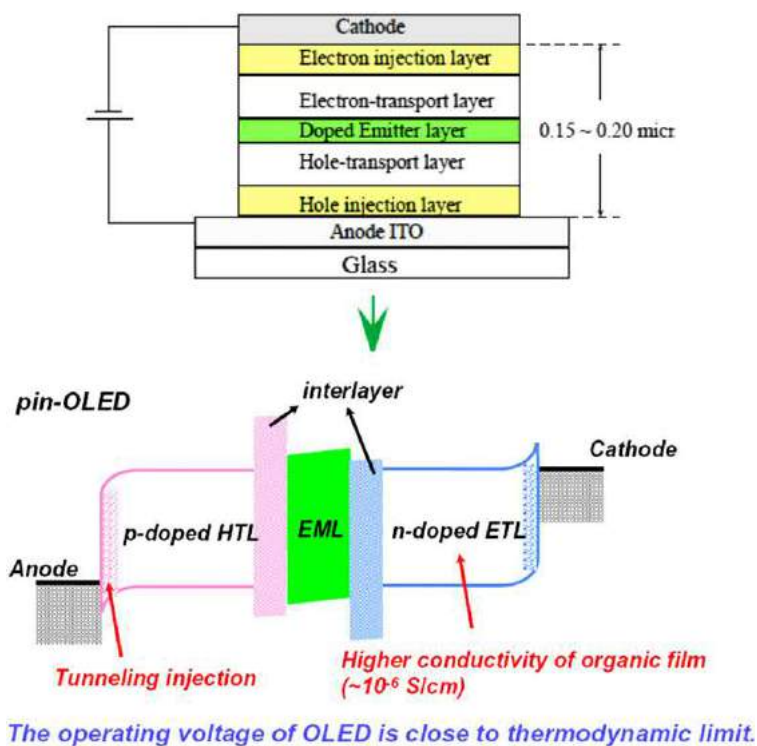


Fig. 1 A typical bottom-emission OLED architecture (top) and energy profile of *p-i-n* OLED upon positive bias (bottom).

(mostly Ir) red, green or even some light blue phosphorescence EL materials have been reported. Meanwhile, hosts derived from carbazole with high triplet energy as well as many triplet exciton blocking materials to prevent luminescence quenching have also been developed. In dopant/host emitter device with appropriate blocking layer, EL power efficiency of >100 lm/W for green $\text{Ir}(\text{ppy})_3$, >25 lm/W for red $\text{Ir}(\text{mphen})_2(\text{acac})$ equivalent to $\eta_{\text{EXT}} \sim 25\%$ and near 50 lm/W for sky-blue Flrpic was reported at brightness of 1000 cd/m^2 . Mainly due to the high EL efficiency (hence low power consumption), today's AMOLED display products in the marketplace all have adopted phosphorescence EL materials for their green and red emitters. The blue phosphorescent material owing to its intrinsic instability and lack of color saturation issues remains to be the *Holy Grail* that has been pursued relentlessly. As a result, display industry presently favors the use of the deep blue anthracene-based fluorescent materials especially the deep blue with CIE_y (<0.10) as dopant for its AMOLED products.

The 3rd generation of luminescence was discovered in 2011 which is based on the up-conversion of triplet exciton into its singlet excited state leading to efficient thermally activated delayed fluorescence (TADF). This type of materials which have the advantage of containing no heavy metal like Ir or Pt (which is expensive and rare) and the potential of harvesting triplet exciton as well as the possibility of breaking the Ir-materials patent gridlock have quickly caught the fancy of a great many researchers. In just a few years thereafter, green 4CZIPN ($\lambda_{\text{max}}^{\text{PL}}$ 507 nm) and DACT-II ($\lambda_{\text{max}}^{\text{PL}}$ 516 nm), greenish blue ($\lambda_{\text{max}}^{\text{PL}}$ 480 nm) SpiroAC-TRZ and a near blue ($\lambda_{\text{max}}^{\text{PL}}$ 460 nm) DMAC-DPS with $\text{CIE}_{x,y}$ (0.16, 0.20) have been reported to achieve η_{EXT} as high as 19.3%, 29.6%, 36.7% and 19.5%, respectively. All these TADF materials have a common structural feature containing a twisted non-coplanar configuration formed by steric crowding between an electron donor (usually arylamine) and an acceptor (e.g., CN, SO_2 , $\text{C}=\text{O}$, $\text{P}=\text{O}$) through certain π -conjugation (such as benzene ring) which has been shown to possess unusually narrow singlet-triplet state bandgap (<0.1 eV) conducive for room temperature intersystem crossing.

In 2015, a series of TADF molecules have been adopted as an assisted dopant (such as ACRSA) in OLEDs to transfer its singlet excitons to the singlet state of a more fluorescent dopant emitter (such as TBPe) via Förster energy transfer process to produce what is called "hyperfluorescence". To differentiate from TADF emitter, a 3.5th generation of OLED materials has been declared after its discovery which opens up a golden opportunity for further research in the pursuit of highly efficient yet still stable emitters particularly for the blue color. This is because the twisted molecular configuration associated with TADF molecules usually have a rather large FWHM which is adverse to the color purity required for display applications. Utilization of conventional fluorescence molecule as the end emitting dopant can alleviate the problem so as to regain its color purity yet still maintain high EL efficiency as well as long device lifetime. Although TADF materials are non-heavy-metal containing, recent study showed that most TADF OLEDs still suffered from fast degradation and "roll-off" of efficiency at high brightness particularly acute for the blue. This is mainly attributed to the synergetic oxidation of the TADF emitters in the devices, including at least an electrochemical oxidation and a photo-oxidation processes. Evolution of OLED materials development and a few examples of key molecular structures are illustrated in Fig. 2.

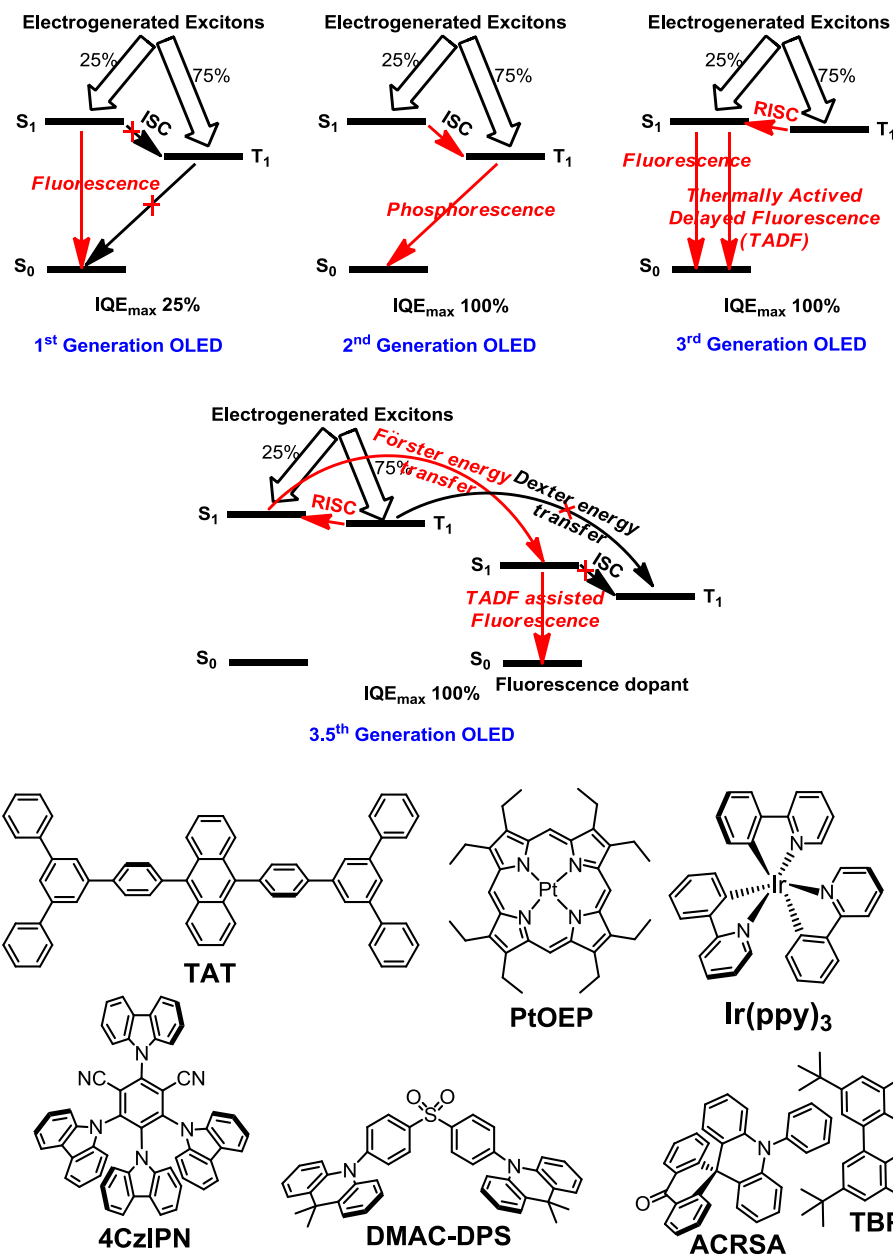


Fig. 2 Evolution of OLED materials development and examples of key molecular structures.

Devices

All good OLED materials invariably need good device architecture to maximize their electroluminescence (EL). This is particularly true in regard to overall performance involving key measurement parameters like stability, power consumption, luminance efficiency and color. OLED device structures can be roughly classified into bottom-emission and top-emission (TOLED) depending upon whether the bottom anode (ITO) or the top cathode is made transparent (with its anode reflective), respectively. If both electrodes are transparent, a *transparent* OLED can also be made. For certain *n*-type TFT driving needs, there have also been developed the corresponding bottom-emission and top-emission *inverted* OLED where the anode and cathode are reversed instead.

For driver electronics, there are passive-matrix (PM-OLED) used mostly for low-information content, small size alphanumeric displays (because of its line-by-line scan requiring very high peak-current density and limitation for high resolution) and the active-matrix TFT driving (AM-OLED) which is the most popular. Low temperature poly-silicon (LTPS) TFT with its high *p*-channel mobility ($\sim 100 \text{ cm}^2/\text{V s}$) is best suited for small to medium displays like the smartphones and tablets while Oxide (IGZO) TFT is particularly suitable for large size displays like OLED TV. To increase the aperture ratio for high-resolution, it is most advantageous to couple LTPS-TFT backplane with a TOLED structure in which its cathode is made semi-transparent (e.g., Ag:Mg 15 nm) and its

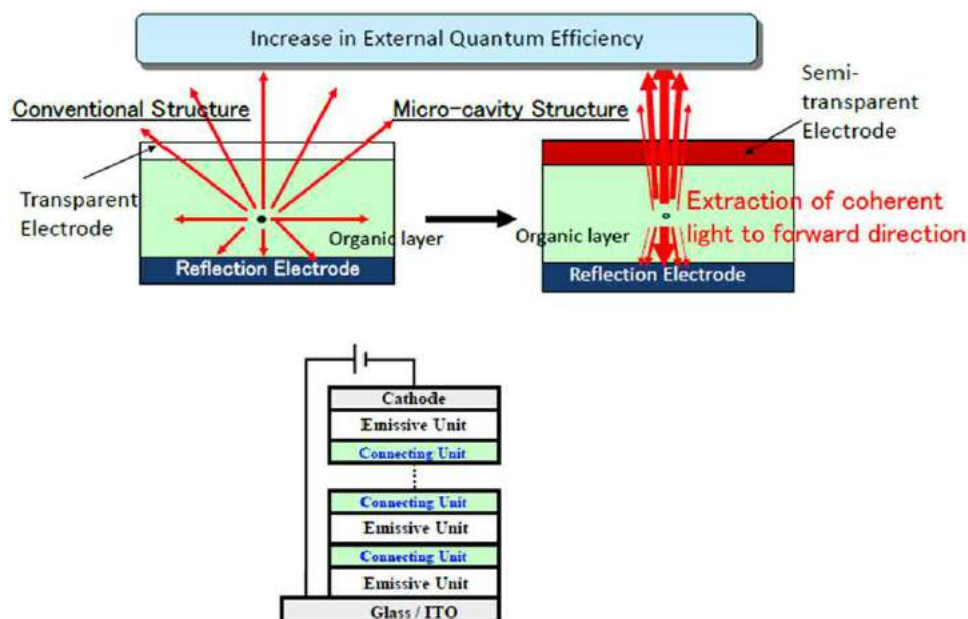


Fig. 3 Microcavity effect of TOLED (top) and a typical tandem OLED structure (bottom).

anode reflective (Ag) as exemplified in Fig. 3 (top). There's also an added advantage in its capability of creating a microcavity within the organic media between the two metallic electrodes one of which is semi-transparent and the other reflective. Its optical effect in terms of constructive (*antinode*) and destructive (*node*) interferences which can be modeled as a Fabry-Perot cavity embedded light source within 2 mirrors, is the dramatic spectral narrowing and intensity enhancement commonly observed in a TOLED. By clever adjustment of the optical length L of the μ -cavity (which is usually done by adjusting the HTL to minimize impact on drive voltage) for each of the color pixel according to the first order approximation of $[2L/\lambda_{max} + \phi/2\pi = m]$ (m =integer) and [full width at half maximum ($FWHM$) = $\lambda^2/2L \times 1 - (R_1R_2)^{1/2}/\pi(R_1R_2)^{1/4}$] where L is the optical length and R_1R_2 are the reflectance of the 2 mirrors, greatly enhanced color gamut by reducing FWHM of each of the R, G, B emission wavelength (λ) and EQE (η_{EXT}) of AM-OLED can be readily achieved.

Another key development in OLED device is the discovery of tandem OLED where multiple OLEDs can be coupled together by transparent connecting charge-generation layer (CGL) which functions as a reversed p/n degenerate contact to create and inject electron and hole back to the ETL and HTL, respectfully for recombination as depicted in Fig. 3 (bottom). As a consequence, luminance of tandem OLED (L_n) can be enhanced at the same driving current density according to the equation of $[L_n = n \times L]$ where n is the number of stacked units and L is the luminance of each single unit. This in turn means that the lifetime of tandem OLED can also be greatly enhanced because the intrinsic stability of organic luminescent materials is singularly dependent upon the driving condition and inversely proportional to the total amount of electric charge passed through the device. Thus, brightness of 3000 cd/m² required for most general indoor lighting is readily attainable with relatively small driving current density.

Recently, due to the high interest of developing OLED as a green and healthy *area* indoor solid state lighting alternative to LED, a great deal of effort has been aimed to extract more light trapped inside OLED. Detailed analysis by rigorous electromagnetic modeling of optics have revealed 4 types of light entrapment, namely radiation mode, substrate mode, waveguide mode and Plasmon mode as depicted in Fig. 4 (top). It also shows that coupling into different modes is highly dependent on the position of emission center which oscillates as illustrated in Fig. 4 (bottom). The second *antinode* appears to give the highest outcoupling efficiency while the surface Plasmon mode can be minimized by fabricating OLED emitter away from its reflective metallic electrode. To date, the most noteworthy light extraction technology is the so-called internal enhancement layer (IES) which is a high-index coupling layer embedded with light scattering particles incorporated in-between the ITO and the glass substrate where η_{EXT} as high as 45% has been achieved in white OLED.

Applications

As display technology, OLED has considerable attributes that are unsurpassed by others. There have been a great many super AMOLED products (e.g., smartphones, tablets, monitors, VR/AR microdisplays, wearables, TVs, etc.) that have successfully entered the consumer electronics market and indeed, many of which had won Best Display Awards (including most recently the 2016 "Best of Innovation" for a 77-in 4K OLED TV). It has also been declared "the best panel ever" as tested by the independent Display Mate Labs in major measurements categories like winning the "highest color accuracy, infinite contrast ratio, lowest screen reflectance, and smallest brightness variation with viewing angle."

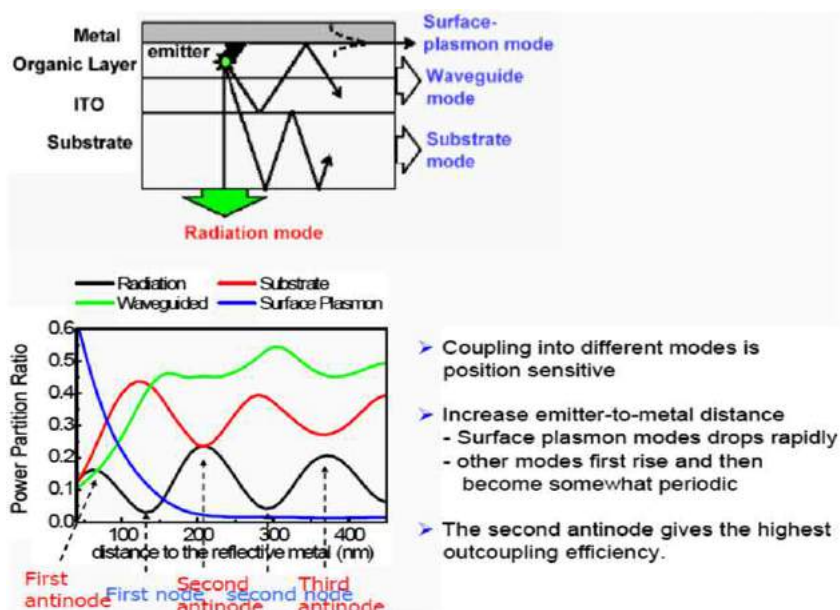


Fig. 4 Electromagnetic modeling of light generated inside OLED (top) and their couplings with respect to the position of emission (bottom).



Fig. 5 Examples of OLED displays and lighting.

As predicted by DisplaySearch and IHS, OLED display revenues will grow to about \$44B in 2020 in which LTPS-TFT AMOLED smartphones will be the main driver overtaking TFT-LCD to become the leading display technology in mobile sector while OLED TV business will begin to grow significantly after 2016. In addition, owing to the tremendous improvement of power efficiency ($> 100 \text{ lm/W}$) as well as lifetime recently, WOLED lighting which has the potential of emitting warm light across large surfaces

and bringing to bear novel *form factor* into the lighting sector is also forecasted to gain market share slowly and eventually to become competitive with the existing LED solid state lighting particularly in indoor illumination. A glossary of several OLED display and lighting products is reproduced in [Fig. 5](#).

Perspectives

After more than 30 years of R&D, OLED technology has finally begun to hit the big time in the display market. It is expected that flexible OLED (FOLED) will be a “killer application” for which OLED industry has long been searching. Indeed, we have witnessed the successful launching of many unbreakable and bendable FOLEDs which are fabricated on thin plastics and continue looking forward to the ultimate realization of foldable and rollable OLEDs in years to come. But in order to become a viable and sustainable business to compete effectively with LCD as well as a host of other new emerging displays like those of Quantum Dot, microLED and laser, many challenges must be overcome. These include issues of deep blue emitter efficiency and lifetime, large area fabrication, mass production, cost down, thin-film encapsulation, robust RGB pixelation (to achieve high yield and resolution), improvement of luminance and power efficiency, etc. Control of optics to maximize out-coupling efficiency of OLED (without impacting on image quality for display application) will remain to be an interesting topic of research and further study in the future.

It suffices to state that the future of OLED as display and lighting is extremely bright and colorful, which presents great opportunities for numerous invention and innovation in R&D as the technology matures and continues to improve over time. We will witness significant growth of OLED investment and business as well as proliferation of AMOLED, FOLED, wearable and VR/AR products like smart phones, PMP, tablets, ultrabooks, monitors and TVs in the marketplace in the foreseeable future. By then, it is predicted that many area OLED lighting luminaire products will also be with us as one of the major green, energy-saving, most healthy (no glare nor blue-light hazard) solid state lighting fixtures in our home. Perhaps, the last question to be pondered is: Will Roll-to-Roll printing of OLED not be too far?

It is impossible to describe OLED science and technology in any breath or depth in these few pages. Interested readers are encouraged to read further with the list of books and references cited in the following.

Acknowledgement

We gratefully thank M.-T. Lee of AU Optonics (Taiwan), Y.-S. Tyan of First-O-Lite (China) and C. C. Wu of National Taiwan University for permission of using excerpts of some of their presentation and lecture notes for graphic illustration in this review.

See also: OLED/Introduction

Further Reading

- Baranoff, E., Yum, J.-H., Graetzel, M., Nazeeruddin, Md. K., 2009. *J. Organometallic Chem.* 694 (17), 2661–2670.
 Han, T.-H., Jeong, S.-H., Lee, Y., *et al.*, 2015. *J. Information Display* 16 (2), 71–84.
 Kaji, H., Suzuki, H., Fukushima, T., *et al.*, 2015. *Nat. Commun.* 6, 8476.
 Kim, S.T., 2006. Keynote in IMID 2006, August 23, 2006, Daegu, Korea.
 Ko, S.H. (Ed.), 2011. *Organic light emitting diode- material, process and devices* (Book, 2011).
 Lee, J.-H., Kim, J.-J., Lee, S., *et al.*, 2013. *J. Information Display* 14 (1), 39–48.
 Liu, C.-C., Liu, S.-H., Tien, K.-C., *et al.*, 2009. *Appl. Phys. Lett.* 94, 103302.
 Liu, J., Chen, C.-T., Chen, C.H., 2015. *Introduction to organic light-emitting diodes (OLED)*. In: Kriss, M. (Ed.), *The Handbook of Digital Imaging*. John Wiley & Sons, Ltd., pp. 577–626.
 Mazzeo, M. (ed), 2010. *Organic Light Emitting Diode* (Book, 2010).
 Nakanotani, H., Adachi, C., 2014. *Nat. Commun.* 5, 4016.
 Raychaudhuri, P.K., Madathil, J.K., Shore, J.D., Van Slyke, S.A., 2004. Performance enhancement of top- and bottom- emitting organic light-emitting devices using microcavity structures. *J. Soc. Inf. Display* 12, 315–321.
 So, F., Adachi, C. (Eds.), 2013. *Organic light emitting materials and devices*. In: XVII SPIE Proceedings, vol. 8829, San Diego, CA.
 Tyan, Y.-S., Rao, Y.-Q., Ren, X.-F., *et al.*, 2009. *Proceedings of SID 2009*, p. 895.
 Uoyama, H., Goushi, K., Shizu, K., Nomura, H., Adachi, C., 2012. Highly efficient light-emitting diodes from delayed fluorescence. *Nature* 492, 234–239.
 Zhang, Q., Li, B., Huang, S., *et al.*, 2014. *Nat. Photonics* 8, 326.

Optical Characteristics of Display Devices

Han-Ping D Shieh, National Chiao Tung University, Hsinchu, Taiwan

© 2018 Elsevier Ltd. All rights reserved.

Definition of a Display

A display can be defined as an object that transfers information to the viewer and/or an output unit that gives a visual representation of data (Lee *et al.*, 2008; Yeh and Gu, 2010; Tsujimura, 2017; Kobayashi *et al.*, 2009).

Optical Characteristics of a Display

Considering that the information provided by a display is delivered via light, the optical characteristics of a display are of great importance. A display has, in general, the optical characteristics of luminance, reflectivity, contrast ratio, color, surface gloss, resolution, and image sharpness (sometimes called small area contrast). Measurement methods can be obtained from the Electronics Industries Alliance (EIA) JT-31 (Optical Characteristics of Display Devices), Society of Automotive Engineers (SAE), International Organization for Standardization (ISO), and Video Electronics Standards Association (VESA) Standards, to name a few (see Relevant Website section).

Radiometric and Photometric Units

To measure light, a certain system of units is needed. Generally, the radiometric or photometric units are adopted for different purposes. The radiometric units refer to a measurement of any portion of the electromagnetic spectrum, such as visible, infrared, and ultraviolet light. In this system, the radiant power is in Watt (W), which is the total radiant energy per unit time (Joule per second, J/s). Different from the radiometric units, the photometric units only deal with visible light and consider the human perception. As the sensitivity of the human vision system varies with wavelengths, the weights of different wavelengths are also different. In the photometric system, the most basic quantity is luminous intensity, which is a measure of the wavelength-weighted power emitted by a light source in a particular direction per unit solid angle. The International System of Units (SI unit) of luminous intensity is candela (cd) (in the English system, cp), which is also an SI base unit. Table 1 lists some common radiometric quantities with their units and corresponding photometric quantities with units.

Luminance

The luminance of a display is the output luminous intensity per unit area in a unit solid angle, which is given in units of candela/square meter or cd/m^2 (in the English system fL), as shown in Eq. (1) and Fig. 1 (Warren, 2007; Wyszecki and Stiles, 1982). Luminance is one of the most important specifications of a display device, which is closely associated with a display's perceptive brightness (Booker, 1981; Ronchi *et al.*, 1980). Generally, a computer monitor may have a luminance of several hundred nits while an outdoor display may have several thousands (Gong *et al.*, 2012; Reinhard *et al.*, 2010). In addition, the luminance may vary with the region on a screen (spatial luminance profile) and viewing angle (angular luminance profile). To measure luminance of a display, a photometer, photometer/colorimeter, or spectroradiometer can be adopted.

$$L = \frac{d\Phi}{d\Omega \cdot dS \cdot \cos\theta} \quad (1)$$

where L is luminance, Φ is luminous flux, Ω is solid angle, S is area, and θ is viewing angle.

Table 1 Common quantities and units in the radiometric and photometric systems

Radiometric system		Photometric system	
Quantity	Unit	Quantity	Unit
Radiant energy	J	Luminous energy	$\text{lm} \cdot \text{s}$
Radiant intensity	$\text{W} \cdot \text{sr}^{-1}$	Luminous intensity	$\text{cd} (\text{lumen} \cdot \text{sr}^{-1})$
Radiant flux	$\text{W} (\text{J} \cdot \text{s}^{-1})$	Luminous flux	$\text{lumen} (\text{cd} \cdot \text{sr})$
Radiance	$\text{W} \cdot \text{sr}^{-1} \cdot \text{m}^{-2}$	Luminance	$\text{nit} (\text{cd} \cdot \text{m}^{-2})$
Irradiance	$\text{W} \cdot \text{m}^{-2}$	Illuminance	$\text{lux} (\text{lm} \cdot \text{m}^{-2})$

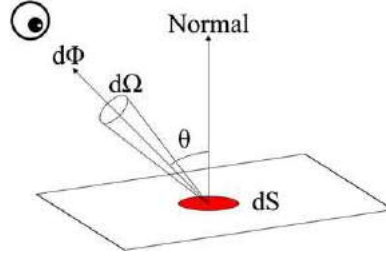


Fig. 1 Calculation of luminance perceived from a direction θ relative to the normal. $d\Phi$, $d\Omega$, and dS are infinitesimal luminous flux, solid angle, and area, respectively.

Illuminance

The illuminance is the total luminous flux incident on a surface, per unit area, as given in Eq. (2) (Warren, 2007; Wyszecki and Stiles, 1982). Its units are lumen per m^2 or lux (and in the English system Ft.cd). This is measured using illuminance meters. The luminance of a display may also vary with the region on the screen.

$$E = \frac{d\Phi}{dS} \quad (2)$$

where E is illuminance, Φ is luminous flux and S is area.

Reflectivity

The reflectivity of a display is the ratio of the reflected light to the incident light. For CRTs, the reflectivity is measured with incident light at 45 degrees, while a smaller angle of 30 degree is used in some cases for LCDs (Keller, 1997). The reflected light from the surface is measured with a detector normal to (perpendicular to) the surface.

Color

A normal human vision system has three kinds of cone cells, which respectively have sensitivity peaks in short (S, 420 nm–440 nm), middle (M, 530 nm–540 nm), and long (L, 560 nm–580 nm) wavelengths. According to the additive color-mixing principle, various color systems have been developed from the working mechanism of the human color vision. With the aid of a color system, although color perception is different from different individuals, the standard responses to a color stimuli can be evaluated and we can measure colors of different objects and display images. Color can be measured with a colorimeter or spectroradiometer (Smith, 1978; Ohno, 2000). The output data are given in various color systems. Color systems, such as the 1931 CIE (Commission International de l'Eclairage) Color System (x, y), the 1960 CIE (u, v) or 1976 CIE (u', v') color coordinates are all used (Wikipedia, 1931, 1960; see Relevant Website section). All these color systems can be converted to each other by simple mathematical relations and these conversions are usually available to the user in the data output. The choice of the color space/system is dependent on application but the system in most general use for displays is the 1976 CIE (u', v'), where the values of u' and v' can be converted from tristimulus values XYZ in 1931 CIE (x, y) by Eq. (3) (Ford and Roberts, 1998). This color system has become more widely used because equally perceived color differences in this color system are at almost equal distances in the color space (McGuire, 1992). In reality, the quest for a system with equal distances equating to equal color differences can be achieved through a complicated matrix approach in which the constants used in the matrix vary from point to point in the color space.

$$\begin{aligned} u' &= \frac{4X}{X + 15Y + 3Z} = \frac{4x}{-2x + 12y + 3} \\ v' &= \frac{9Y}{X + 15Y + 3Z} = \frac{9y}{-2x + 12y + 3} \end{aligned} \quad (3)$$

Gloss

Gloss is a measure of surface reflectivity under Snell's Law criteria (equal incidence and reflection angles). The measurement angles of 20, 45, 60, and 85 degrees are used to measure the gloss of common surfaces (Keller, 1997). Display glass surface measurements tend to use 60 degree gloss angle. The surface of a display can be made rougher to scatter incident light away from the viewer or it can be optically coated to reduce the reflected light to achieve improved viewing. Some times both methods are used (Dearborn Glass Co, 1971). Reduced gloss surfaces are also involved with the reduction in surface glare. The level of increased roughness of the surface affects the user's ability to see details.

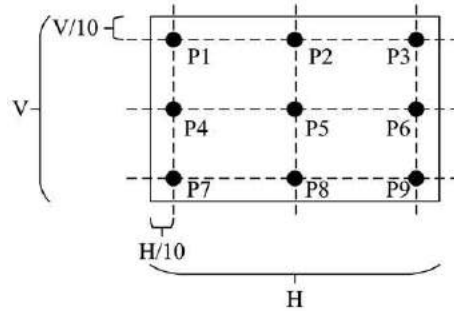


Fig. 2 9-point testing of the uniformity of a display, where the black dots denote the testing points.

Uniformity

The uniformity of a display is how consistent an image appears on the screen in terms of luminance, color, etc. For instance, luminance uniformity is defined as the ratio of the smallest luminance on the screen to the largest, which is usually required to be higher than 0.75 for a qualified display (Keller, 1997). To measure the luminance uniformity (or uniformity in terms of other metrics), a 9-point testing is often implemented, which measures nine testing points uniformly distributed on the screen with the aid of certain test charts, as shown in Fig. 2. Sometimes, a similar 16-point testing is also adopted. Also based on the 9-point testing, the chromatic uniformity in 1976 CIE (u' , v') color space can be measured, as given by Eq. (4).

$$\Delta u'v' = \frac{\sum \sqrt{(u'_i - u'_0)^2 + (v'_i - v'_0)^2}}{N} \quad (4)$$

where (u'_0 , v'_0) is the chromaticity coordinate of a reference, such as the central point of the screen or a designated color.

Contrast Ratio

When people talk about contrast they generally mean “contrast ratio.” The contrast ratio of an area on a display is a unitless parameter that is defined as the ratio of the luminance of the brightest color to that of the darkest color (Lee *et al.*, 2008; Yeh and Gu, 2010). Contrast ratio is a general concept that has not been officially standardized across different types of displays and measurement methods. For LCDs or other modern displays, the most widely adopted conceptions are static contrast ratio and dynamic contrast ratio.

Static contrast ratio is the fundamental and inherent characteristic of a display that is the luminosity ratio of the brightest color to the darkest that the display is able to produce simultaneously at any instant of time. Two methods are usually used to measure the static contrast ratio, as (1) native on/off method, which sequentially adopts a full black signal and then a full white signal, and (2) ANSI method, which adopts a checkerboard pattern containing fully black and white colors simultaneously (Pavlovych and Stuerzlinger, 2005). The native on/off contrast ratio is always a little higher than the ANSI one, so commercial products are always characterized by ANSI contrast ratio. In LCDs, certain backlight intrinsically leaks through LC cells even if a black signal is displayed; i.e., LC cells cannot be completely closed, which restricts the contrast ratio of an LCD. Generally, modern LCDs can achieve a native on/off contrast ratio of approximately a thousand to one. Next-generation displays, such as OLED and micro-LED displays (Pavlovych and Stuerzlinger, 2005; Putman, 2016), directly emit light from the screen without backlight; thus the pixels actually produce “no light” for a black signal, and then, the contrast ratio can be way much higher.

Dynamic contrast ratio is the luminosity ratio of the brightest to the darkest color over time, which is usually measured by dynamic on/off method. What is different from the native on/off method is that a mechanical iris or electronic signal processing is allowed to help increase the contrast ratio. Local dimming or some brightness boost technologies that can be found in modern displays can also be utilized (Huang *et al.*, 2011). Obviously, dynamic contrast ratio measured in such a manner is (much) higher than static contrast ratio. Some LCD manufacturers announce that their products can even reach a dynamic contrast ratio of higher than 10,000,000:1. However, since software or circuit based contrast ratio enhancement technologies and measurement conditions have not been standardized, dynamic contrast ratio is usually not compared across different display devices.

Additionally, the effect of ambience on contrast ratio measurement should also be considered sometimes. If the measurement is implemented in a room with reflective walls, light threw out from the display will be reflected back by the walls and then degrade the contrast ratio.

Finally, the variation of contrast ratio with viewing angle also needs to be paid attention. Modern display technologies, such as emissive OLED displays or in-plane switching (IPS) LCDs, can realize quite homogenous contrast ratio over a wide range of viewing angles, while twisted nematic (TN) LCDs have a poorer performance, like that shown in Fig. 3.

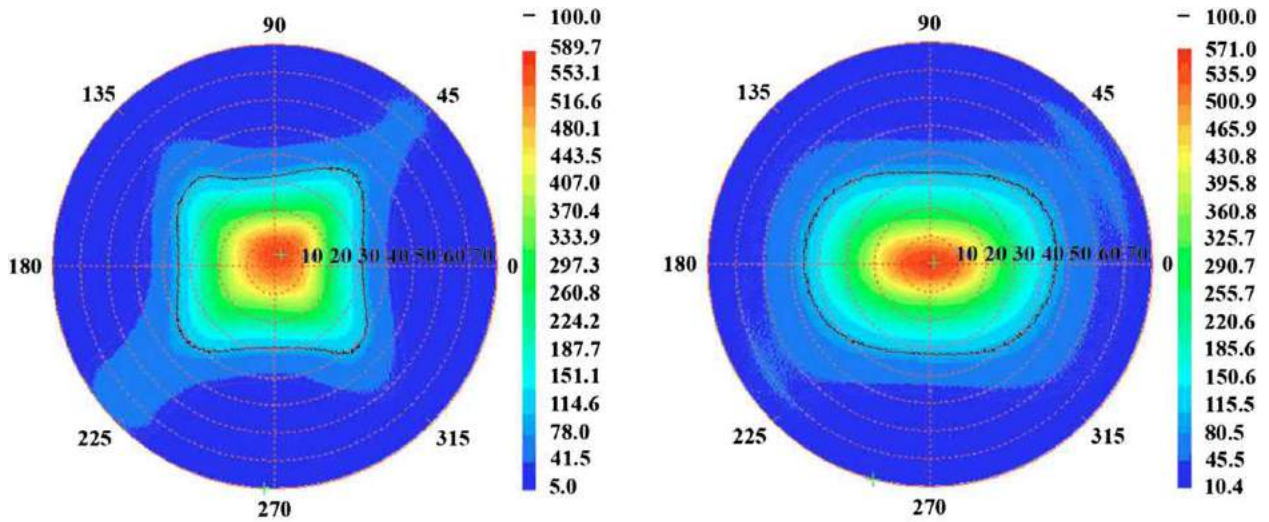


Fig. 3 Iso-contrast contours of angular-dependent contrast ratios of a conventional TN LCD (left) and a viewing-angle-enhanced LCD (Hwang *et al.*, 2009). Reproduced from Hwang, S., Min, J., Lee, M.G., *et al.*, 2009. Novel wide viewing liquid crystal display with improved off-axis image quality in a twisted nematic configuration. *Optical Engineering* 48(11). doi: 10.1117/1.3265524.

Ambient Illumination and Reflection

A display for users to see the information is exposed in an ambient illumination, like bright daylight; hence ambient contrast ratio (ACR) that takes reflected ambient light into account is usually used. ACR of a display with an overall reflectivity R can be calculated with Eq. (5) (Singh *et al.*, 2012; Qin *et al.*, 2015). It can be seen that higher ambient luminance contributing both to the numerator and the denominator leads to a lower ACR.

$$ACR = \frac{L_{on} + R \cdot L_{amb}}{L_{off} + R \cdot L_{amb}} \quad (5)$$

where L_{on} , L_{off} , and L_{amb} is luminance of the brightest color, the darkest color, and the ambience.

With the development of flat panel displays, displays are widely used in portable electronic devices. It is a common scenario that a user watches a display in the bright daylight; thus the "specular reflection" (Snell's Law type of reflection) on the front glass of the display will degrade the contrast ratio severely. This reflection is reduced by methods such as anti-glare (AG) and anti-reflecting (AR) surfaces applied to the front surface of a display. Surface roughness is used in AG surfaces to scatter a portion of the specular light away from the viewer. Surface roughness can be achieved through surface abrading, surface plastic coating, surface silicate coating, and surface acid etching to name a few methods. AR surfaces use optical interference techniques to reduce the reflected light to the viewer. In some cases, both methods are applied to a display surface. If an AG coating is used, the image sharpness will be affected by the reduction in specular reflection through the use of surface roughness; thus ambient light reflection and image sharpness should be traded off to acquire the optimum image quality. On the other hand, AR coatings improve glare reduction without scattering and thus give higher resolution capability to the displays but this is usually obtained at a higher price than the AG coating (Chen, 2001).

Moreover, like OLED displays, the ambient light reflection is caused by the metallic electrodes inside the display, hence AG or AR coating on the front surface does not take effect. The most common approach for such an internal reflection is to use a linear polarizer along with a quarter wavelength retarder, so that the ambient light can be absorbed before it exits the display, as shown in Fig. 4 (Singh *et al.*, 2012; Qin *et al.*, 2015). Nevertheless, the use of a polarizer cuts off at least half of the emissive light, and a considerable thickness of the polarizer plus the retarder is not expected in flexible displays. Therefore, some new techniques are under development, such as "black cathode (Xie and Hung, 2004)," microlens array with black matrices (Qin *et al.*, 2015), etc.

Bidirectional Reflectance Distribution Function

An effective tool is bidirectional reflectance distribution function (BRDF), which can describe reflection properties at different viewing angles versus light with different incident angles (Marschner *et al.*, 2000). Fig. 5 and Eq. (6) explains how BRDF works with four variables. Its value equals to the ratio of the reflected radiance in the outgoing direction to the irradiance in the incoming direction. BRDF can comprehensively depict the reflective property of a surface, so that it can be widely utilized in optical

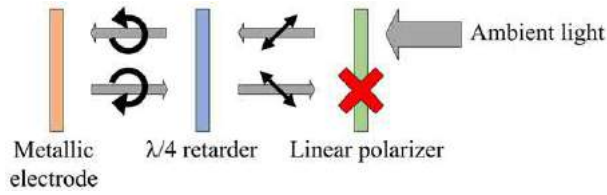


Fig. 4 Elimination of ambient light reflection in a display containing metallic electrodes by using a linear polarizer with a quarter wavelength retarder. The straight and circular arrows demote linearly and circularly polarized light, respectively. The directions of the arrows denote the polarization direction.

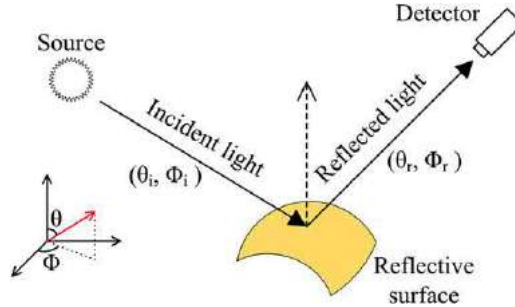


Fig. 5 Illustration of the definition of BRDF. θ and Φ denote the direction. The subscripts “i” and “r” denote incoming and reflected light, respectively.

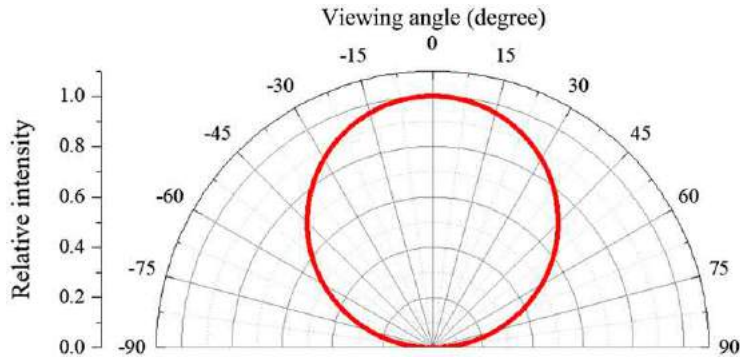


Fig. 6 Angular luminous intensity distribution of a Lambertian surface in a polar coordinate system.

design, computer graphic, computer vision, etc. Common BRDF models include Lambertian, Lommel–Seeliger, Phong etc. (Langovoy and Wübbeler, 2013)

$$BRDF(\theta_i, \Phi_i, \theta_r, \Phi_r) = \frac{L(\theta_r, \Phi_r)}{E(\theta_i, \Phi_i)} \quad (6)$$

where L donates radiance and E donates irradiance.

Lambertian Surface

An ideal diffusive surface that is important in displays obeys Lambert's cosine law, so called Lambertian surface (Warren, 2007). A Lambertian surface reflects light with luminous intensity proportional to cosine of the viewing angle (the angle between the direction of reflected light and the surface normal), whose luminous intensity profile is shown in Fig. 6. More important, its luminance is isotropic regardless the viewing angle, resulting in appearing equally bright to observers in all directions. BRDF of Lambertian surface is a special case that BRDF is a constant of ρ/π where ρ is the albedo.

Viewing Angle and Viewing Cone

Viewing angle is defined as the maximum angle, within which a display can provide acceptable visual performances (luminance, chromaticity, etc.) to the observers (Lee *et al.*, 2008; Yeh and Gu, 2010; Teunissen *et al.*, 2008). A display may perform differently in different directions, hence horizontal and vertical viewing angles are usually used to gauge a display. In fact, the most frequently used viewing angle is defined by the luminance. According to different definitions, the viewing angle in terms of luminance can be defined as the angle where the luminance descends to one half or one third of the on-axis value. To be more comprehensive, a more complicated concept, as viewing zone, is sometimes adopted for a display. Viewing cone is defined as “the range of viewing directions that can satisfactorily be used for the intended task without reduced visual performance” in ISO standards, which is actually the multitude of viewing angles in all viewing directions, leading it to appear in an inverted “cone” with its apex on the display. According to different usage scenarios, different aspects of performances, such as luminance, chromaticity, contrast, can be adopted to define the viewing cone. To conveniently represent a viewing cone, a contour or heat map in a polar coordinate system is usually recommended.

Resolution and Image Sharpness

The resolution of a television, computer monitor, cellphone display etc. is a specification that denotes the distinct pixel amount in each dimension that can be displayed. Usually, resolution with the unit of pixel is expressed in width $\times$ height. For example, 1920×1080 means 1920 pixels in the horizontal direction and 1080 pixels in the vertical direction. For modern displays with a fixed pixel array, such as LCD, digital light processing (DLP) projector, OLED display, etc., the display resolution is directly the number of physical pixels. Some common display resolutions are shown in Table 2. The resolution of a display also refers to the image sharpness of it; that is, the higher resolution, the finer image detail that can be displayed.

Pixels Per Inch

Sometimes, to give not only the total pixel number, but also the pixel density, another parameter associated with display resolution, as pixels per inch (PPI), is also used. By giving PPI, the viewing distance, over which the human eyes cannot distinguish between pixels of a display, can be determined. For instance, a pixel density around 400 PPI is satisfactory for high-end smartphones; virtual-reality (VR) and augmented-reality (AR) devices usually call for a PPI larger than 600; and light field displays even requires a PPI of several thousands. Table 3 shows PPIs of some typical consumer electronic products released these years.

Color Temperature

Color temperature originally comes from ideal black-body radiators that emit light with different hues at different temperatures, and is adopted to evaluate the hue of light sources in Kelvin (K) (Wyszecki and Stiles, 1982). The Planckian locus in Fig. 7 shows the chromaticities of a black-body radiator of different temperatures on CIE 1931 (x, y) chromaticity diagram. Generally speaking, the lower the color temperature is, a white light appears more reddish; the higher the color temperature is, the white is more bluish. For stimuli that do not exactly follow the form of black-body radiators, the concept of correlated color temperature (CCT) is commonly used, which is the temperature of a Planckian radiator whose perceived color most closely resembles that of the given stimulus at the same brightness and under specified viewing conditions. In Fig. 8, chromaticity coordinates having the same CCT are marked with black lines, revealing that different colors may have the same CCT.

Under NTSC, PAL TV norms, a display should reproduce electrically black and white signals with the minimal color saturation at a color temperature of 6500K (The value may differ in some other norms). By considering that it is high-cost to precisely preprogram or calibrate a display's color temperature to be 6500K or other designated values, many consumer products have

Table 2 Some common display resolutions

Standard name	Aspect ratio	Width (pixel)	Height (pixel)
SVGA	4:3	800	600
XGA	4:3	1024	768
WXGA	16:9	1280	720
HD	16:9	1366	768
HD +	16:9	1600	900
UXGA	4:3	1600	1200
FHD	16:9	1920	1080
WQHD	16:9	2560	1440
4K UHD	16:9	3840	2160
8K UHD	16:9	7680	4320

Table 3 PPIs of some typical consumer electronic products released these years

<i>Product</i>	<i>Release year</i>	<i>Diagonal size (inch)</i>	<i>PPI</i>
Smartphones			
Apple iPhone 4	2010	3.5	330
Samsung Galaxy S II	2011	4.3	216
Apple iPhone 6 Plus	2014	5.5	401
Google Nexus 6	2014	6	490
Sony Xperia XZ Premium	2017	5.8	801
Tablets and laptops			
Apple MacBook Air 13"	2010	13.3	128
Microsoft Surface RT	2012	10.6	148
Amazon Kindle Fire HD	2013	7	216
Apple iPad Air	2013	9.7	264
Apple MacBook Pro Retina 15"	2015	15.4	220

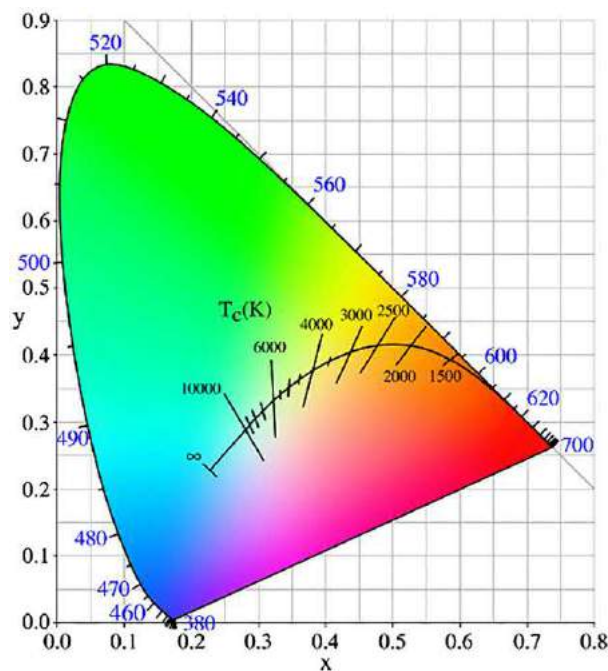


Fig. 7 Planckian locus on the CIE 1931 (x, y) chromaticity diagram that shows chromaticities of black-body radiators corresponding to different temperatures.

deviations from this requirement, and consequently produce some chromatic differences. To meet different personal preferences of color reproduction, some displays also provide a function of color temperature adjustment.

Color Triangle

In an additive color system, the gamut of all the possible combinations of all the primary colors can be plotted in a polygon with vertices defined by these primaries (Wyszecki and Stiles, 1982; Smith, 1978). For three primaries, it is a triangle, which is simultaneously the chromaticity gamut when the amounts of the primaries are constrained to be nonnegative, as the Maxwell's triangle shown in Fig. 8. In a Maxwell's triangle, corners of the triangle are three primary colors – red (R), green (G), and blue (B), and other colors are represented by distances from each of the three sides of the triangular to a point inside. After the CIE system was developed, color triangles are used along with chromaticity diagrams. Generally, the sum of the three chromaticity values in modern chromaticity diagrams has a fixed value and only two of the three values are depicted, hence color triangles are plotted in a Cartesian coordinate system. Fig. 9 shows color triangles of some typical color gamuts on CIE 1931 (x, y) chromaticity diagram.

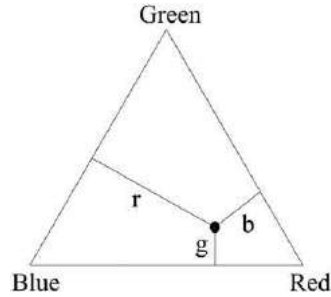


Fig. 8 Maxwell's color triangle, in which a color point (the black dot) can be represented by the distances from each of the three sides, as r , g , and b .

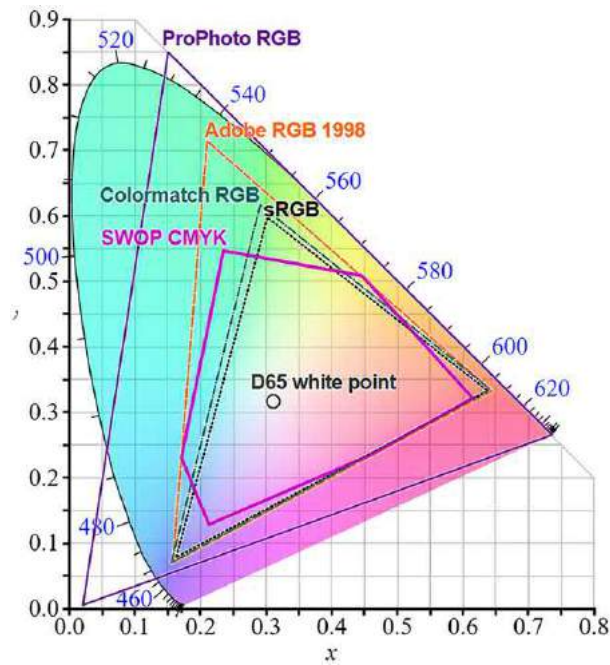


Fig. 9 Color triangles of some typical RGB and CMYK (cyan, magenta, yellow, black) color gamuts, where the name of the color gamuts are marked in the figure.

Response Time

Response time of a display is the time for a pixel takes to change status and is usually with the unit of millisecond (Lee *et al.*, 2008; Tsujimura, 2017). Generally, the “change” is defined as that a pixel goes from black to white and back to black again, hence response time is the sum of T_r (time rising) and T_f (time falling). Sometimes, grey-to-grey (GTG) response time that represents the time for a pixel takes to change from a grey value to another is also adopted to acquire a better-looking specification. A long response time may create comet-like tails while objects are quickly moving on a display, so called trailer effect or motion blur; therefore, response time is important to fast moving contents such as sports or video games. LCDs, as the most widely used display technology today, have an inherent issue of low response time, because the liquid crystals need some time to twist. Modern LCDs can achieve a response time of several milliseconds. Blue phase mode LCDs can even decrease the response time to about one millisecond (Ge *et al.*, 2009).

Refresh Rate and Frame Rate

For displays, refresh rate is the number of times in a second that a display hardware updates its buffer, as known as temporal resolution, while frame rate is the number of times in a second that the image source provides an entire frame of new data to a display. Sometimes, the refresh rate may not equal to the frame rate. For example, an LCD operated under a refresh rate of 60 Hz is

provided with 30 frames per second, so that its refresh rate (60 Hz) is unequal to the frame rate (30 Hz). Since pixels of an LCD do not flash from a frame to another, a minimal refresh rate is not required to eliminate the refresh-induced flicker, whereas this is important for CRTs, whose refresh rate usually needs to be higher than 85 Hz to eliminate the flicker. On the other hand, a high refresh rate of LCDs can smooth the motion objects in the image. Generally, a monitor or TV LCD for regular applications has a refresh rate of 60 Hz. Moreover, LCDs dedicated to video games usually have a refresh rate of 120 Hz, or even 180 Hz to match the high frame rate provided by a graphics card. For some special applications, such as field sequential color displays (Huang *et al.*, 2011), the refresh rate needs to be 240 Hz or even higher. For LCDs with a high refresh rate, the pixel response time should be short enough to realize a rapid grey-to-grey conversion; otherwise, the images will be distorted.

See also: Color Formation of LCD – Spatial and Temporal

References

- Booker, R.L., 1981. Luminance–brightness comparisons of separated circular stimuli. *JOSA* 71 (2), 139–144.
- Chen, D., 2001. Anti-reflection (AR) coatings made by sol–gel processes: A review. *Solar Energy Materials and Solar Cells* 68 (3), 313–336.
- Dearborn Glass Co, 1971. Method of producing glare-reducing glass surface, U.S. Patent 3,616,098, issued October 26.
- Ford, A., Roberts, A., 1998. *Colour Space Conversions*, 1998. London: Westminster University, pp. 1–31.
- Ge, Z., Gauza, S., Jiao, M., Xianyu, H., Wu, S.-T., 2009. Electro-optics of polymer-stabilized blue phase liquid crystal displays. *Applied Physics Letters* 94 (10), 101104.
- Gong, R., Haisong, X., Wang, B., Ronnier Luo, M., 2012. Image quality evaluation for smart-phone displays at lighting levels of indoor and outdoor conditions. *Optical Engineering* 51 (8), doi:10.1117/1.OE.51.8.084001.
- Huang, Y.P., Lin, F.C., Shieh, H.P.D., 2011. Eco-displays: The color LCD's without color filters and polarizers. *Journal of Display Technology* 7 (12), 630–632.
- Hwang, S., Min, J., Lee, M.G., *et al.*, 2009. Novel wide viewing liquid crystal display with improved off-axis image quality in a twisted nematic configuration. *Optical Engineering* 48 (11), doi:10.1117/1.3265524.
- Keller, P.A., 1997. *Electronic Display Measurement: Concepts, Techniques, and Instrumentation*. Wiley-Interscience.
- Kobayashi, S., Mikoshiba, S., Lim, S. (Eds.), 2009. *LCD Backlights* 21. John Wiley & Sons.
- Langvov, M., Wübbeler, G., 2013. Empirical BRDF models for standard reference materials, EURAMET JRP-i21 Deliverable 4.1.
- Lee, J.H., Liu, D.N., Wu, S.T., 2008. *Introduction to Flat Panel Displays*, 20. John Wiley & Sons.
- Marschner, S.R., Stephen, H.W., Lafortune, E.P.F., Torrance, K.E., 2000. Image-based bidirectional reflectance distribution function measurement. *Applied Optics* 39 (16), 2592–2600.
- McGuire, R.G., 1992. Reporting of objective color measurements. *HortScience* 27 (12), 1254–1255.
- Ohno, Yoshi, 2000. CIE fundamentals for color measurements. In: *NIP & Digital Fabrication Conference*, vol. 2000, no. 2, Society for Imaging Science and Technology, pp. 540–545.
- Pavlovych, A., Stuerzlinger, W., 2005. A high-dynamic range projection system. In: *Photonics North 2005 International Society for Optics and Photonics*, 596920.
- Putman, P.H., 2016. Display Technology: The Next Chapter. *SMPTE Motion Imaging Journal* 125 (3), 30–34.
- Qin, Zong, Yeh, Yen-Wei, Huang, Yi-Pai, David Shieh, Han-Ping, 2015. High ambient contrast ratio OLED display with microlens array and cruciform black matrices. In: *Photonics Conference (IPC)*, 2015, pp. 32–33. IEEE.
- Reinhard, E., Heidrich, W., Debevec, P., *et al.*, 2010. *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*. Morgan Kaufmann.
- Ronchi, L.R., Macii, R., Stefanacci, S.R., Bassan, M., 1980. Brightness and luminance of light-emitting diodes: Influence of size and defocus. *Color Research & Application* 5 (4), 207–211.
- Singh, R., Narayanan Unni, K.N., Solanki, A., 2012. Improving the contrast ratio of OLED displays: An analysis of various techniques. *Optical Materials* 34 (4), 716–723.
- Smith, A.R., 1978. Color gamut transform pairs. *ACM Siggraph Computer Graphics* 12 (3), 12–19.
- Teunissen, K., Qin, S., Heynderickx, I., 2008. A perceptually based metric to characterize the viewing-angle range of matrix displays. *Journal of the Society for Information Display* 16 (1), 27–36.
- Tsujimura, T., 2017. *OLED Display Fundamentals and Applications*. John Wiley & Sons.
- Warren, J.S., 2007. *Modern optical engineering. The Design of Optical Systems*. 205–216.
- Wikipedia, 1931. CIE color space, <https://en.wikipedia.org/wiki/CIE-1931-color-space>.
- Wikipedia, 1960. CIE color space, <https://en.wikipedia.org/wiki/CIE-1960-color-space>.
- Wyszecki, G., Stiles, W.S., 1982. *Color Science*, 8. New York: Wiley.
- Xie, Z.Y., Hung, L.S., 2004. High-contrast organic light-emitting diodes. *Applied Physics Letters* 84 (7), 1207–1209.
- Yeh, P., Gu, C., 2010. *Optics of Liquid Crystal Displays*, 67. John Wiley & Sons.

Relevant Websites

- <https://en.wikipedia.org/wiki/CIELUV>
CIELUV.
- <https://en.wikipedia.org/wiki/Electronic-Industries-Alliance>
Electronic Industries Alliance.
- <https://www.nist.gov/director/pao/nist-general-information>
NIST.
- <http://standards.sae.org/>
SAE.
- <https://www.vesa.org/about-vesa/>
VESA.

Reflective Display Technologies

Han-Ping D Shieh, National Chiao Tung University, Hsinchu, Taiwan

© 2018 Elsevier Inc. All rights reserved.

In addition to transmissive and emissive displays, reflective displays are also widely used in display applications that call for ultra-low power consumption, sunlight readability, flexibility, low cost etc., such as electronic paper (e-paper), digital signages, electronic shelf labels, wearable displays. For instance, e-paper is a typical application, which is predicted to share more than 10% of the traditional printing and publishing market (Wu and Yang, 2001; Luxresearch, 2015). From 2015 to 2020, the overall market of reflective displays will grow by 30 million dollars reportedly (DisplaySearch, 2016). Several mainstream reflective display technologies including electrophoretic, electro-wetting etc., will be introduced in this chapter.

Electrophoretic Display (EPD)

Electrophoretic is an electro-kinetic phenomenon that refers to the motion of dispersed particles in a fluid when an electric field is applied, which was observed in as early as 1807 (Dukhin and Derjaguin, 1974). Based on this phenomenon, electrophoretic displays (EPDs) are developed (Dukhin and Derjaguin, 1974; Comiskey et al., 1998; Kim et al., 2004; Amundson et al., 2001; Inoue et al., 2002; Oh et al., 2009; Lee and Liang, 2003; Liang et al., 2003; Huang et al., 2013; Li et al., 2016). Generally, an EPD contains two electrode layers sandwiching an ink layer where white particles (usually positively charged titanium dioxide) and black particles (usually negatively charged carbon black) are suspended, as shown in Fig. 1(a). By applying an appropriate voltage across the electrode layers, the particles will move under the electrostatic force. When the white particles gather at the top electrode, the ambient light will be diffusely reflected and a white state will be observed. In the opposite case, the black particles are pushed to the top electrode, hence the ambient light is absorbed and a black state is given. To form an image, active matrix driving technology is usually utilized to separately drive a number of regions of the display. In addition to white and black particles with opposite charges, another alternative principle of EPD is to use particles with opposite charges and colors respectively at both sides, as shown in Fig. 1(b). When an electric field is applied, the particles will be rotated and moved by the field stress; hence, the color appearing to the observers changes. For an actual EPD, the driving waveform, which forces the charged particles to move towards or away from the observers, plays an important role in sequentially erasing the original image, activating the particles, and displaying a new image. Therefore, the driving waveform algorithm should be carefully optimized for driving time, grayscale, flicker and ghost.

The most significant advantage of EPD is its power saving feature introduced by the “bistability,” which denotes that a system has two stable equilibrium states. For displays, “bistability” means a display can retain a still image without power consumption in two states (usually black and white). In an EPD, positions of the charged particles are retained when the applied voltage is removed. Consequently, the displayed image can be kept when the power source is removed and the display consumes little power when it shows still images. In contrast, in traditional LCDs, the backlight needs to be lightened and the driving circuit needs to be rapidly refreshed all the time. Next, EPDs use basically the same pigments as the regular ink for printed paper to reflect the ambient light, hence it has the same satisfactory contrast ratio and viewing angle as printed paper, especially in the bright sunlight. Thirdly, the displayed images of EPDs are essentially generated by reflected ambient light, hence the healthily harmful light possibly in transmissive and emissive displays can be prevented. Fourthly, the resolution of EPDs can directly benefit from active matrix driving technology that is fast growing (Amundson et al., 2001). Additionally, all components of an EPD can be made flexible to realize a bendable or even rollable display. Therefore, EPDs are now widely used for e-readers and gain increasing attention for architecture displays, wearable displays, electronic signages etc. However, the response time of EPDs of hundreds of milliseconds somewhat restricts their application for animation or video contents. In addition, due to the difficulty of accurately controlling the charged particles, the bit number of gray-level of an EPD is a little poor in comparison with conventional LCDs. For example, 6-bit gray-level is quite prominent for an EPD while an entry-level LCD can easily achieve 8- or even 10-bit gray-level. This issue of bit number further affects the color rendering capacity. Currently, an active research area of EPD is to enhance the color gamut to expand the application scenarios.

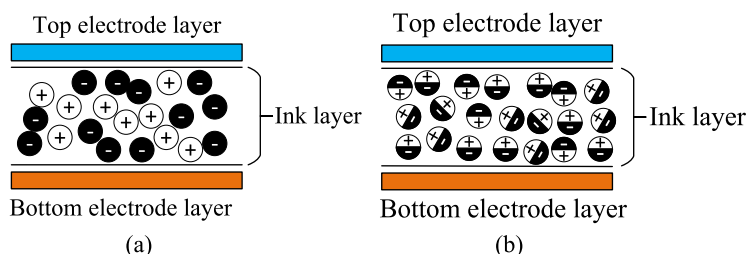


Fig. 1 Basic principle of EPD: (a) using black and white particles with opposite charges; (b) using opposite charges and colors at both sides of a single particle.

Microencapsulated Electrophoretic Display

In an EPD, to solve the issues of lateral drift caused by the gravitation and particle agglomeration caused by the direct contact between the particles and electrodes, the particles suspended in a fluid are enclosed in tiny microcapsules with a scale of dozens of microns, so called microencapsulated EPD (Oh *et al.*, 2009). Microencapsulated EPDs can achieve improved reliability and realize direct coating of microencapsulated electrophoretic materials on the substrate, which leads to potential flexibility. Fig. 2 shows the principle of a monochrome microencapsulated EPD using white and black pigments with opposite charges. The applied voltage pushes the charged white or black pigments to the top electrode to make the display appear white or black to the observers. The voltage applied to each microcapsule can also be bifurcated, then the mixture of white and black particles can realize different gray levels. If color filters are added onto these microcapsules, reflected ambient light passing through the color filters can realize a chromatic microencapsulated EPD (Oh *et al.*, 2009), as shown in Fig. 3.

Microcup[®] Electrophoretic Display

Another practical EPD technology uses microcups to enclose the particles and fluid, so called microcup[®] EPD (Lee and Liang, 2003; Liang *et al.*, 2003; Huang *et al.*, 2013). An example of monochrome microcup[®] EPD with two types of pigment is shown in Fig. 4. The size of a microcup, which can be prepared by a lithographic or embossing process, may vary from 50 to 180 μm in diameter, 12 to 40 μm in height, and 8 to 25 μm in partition width, depending on the application. A unique feature of microcup[®] EPD is that the microcups can be easily imprinted onto a flexible substrate and driving backplane with a mature roll-to-roll (R2R) processing, thus flexible microcup[®] EPDs can be produced at relatively low cost rather than microencapsulated EPDs. Additionally, microcup[®] EPDs can also be made chromatic by using chromatic fluid or a three-pigment-ink system (Oh *et al.*, 2009; Lee and Liang, 2003; Liang *et al.*, 2003; Huang *et al.*, 2013; Li *et al.*, 2016), as shown in Fig. 5. Benefiting from no use of color filters, the reflectivity can be made quite high for a chromatic microcup[®] EPD.

Electro-Wetting Display (EWD)

The electro-wetting phenomenon refers to a modification of the wetting properties of a surface, which is typically hydrophobic, with an applied electric field (Lippmann, 1875). In electro-wetting, two terms including surface tensions and electrostatics are involved to acquire a force equilibrium, as given by Eq. (1).

$$\gamma_{LV}\cos\theta = \gamma_{SV} - \gamma_{SL} + \frac{\epsilon_0\epsilon_r}{2d}V^2 \quad (1)$$

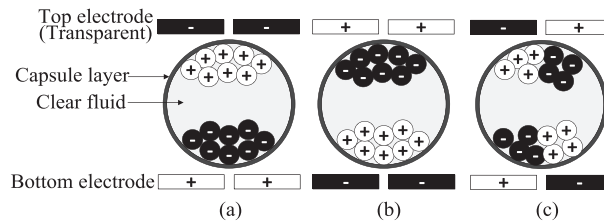


Fig. 2 Principle of a monochrome microencapsulated EPD where (a), (b) and (c) show the cases of white, black and gray visible to the observer, respectively.

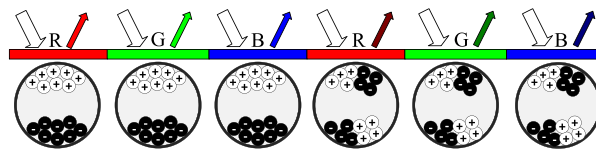


Fig. 3 Principle of a chromatic microencapsulated EPD where red (R), green (G) and blue (B) color filters are added.

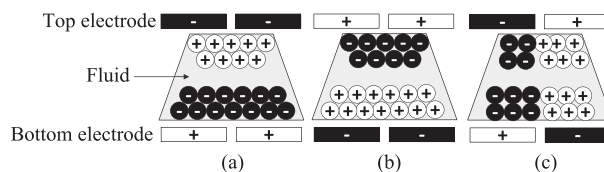


Fig. 4 Principle of a monochrome microcup[®] EPD where (a), (b) and (c) show the cases of white, black and gray, respectively.

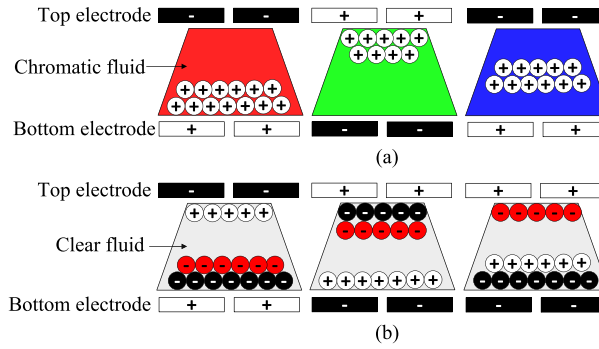


Fig. 5 Principle of a chromatic microcup[®] EPD: (a) using chromatic fluid and control the position of particles in a microcup to generate colors with different brightness; (b) using a three-pigment-ink system to generate white, black and one more color, which is usually used to enable electronic signages.

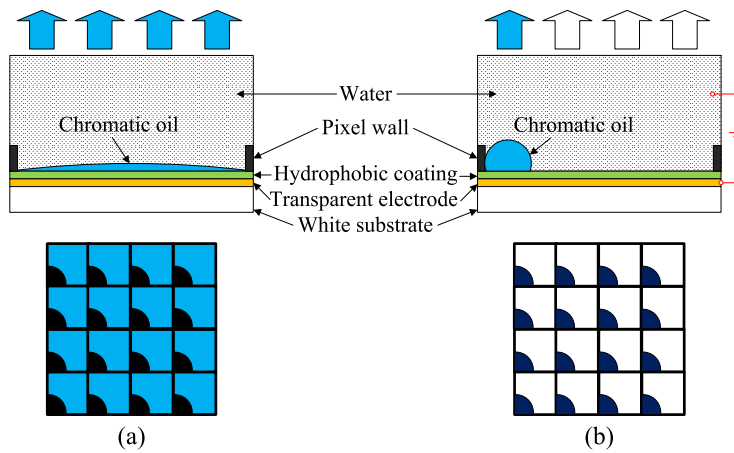


Fig. 6 Principle of a typical EWD using chromatic oil and water: (a) colored (blue) pixel; (b) white pixel where part of the electrode is omitted in the lower left corner of each pixel to control the oil motion.

where γ_{LV} , γ_{SV} , γ_{SL} is respectively surface tension of the liquid/vapor, solid/vapor, and solid/liquid interface; θ is the contact angle; ϵ_r is dielectric constant; d is the thickness of the hydrophobic insulator.

Electro-wetting has been utilized for a variety of applications, such as displays, optical filters, fiber optics, adaptive lenses, lab-on-a-chip. Among these applications, electro-wetting display (EWD) is one of the most important, which was firstly proposed in 2003 (Hayes and Feenstra, 2003). Fig. 6 shows the principle of a typical EWD (Hayes and Feenstra, 2003; Feenstra and Hayes, 2009; Hayes et al., 2004; Blankenbach, 2008; Burns et al., 2005; Charipar et al., 2015; You and Steckl, 2010). In a pixel, the non-polarized oil containing chromatic dye is dropped on a hydrophobic electrode with an insulating coating, and enclosed by hydrophilic ribs. The oil and electrode are immersed in a polarized water solution. When no voltage is applied, the oil forms a homogeneous film between the water and the insulating coating to acquire the minimum Gibbs free energy, resulting in a colored pixel, as shown in Fig. 6(a). When a voltage is applied across the hydrophobic insulating coating, an electrostatic force is added to the original force equilibrium. To acquire the minimum Gibbs free energy again, the water is moved into contact with the insulating coating; thus, the oil is pushed aside and the white substrate is exposed to form a white pixel, as shown in Fig. 6(b). For a typical pixel size of about 200 μm , the interfacial tension force is more than 1000 times stronger than the gravitational force, thus the displayed images are stable in all orientations. Moreover, the state of a pixel can be continuously tuned between on and off states by setting a certain balance between the electrostatic and surface tension forces. Consequently, different gray levels can be analogously acquired for an EWD. If pulse width modulation (PWM) is adopted, different gray levels can also be digitally realized. Certainly, black or transparent oils can also be utilized in combination with a substrate of different colors, and the working principle is similar except the pixel color.

The switching time between on and off states of an EWD is typically several milliseconds; hence, an EWD is sufficiently fast to show video contents. Current EWDs can go through over a billion times of switches without a degradation of electro-optical performances, revealing a long working life of EWD for video contents. Moreover, under a low operation voltage of around 20 V, the power consumption of an EWD is just about one third of that of conventional transmissive LCDs, leading to an attractive feature of power saving. Due to the reflection of ambient light via the oil or substrate, reflectivity, contrast ratio, viewing angle, sunlight readability etc. of EWDs is comparable to those of regular printed paper. Finally, since the electro-optical performances of

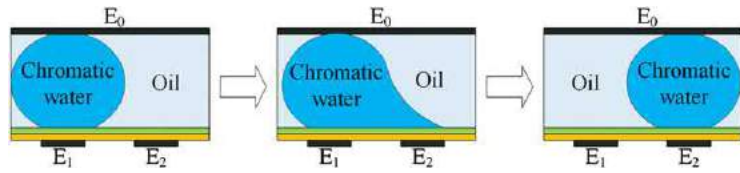


Fig. 7 Principle of a bistable EWD utilizing droplet movement. E_0 is common electrode, and E_1 , E_2 are structured control electrodes, which are applied opposite waveforms in phase while moving the droplet.

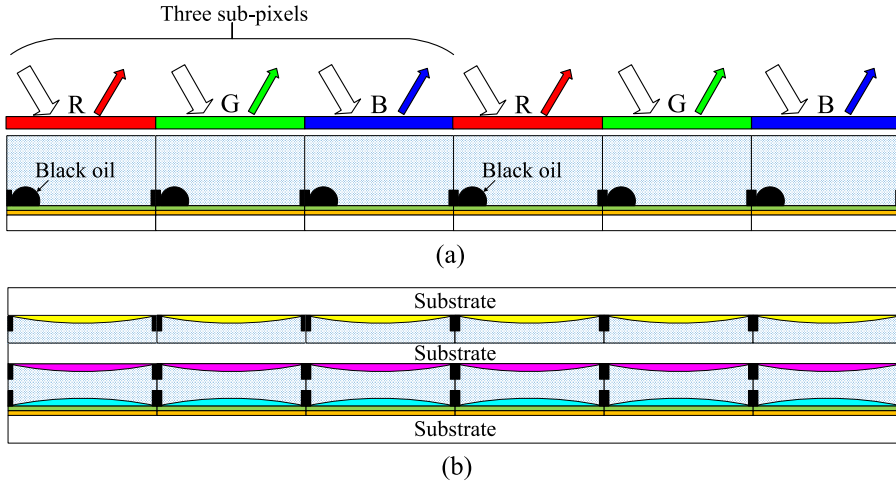


Fig. 8 Principle of a chromatic EWD: (a) adding a color filter onto each pixel containing black oil; (b) stacking three monochrome layers (cyan, magenta and yellow).

an EWD pixel is insensitive to the cell-gap, which is hard to be homogeneous across the screen for a flexible display, the fabrication of flexible EWDs is relatively easy and low-cost (Burns *et al.*, 2005). Furthermore, EWDs and conventional LCDs share around 90% of the production process, leading to a promising prospect of commercialization. However, an EWD utilizing droplet (oil) contracting and relaxing is not bistable, hence it consumes power when still images are shown.

To make EWD adapted for applications that need a bistability, such as e-readers and digital signages, some researchers have realized bistable EWDs (Charipar *et al.*, 2015) by utilizing droplets moving from one position to another, as shown in Fig. 7. Nevertheless, the response time of a bistable EWD is lower than that of a non-bistable one.

EWDs can also be adopted to realize low-cost and power-saving color displays (You and Steckl, 2010). The first approach is simply adding color filters onto each pixel containing black oil to form a chromatic sub-pixel, as shown in Fig. 8(a). As the manufacturing process for traditional LCDs using color filters can be directly utilized here, the cost is considerably low while a higher optical efficiency than that of LCDs can be offered because of no polarizers. The second approach is stacking three monochrome layers that can be separately controlled, as shown in Fig. 8(b). By adjusting the driving voltage of each layer, a desired color can be generated via subtractive color-mixing principle. This triple-layer architecture can directly adopt the processes for single-layer to suppress the cost, in addition to a much higher optical efficiency. Benefiting from the triple-layer structure, the color rendering capacity of such an EWD is much higher than that of other reflective displays. Furthermore, due to the subtractive color-mixing principle, the color gamut and brightness have a win-win relationship for a triple-layer EWD, while a trade-off needs to be endured for a conventional LCD adopting additive colors.

Other Reflective Display Technologies

In addition to EPD and EWD, there are several other reflective display technologies, such as quick response liquid powder display (QR-LPD) (Hattori *et al.*, 2004), cholesteric liquid crystal display (ChLCD) (Tamaoki, 2001), interferometric modulator display (IMOD, trademarked mirasol[®] by Qualcomm) (Bita *et al.*, 2012), memory LCD (Tamaki *et al.*, 2014). Although they have not been widely commercialized, several attractive features can be found in these technologies. Here their principles and features are briefly summarized in Tables 1 and 2.

- (1) QR-LPD[®]: A QR-LPD[®] is based on electrophoresis, where high-fluidity black and white “liquid powders” at nanometer scale are charged and suspended in the air to be driven by two electrodes sandwiching the powders. When a negative voltage is applied to the upper transparent electrode, the black powders with positive charges aggregate on the top, and then bring

Table 1 Principles and features of several reflective display technologies (1)

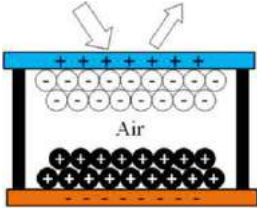
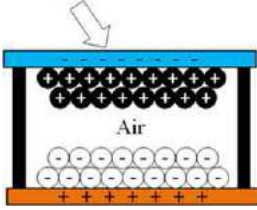
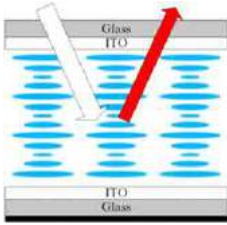
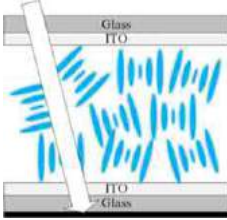
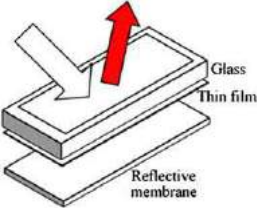
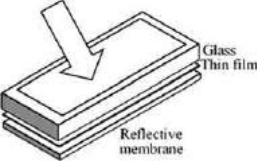
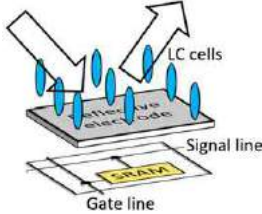
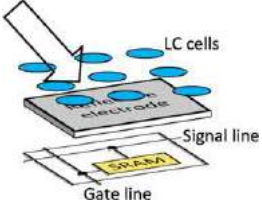
Technology	QR-LPD	ChLCD
Principle	White state  Black state 	 
Advantages	Fast response (<1 ms), low temperature driving, wide viewing angle, power saving due to bistability	No use of color filter or polarizer, wide color gamut, power saving due to bistability
Disadvantages	High driving voltage, low contrast ratio	Slow response (~ 300 ms), low resolution, high cost, high driving voltage
Leading companies	Bridgestone	Fujitsu, Fujitsu Xerox, Kent Displays, ITRI

Table 2 Principles and features of several reflective display technologies (2)

Technology	IMOD	Memory LCD
Principle	White state  Black state 	 
Advantages	Fast response (~ 10 μ s), power saving due to bistability, high contrast ratio, high reflectivity ($>60\%$), high robustness	Low power consumption due to bistability
Disadvantages	Low resolution of gray scale, high cost	Limited resolution and color rendering capacity due to the complication of SRAM circuit integration
Leading companies	Qualcomm	Sharp

about a black appearance, while a positive voltage applied to the upper electrode produces a white appearance in the opposite case. Reflection by the white powders makes a QR-LPD to have a paper-like appearance and wide viewing angle. Due to the high fluidity of the liquid powders induced by their surface morphology and electronic property, response time of

- a QR-LPD is ultra-fast, as shorter than 1 ms. Additionally, the electric field forces between the electrodes and liquid powders give a QR-LPD a bistability feature, resulting in ultra-low power consumption. By optimizing the charge amounts for liquid powders of different colors, a colorful QR-LPD can be realized.
- (2) ChLCD: A cholesteric liquid crystal with chirality caused by a helical structure is contained in a ChLCD. The liquid crystal is arranged in layers with no positional ordering within layers, while director axes of the layers vary periodically, leading to Bragg reflection to the incoming light. When the incident wavelength is between $n_o p$ and $n_e p$, where n_o is the ordinary refractive index, n_e is the extraordinary refractive index, and p is the cholesteric pitch over which the cholesteric liquid crystal molecules fully rotate 360° along their helical, a certain light wavelength is selectively reflected. In this way, it is possible to realize vivid color images without color filters and polarizers. On the other hand, a focal conic domain configuration makes the light pass through and absorbed by an absorption layer beneath to make a dark state. A ChLCD is also bistable (planar state and focal conic state), thus little power is consumed to maintain a still image. To achieve a full-color ChLCD, a stacked-panel technology can be adopted.
 - (3) IMOD: Vivid colors on a butterfly's wings or a peacock's feathers, which are generated by light interference, are imitated. In an IMOD, each pixel is actually a micro-electromechanical system (MEMS) device that contains an air gap sandwiched by a thin-film-coated glass substrate on the top and a reflective conductive membrane at the bottom. The thickness of the air gap at micrometer scale can be tuned by applying a voltage on the top glass substrate and bottom membrane. In this way, a constructive or destructive interference occurs; and then the incoming light with a certain wavelength can be selectively reflected or collapsed. To build a full-color pixel, grids of red, green and blue color subpixels are combined and gray levels are produced by using spatial dithering. An IMOD calls for very little power to maintain either one of its two states to realize a bistable display. Moreover, the interference light from the ambience leads to a great sunlight readability.
 - (4) Memory LCD: Memory LCD, as known as memory-in-pixel (MIP) reflective LCD, adopts reflective electrodes under LC cells and integrates a static random access memory (SRAM) circuit and LC driver into each subpixel (under the reflective electrodes) using conventional TFT technologies. When a still image is displayed, benefiting from the SRAM circuit, only the pixel circuit is driven, leading to no charging current to the data line and a very small amount of charging current to the pixel capacitor. Therefore, a memory LCD is bistable-like and the power consumption of it is ultra-low. Nevertheless, due to the considerable complication of integrating a memory circuit into each subpixel, the resolution of a memory LCD is highly limited. In addition, the bit number of the in-pixel memory is also limited, as only one or two bit; thus, the color rendering capacity is also low in comparison with other reflective display technologies.

References

- Amundson, K., Ewing, J., Kazlas, P., *et al.*, 2001. Flexible, active-matrix display constructed using a microencapsulated electrophoretic material and an organic-semiconductor-based backplane. *SID Symposium Digest of Technical Papers*, vol. 32(1), pp. 160–63.
- Bitá, I., Govil, A., Gusev, E., 2012. Mirasol[®]—MEMS-based direct view reflective display technology. In *Handbook of Visual Display Technology*. Berlin/Heidelberg: Springer.
- Blankenbach, K., Schmoll, A., Bitman, A., Bartels, F., Jerosch, D., 2008. Novel highly reflective and bistable electrowetting displays. *Journal of the Society for Information Display* 16 (2), 237–244.
- Burns S.E., Reynolds, K., Reeves, W., *et al.*, 2005. Flexible active-matrix displays. *SID Symposium Digest of Technical Papers*, vol. 36 (1), 19–21.
- Charipar, K.M., Charipar, N.A., Bellemare, J.V., Peak, J.E., Piqué, A., 2015. Electrowetting displays utilizing bistable, multi-color pixels via laser processing. *Journal of Display Technology* 11 (2), 175–182.
- Comiskey, B., Albert, J.D., Yoshizawa, H., Jacobson, J., 1998. An electrophoretic ink for all-printed reflective electronic displays. *Nature* 394 (6690), 253–255.
- DisplaySearch, 2016. E-paper Display Revenues to Reach US\$9.6B by 2018. [https://www.thefreelibrary.com/E-paper+display+revenues+to+reach+US\\$249.6B+by+2018%3a+DisplaySearch.-a0208133110](https://www.thefreelibrary.com/E-paper+display+revenues+to+reach+US$249.6B+by+2018%3a+DisplaySearch.-a0208133110).
- Dukhin, S.S., Derjaguin, B.V., 1974. *Electrokinetic Phenomena*. New York: J. Wiley and Sons.
- Feenstra J. and Hayes R., 2009. Electro-wetting Displays, Liguavista White Paper. <http://www.liquavista.com/media/772/LQV0905291LL5-15.pdf>.
- Hattori, R., Yamada, S., Masuda, Y., Nihei, N., 2004. A novel bistable reflective display using quick-response liquid powder. *Journal of the Society for Information Display* 12 (1), 75–80.
- Hayes R.A., Feenstra, B.J., Camps, I.G.J., *et al.*, 2004. A high brightness colour 160 PPI reflective display technology based on electrowetting. *SID Symposium Digest of Technical Papers*, vol. 35 (1), pp. 1412–1415.
- Hayes, R.A., Feenstra, B.J., 2003. Video-speed electronic paper based on electro-wetting. *Nature* 425 (6956), 383–385.
- Huang, Y.-P., Lin, F.-C., Wu, S.-L., Yang, P.-R., 2013. Mechanism and improvement of charged-particles transition in microcup electrophoretic displays. *Journal of Display Technology* 9 (8), 619–625.
- Inoue, S., Kawai, H., Kanbe, S., Saeki, T., Shimoda, T., 2002. High-resolution microencapsulated electrophoretic display (EPD) driven by poly-Si TFTs with four-level grayscale. *IEEE Transactions on Electron Devices* 49 (9), 1532–1539.
- Kim, M.K., Am Kim, C., Ahn, S.D., Kang, S.R., Suh, K.S., 2004. Density compatibility of encapsulation of white inorganic TiO₂ particles using dispersion polymerization technique for electrophoretic display. *Synthetic Metals* 146 (2), 197–199.
- Lee, H., Liang, R.C., 2003. SiPix Microcup[®] electronic paper – an introduction. *Advanced Display* 3, 4–9.
- Li, G.X., Qin, S.C., Feng, Y.Q., Fang, S., Meng, S.X., 2016. Preparation and characterization of microcapsule-encapsulated colored electrophoretic fluid in trifluorotoluene system for electrophoretic display. *Journal of Display Technology* 12 (10), 1145–1151.
- Liang, R.C., Hou, J., Zang, H., Chung, J., Tseng, S., 2003. Microcup[®] displays: electronic paper by roll-to-roll manufacturing processes. *Journal of the Society for Information Display* 11 (4), 621–628.
- Lippmann, M.G., 1875. Relation entre les phénomènes électriques et capillaires. *Annales de Chimie et de Physique* 5, 494.
- Luxresearch, 2015. Picking the win, place and show in the emerging display technologies horserace. <http://blog.luxresearchinc.com/blog/tag/reflective-displays/>.
- Oh, S.W., Kim, C.W., Cha, H.J., Pal, U., Kang, Y.S., 2009. Encapsulated-dye all-organic charged colored ink nanoparticles for electrophoretic image display. *Advanced Materials* 21 (48), 4987–4991.

- Tamaki, M., Fukunaga, Y., Mitsui, M., *et al.*, 2014. A memory-in-pixel reflective-type LCD using newly designed system and pixel structure. *Journal of the Society for Information Display* 22 (5), 251–259.
- Tamaoki, N., 2001. Cholesteric liquid crystals for color information technology. *Advanced Materials* 13 (15), 1135–1147.
- Wu, S.-T., Yang, D.-K., 2001. *Reflective Liquid Crystal Displays*. New York: Wiley.
- You, H., Steckl, A.J., 2010. Three-color electro-wetting display device for electronic paper. *Applied Physics Letters* 97 (2), 023514.

In Situ Optical Tissue Diagnostics/Miniaturized Optoelectronic Sensors for Tissue Diagnostics

Seung Yup Lee, Georgia Institute of Technology/Emory University, Atlanta, GA, United States
Yooree Grace Chung and Mary-Ann Mycek, University of Michigan, Ann Arbor, MI, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Optical sensing technology has been widely employed in a variety of biomedical applications due to its noninvasive and nonionizing nature (Mycek and Pogue, 2003). The core components comprising the optical sensor are a light source and a detector. The light source excites a sample with different intensities, wavelengths, time-scales, and coherences of light suited to the biophysical characteristics being investigated. The detector receives the re-emitted light that has interacted with tissues and typically converts a photon signal into an electrical signal to be read by an electronic recording system. Conventional light sources include various types of lasers and lamps to generate photons. Light detecting devices commonly used in tissue optical sensing include photomultiplier tubes (PMTs), silicon photodiodes (PDs), and image sensors such as complementary metal-oxide-semiconductor (CMOS) and charged-couple device (CCD).

Typically, these optical components are bulky, resulting in a large footprint of the final instrument that integrates all the optical and electronic hardware units (Pitts and Mycek, 2001; Lloyd *et al.*, 2010). The large size of these photonic medical devices can limit their mobility, requiring patients to visit the clinical location of the device. Further, if an optical fiber is employed to deliver light to tissue, disadvantages could include inefficient coupling, limited bend radius, the possibility of breakage and low spatial information.

During the past couple of decades, advances in techniques for handling new semiconductive materials, micro-/nano-fabrication technologies, and tiny electronics with high computing power have revolutionized the engineering of miniaturized optoelectronic sensors. Compared to fiber-based optical sensors, miniaturized optoelectronic sensors offer substantial benefits. Dramatic reductions in size, cost and power consumption have great potential to make light-based sensing more ubiquitous than ever. Beyond implementing point-of-care diagnostics, these sensors can be wearable, enabling continuous, noninvasive and object monitoring anytime and anywhere. The end terminal of this “fiber-less” sensor is an electrical signal that can be wirelessly transmitted, thereby providing subject’s mobility with tremendous flexibility. Miniaturized sensors also enable optical interrogation in human organs where traditional fiber-optic-based approaches fulfill very limited tasks.

In this article, we will review the basic construction of miniaturized sensors, major optoelectronic components and assembly techniques, data processing algorithms and their tissue diagnostics applications.

Optoelectronic Sensing System

Fig. 1 illustrates a block diagram of the miniaturized optoelectronic sensing system. The system can be divided into two parts. One is a sensing unit comprised of optoelectronic modules that send and receive photons to and from biological tissues, in which the conversion process occurs between photonic and electrical signal. The other is a control electronics unit that controls the excitation and reading sequence via an electrical signal sent or received by the sensing unit. Depending on the sensor design and application, only the sensing unit can be constructed in a miniaturized form (under a few millimeter scale), communicating with relatively large control electronics located on a remote site via physical wiring. The detailed technique for sensor unit fabrication is discussed in a later section. Alternatively, the two parts can be integrated in one platform utilizing the CMOS microfabrication technique of electrical circuitry or direct use of minimal electronic bare chips. This integrated optoelectronic sensor can independently work as a separate sensor node.

The sensing unit consists of an optoelectronics source, an optoelectronics detector and micro-optic components such as micro-lenses or color filters. The optoelectronic source is powered to emit light at a predetermined intensity given voltage and current supplied by the control electronics. The use of a micro-lens on the source’s emission aperture adjusts the divergence of the beam. The incident photon interacts with a variety of micro-organisms, and its properties are altered (e.g., intensity – attenuated by natural chromophores such as hemoglobin, wavelength – fluorescence molecules) as it propagates through the tissue. The optoelectronic detector captures remitted light exiting the tissues, which contains information about tissue morphology, physiology or metabolism. Before light enters the detector, a micro-lens can determine an acceptance angle, and a micro-fabricated filter selects the wavelength range. We will not discuss the types and technologies of the micro-lens and optical filters as they are beyond the scope of this article.

Optical contrasts used by optoelectronic sensors are classified into two groups: exogenous and endogenous. Exogenous agents such as fluorescence protein, oxygen-sensitive dye and gold nanoparticle can be inserted or labeled with a specific target molecule to enhance the signal. For label-free sensing, the optoelectronic sensor can employ endogenous contrasts. In reflectance mode, scattering from micro-organelles such as nuclei and mitochondria, and absorption by natural chromophores such oxy- and deoxy-

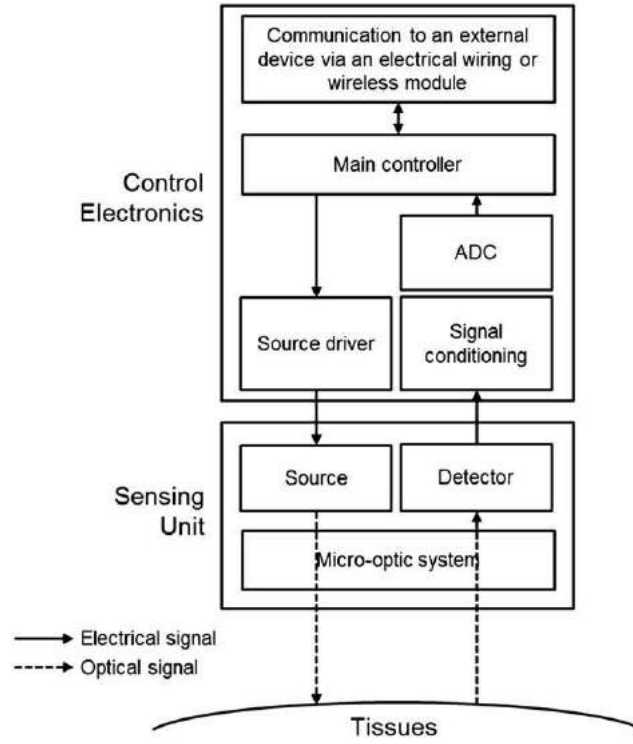


Fig. 1 Schematic diagram of an optoelectronic sensor for tissue diagnostics. ADC: analog-to-digital converter.

hemoglobin, bilirubin and beta-carotene can alter the detected reflectance. In fluorescence mode, metabolic cofactors such as nicotinamide adenine dinucleotide (phosphate) (NAD(P)H) and flavin adenine dinucleotide (FAD) or extracellular matrix (e.g., collagen) can serve as intrinsic fluorescence sources (Lloyd *et al.*, 2013).

In general, there are three different types of techniques in optical sensing relating to domain: continuous, frequency domain and time domain. Most current miniaturized optoelectronic sensors are based on continuous sensing (only light intensity measurements) due to the technical complexity in miniaturization of necessary components for frequency- and time-domain sensing. Thus, we premise the continuous sensing technique when we discuss design, data analysis and applications of the optoelectronic sensors.

Designing optoelectronic sensors for tissue diagnostics considers numerous engineering factors including electrical, optical, thermal, and mechanical aspects. The signal-to-noise ratio (SNR) is one of the most important metrics to achieve the basic function of the designed sensor. SNR is defined as

$$\text{SNR(dB)} = 20 \log \frac{\mu}{\sigma}. \quad (1)$$

Here, μ is the mean and σ is the standard deviation of the repetitive measurements on a sample. Typically, tissue-simulating phantoms mimicking optical properties of targeted biological tissues are employed to obtain the SNR of the sensor. SNR is determined by various elements such as source intensity, detector sensitivity, source–detection separation and electronics design (source driver and signal conditioning circuit). The maximum of excitation intensity is ideally desirable for highest SNR, but it needs to be carefully selected to prevent thermal damage of interrogated tissues. In applications requiring sensor implantation or direct contact with human tissues, a biocompatible encapsulation of the sensor is indispensable to avoid tissue compromise via physical and chemical isolation. At the same time, this encapsulation protects the sensor against the watery environment of tissues.

Thermal stability is another critical factor to be considered in optoelectronic sensor design for two main reasons. Because the operation of the optoelectronic components is highly sensitive to temperature, the optical performance of the designed sensor could vary without careful design of heat sink or transfer. In a miniaturized form, a heat sink might not be a feasible option. Ideal design would have negligible or stabilized temperature increase from heat dissipation of source operation, even during long-term use. Optical and thermal simulation tools are useful for evaluating the initial feasibility of the design.

Optoelectronic Components

Laser diode (LD)

A laser diode (LD), also known as an injection diode laser, is a forward-biased semiconductor diode that emits coherent light when electrons and holes are stimulated by an electrical current into the p–n junction (Fig. 2(a)). LDs consist of two stacked layers of

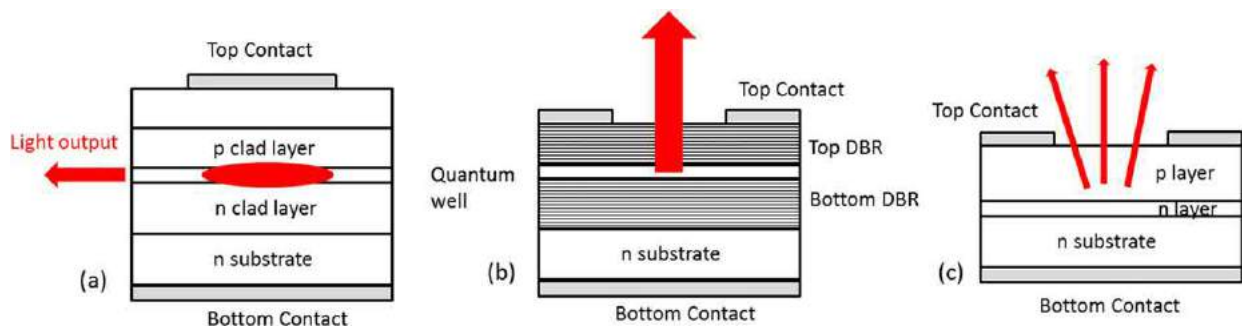


Fig. 2 Illustration of basic structure of (a) laser diode (LD), (b) vertical-cavity surface-emitting laser (VCSEL), and (c) light-emitting diode (LED).

semiconductors as the optical gain medium. The top layer is a p-doped semiconductor, and the bottom layer is an n-doped semiconductor. The layer in between the two semiconductors is the p–n junction, known as the intrinsic region. The intrinsic region is the resonant optical cavity of the laser. When a current is passed through the semiconductors, electrons from the n-type semiconductor and holes from the p-type semiconductor flow into the intrinsic region. Recombination of an electron and a hole releases energy in the form of a photon of light. This photon remains in the intrinsic region and stimulates further release of electrons and holes. The newly formed photons will then be in the same phase, direction and polarization, thereby producing coherent light. Light emerges from one end of the diode and diffracts but can be formed into a straight beam with the use of a collimating lens. Because laser beams are concentrated to a small point, LDs are useful for targeted medical applications that require greater penetration.

Vertical-cavity surface-emitting laser (VCSEL)

A vertical-cavity surface-emitting laser (VCSEL) is a forward-biased semiconductor diode that emits light perpendicular to the laser's top surface as compared to a conventional edge-emitting laser that emits light in the parallel direction (**Fig. 2(b)**). A VCSEL consists of three main layers: the upper p-type distributed Bragg reflector (DBR), the active region with quantum wells, and the lower n-type DBR. The two DBR layers are highly reflective mirrors with a reflectivity of 99.5–99.9%. Photons formed in the active region oscillate between the two DBR layers until they escape as coherent light perpendicularly from the top of the laser. The perpendicular light beam enables three times as many VCSELs to be produced from the same size wafer as compared to edge-emitting lasers. The perpendicular beam also allows VCSELs to be incorporated into two-dimensional arrays to maximize output power and device reliability. Other advantages of VCSELs over edge-emitting lasers include easy coupling with optical fibers due to circular beams, less power consumption, greater quality control during production, and lower production costs.

Light emitting diode (LED)

A light-emitting diode (LED) is a forward-biased semiconductor diode that produces incoherent light. An LED includes a p–n junction where electrons and holes generated from an electric current are recombined (**Fig. 2(c)**). Energy is released as light and heat. The mechanism of light generation is similar to LDs; however, LEDs do not have stimulated emission or coherent light emission. Because LEDs distribute their power incoherently, they are useful for treatment of larger areas of shallow depth. Low power consumption and heat generation associated with LEDs result in improved device efficiency. LEDs are composed of inorganic semiconductors, commonly gallium arsenide phosphide, with long life times. Failure occurs as a gradual, rather than abrupt, decrease in brightness. The durability, compact size, and low production costs of LEDs make them well suited for medical applications. Unlike LDs that emit a coherent light with a narrow bandwidth (~ 1 nm) in a small divergence angle, LEDs emit an incoherent light with a broader bandwidth (typically 25–50 nm) in a larger divergence angle. LEDs come in smaller packages and have more flexibility in available wavelengths than LDs.

Silicone photodiode

A photodiode (PD) is an optical signal receiver. PDs function as semiconductor diodes that detect light with the use of a p–n junction. Incident light is focused onto the junction of the PD. As light enters the charge-depleted region between the p-type and n-type semiconductors and strikes an atom within the crystal structure of the PD, electron–hole pairs are generated. The built-in electric field of the charge-depleted region causes the electrons to drift toward the cathode and the holes to drift toward the anode, creating a photocurrent. The greater the intensity of incident light, the more electron–hole pairs are created. Thus, the generated photocurrent is dependent on light intensity. The total current of a PD includes dark currents, which must be accounted for in calibration. There are different modes of PD operation: the photovoltaic mode, the photoconductive mode, and avalanche PDs. The photoconductive mode creates voltage by restricting the photocurrent flow. In the photoconductive mode, the PD is reverse biased, which increases the width of the charge-depleted region. This width increase decreases the capacitance of the p–n junction, reduces the response time of the PD and increases the linearity of light intensity measurements. However, the photoconductive mode also increases the amount of dark current and noise. Avalanche PDs are highly reverse biased, multiplying the current, and increasing the response of the PD.

Silicone photomultiplier (SiPM)

A silicon photomultiplier (SiPM) is an avalanche PD array joined together on a silicon substrate. Silicon can detect light in the wavelength range 350–800 nm. The array is composed of PD and quenching resistor microcells arranged in parallel. SiPMs have single-photon sensitivity because each individual microcell can detect an incident photon. When a photon is detected, the microcell avalanches as the silicon becomes conductive and amplifies the current (Geiger discharge process). The quenching resistors then stop the avalanche, and the microcells are reset to detect another photon. The released current from all of the microcells can be measured to detect the intensity of light. Therefore, the number of microcells activated is proportional to the amount of incident light. SiPMs are advantageous due to their sensitivity, low operating voltage, and high timing performance. Because of their single photon sensitivity, SiPMs are useful in low-light applications. Currently, SiPMs are being considered for use in positron emission tomography and single-photon emission computed tomography.

Organic LED and PD (OLED and OPD)

Organic LEDs and PDs are composed of organic semiconductors and polymers. Organic polymers allow for greater flexibility and transparent displays. OLEDs and OPDs have several advantages over inorganic LEDs and PDs. Inorganic-based optoelectronics are difficult to scale for large-area production. Organic optoelectronics, however, can be manufactured with flexible substrates and with solution processing, resulting in cheaper and larger volume production. Inorganic components also tend to be more rigid and bulkier. Flexible organic polymers allow organic parts to be used in a wider variety of medical applications. The flexibility of OLEDs and OPDs and their large scalability also introduce the potential for low-cost, disposable medical sensors (Lochner *et al.*, 2014).

Assembly Techniques

Small footprint components soldered on conventional print circuit board (PCB)

The most basic method of constructing an optoelectronic sensor is assembling commercially available packaged components onto a conventional printed circuit board (PCB). Most optoelectronic-based medical devices that have become a standard of care in clinics fall into this category in a hand-held form, allowing for point-of-care diagnostics. The pulse oximeter is the most successful device. The bilirubinometer, which screens for neonatal jaundice by measuring subcutaneous serum bilirubin concentration, is another good example. Their operating principles will be discussed in detail later in the application section. The smallest size of packaged LEDs and PDs currently available is in the range of ~ 1 mm. These components can be manually or automatically soldered onto the PCB, but are not feasible for applications requiring further miniaturization.

Die-level chip assembled on fabricated board with die-attached material

An alternative to using packaged optoelectronic components is assembling unpackaged die-level chips onto micro-fabricated circuit boards for the applications smaller than 1 mm (Fig. 3(a)). The direct soldering of chips is not feasible due to densely patterned interconnections (a few tens of micrometers between patterned lines) or no solder contact metal on the chips. For chips without solder metals on their anode/cathode electrode, a special metal alloy needs to be patterned in the process of fabricating the board. One bonding option utilizes a eutectic alloy like indium–gold (In–Au), which means the melting temperature of the alloy is lower than that of each metal. One advantage of eutectic bonding, in terms of assembly, is that the eutectic-bonded components are not re-melted by another heat process when even higher temperatures are applied (Fig. 4(a)). For the top electrode, wire-bonding is required to make a connection to the bottom interconnections. Some LED manufacturers supply directly attached LED chips of which cathode and anode are on the same side as solder metals. A simple reflow process bonds both electrodes, thus requiring no wire-bonding of the top electrode. Temperature profiles in each heating process need to be carefully controlled because non-optimized processing of temperatures can deteriorate the performance of the optoelectronic components as shown in Fig. 4(b). Even a bare die of an electronic chip can be directly attached to the flexible substrate board (Kim *et al.*, 2017).

Monolithic fabrication of integrated sensor

Maximum miniaturization of a sensor is achieved by monolithically fabricating an integrated sensor (O'Sullivan *et al.*, 2010). The source and detector are fabricated together using fabrication process flow and patterned by photolithography techniques from a single wafer (Fig. 3(b)). The VCSEL is a suitable semiconductor light source for this type of the integrated sensors because it emits light upward with higher intensity and requires simpler fabrication design compared to LED. The fabrication order requires careful consideration since the performance of each source and detector can be critically affected. For high-quality VCSEL lasing, the VCSEL layer needs to be grown before the PD. One advantage of monolithic fabrication is that source–detector separation can be sophisticatedly determined on the order of 100 μm . For a very short distance, the additional structure blocking a direct light pipelining from source to detector is necessary. An interference filter or absorbing-based filter can be used as a blocker and deposition techniques are compatible with the general semiconductor fabrication process of source/detector.

Data Processing Algorithm for Optoelectronic Sensor

Using physical principles on light–tissue interaction, there are various analysis algorithms depending on specific applications. This section will discuss two methods for analyzing diffuse reflectance: (1) the ratiometric method to measure a simple reflectance ratio

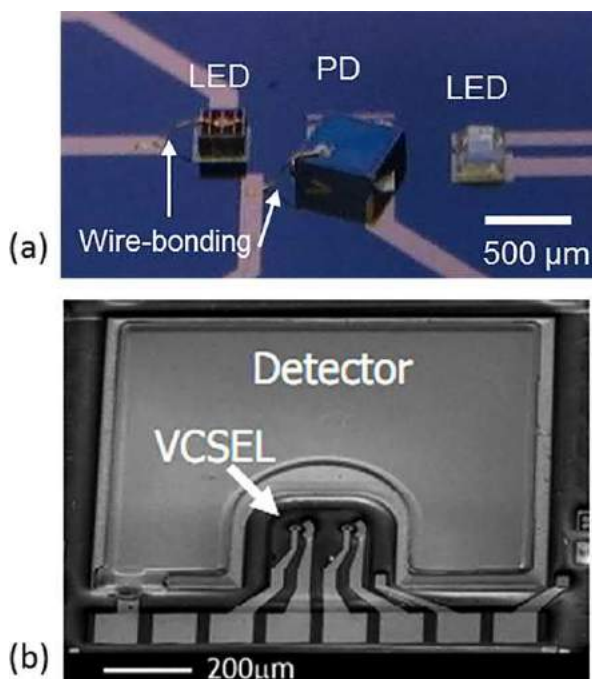


Fig. 3 (a) Die-level chips assembled on the fabricated micro-printed circuit board patterned with solder materials for a needle-compatible optoelectronic sensor assessing reflectance ratios. (b) Monolithically integrated fluorescence sensor. From O'Sullivan, T., Munro, E.A., Parashurama, N., *et al.*, 2010. Implantable semiconductor biosensor for continuous in vivo sensing of far-red fluorescent molecules. *Optics Express* 18, 12513–12525.

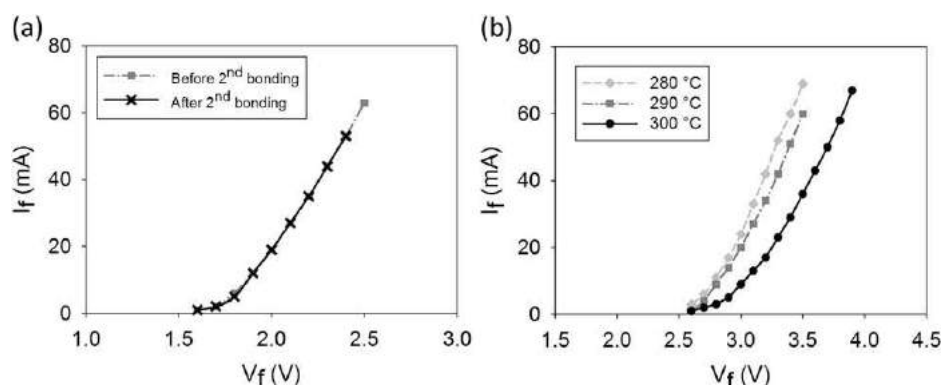


Fig. 4 (a) Forward voltage–forward current characteristics of the eutectic-bonded LED before and after the following heat process for 2nd LED bonding. (b) Forward voltage–forward current characteristics of the bonded die-level LEDs with different temperatures.

between two discrete wavelengths (one is a contrast and the other is a reference), and (2) the quasi-spectral method that uses multiple wavelengths to mimic a continuous spectrum.

Ratiometric analysis

Intensity-based sensing at single wavelength would be the simplest way to analyze diffuse reflectance but requires a rigorous calibration process to address intensity-induced distortions of such varying source intensity. The ratiometric method avoids this process by using one measurement as reference reflectance intensity. Calculation of the ratio in regard to the reference accounts for measurement errors. For example, in human pancreatic tissue assessment, the reflectance ratio, $R_{470 \text{ nm}}/R_{650 \text{ nm}}$ is suitable to distinguish malignant tissues (adenocarcinoma) from non-malignant tissues, including normal tissues and chronic pancreatitis. In a previous optical study on freshly excised human pancreatic tissues, the reflectance intensities at 450–470 nm differed significantly among these tissue types. This was attributed to different scattering properties depending on nuclei size, refractive index, and collagen contents. On the other hand, the reflectance intensities at 630–650 nm in three tissue types were very close, which served as a reference (Chandra *et al.*, 2007).

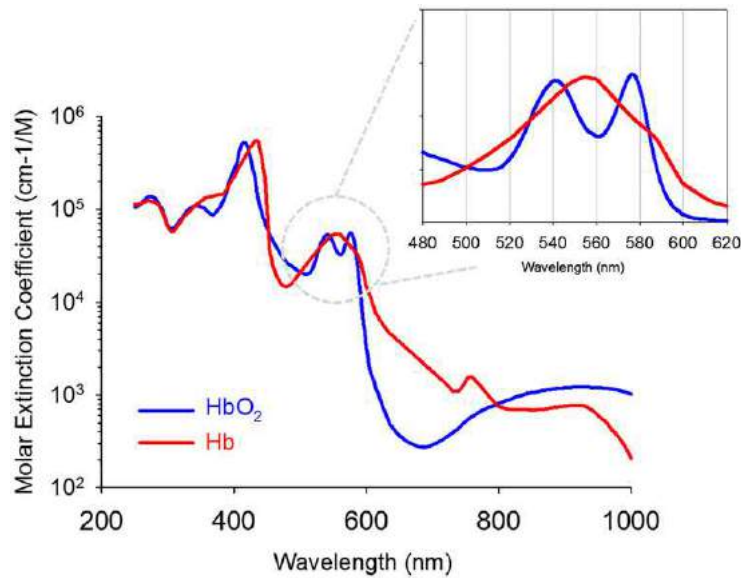


Fig. 5 Wavelength-dependent molar extinction coefficients of oxy- (HbO_2) and deoxy- (Hb) hemoglobin. Inset shows isosbestic points.

More generally, employing different wavelengths related to hemoglobin absorption has potential to assess total hemoglobin (Hb) concentration and tissue oxygenation (Hu *et al.*, 2013). Previous studies demonstrated that the ratiometric analysis based on isosbestic and non-isosbestic points has the capability to estimate tissue hemoglobin concentration and oxygenation. Isosbestic points have the same molar extinction for oxyhemoglobin and deoxyhemoglobin wavelengths, for example, 500, 529, 545, 570 and 584 nm (Fig. 5). Non-isosbestic points are the wavelengths showing the biggest difference in absorbance between oxy- and deoxy-hemoglobin, for example, 516, 539, 560, 577 and 593 nm. However, this study employed a full reflectance spectrum acquired by a fiber-based diffuse reflectance spectroscopy. For the optoelectronic sensors, two LEDs with wavelengths as close to isosbestic and non-isosbestic wavelengths as possible can be selected to enable rapid estimation of blood parameters. Rapid and non-invasive assessment of Hb concentration and oxygenation will have many clinical applications, including cancer detection, anemia detection, perfusion monitoring, etc. The main advantage of the ratiometric analysis is the rapid processing time that potentially enables real-time monitoring.

Quasi-spectral analysis

Unlike conventional optical spectroscopy that takes advantage of a continuous spectrum for evaluating tissues, the number of wavelengths available to optoelectronic sensors is limited due to technical difficulties in manufacturing a miniaturized light-dispersing component. However, multiple monochromatic sources with different wavelengths associated with one detector or, alternatively, a broadband source coupled with multispectral bandpass filter installed onto the detector, make the quasi-spectral analysis feasible. Even with a limited number of wavelengths, traditionally employed analysis algorithms utilizing a full spectrum can be applied to extract biophysically relevant parameters or optical properties directly from sensor measurements.

Analytical models have been employed to analyze diffuse reflectance and fluorescence spectrum obtained from human biological tissues (Wilson and Mycek, 2011). The mathematical equation is based on fundamental principles for light scattering and absorption at a given fiber optics probe geometry (Wilson *et al.*, 2010). A reflectance spectrum is calculated from many input variables including nuclei size and refractive index (Lee *et al.*, 2014). By fitting this mathematical model to the obtained reflectance spectrum, the best parameters generating the least error between the model and the experiment are extracted. The original model uses a wide range of wavelengths in the reflectance spectrum from 400 to 750 nm with 2 nm step for a total of 176 wavelengths. However, using only four discrete wavelengths (460, 520, 560, and 630 nm) can also produce results comparable to the model with 176 wavelengths. The reconstructed model fits from two models applied to the same spectrum obtained from human pancreatic tissues are comparable (Fig. 6(a)). The extracted parameters (scattering amplitude) are also similar in each pancreatic tissue type (Fig. 6(b)).

This preliminary comparison demonstrates the possibility of quasi-spectral sensing using the optoelectronic microprobe for noninvasive assessment of biophysically relevant parameters.

Applications

Clinical standard of care

The pulse oximeter is the most representative optoelectronic sensing device and has become a standard of care in clinics. Arterial oxygen saturation (SpO_2) and pulse rate are noninvasively measured (Chan *et al.*, 2013). Consisting of a small LED and PD, the

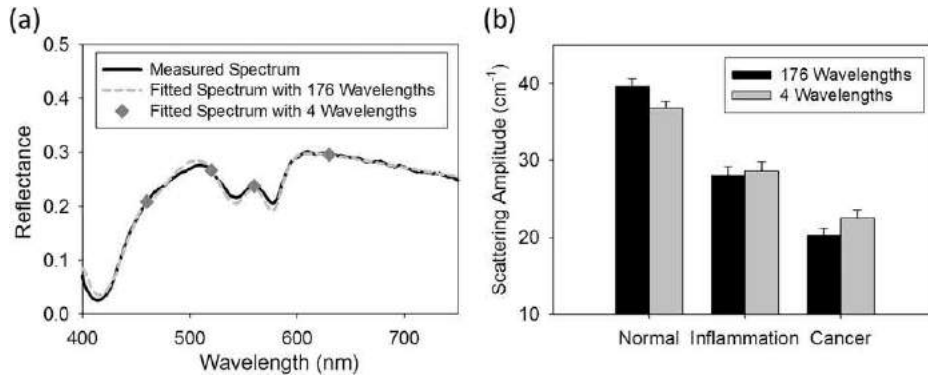


Fig. 6 (a) Reconstructed spectra by fitting procedure using 176 and 4 wavelengths on the measured spectrum. (b) Comparison of the extracted parameters (scattering amplitude) per different tissue types by inversion algorithm using 176 vs. 4 wavelengths.

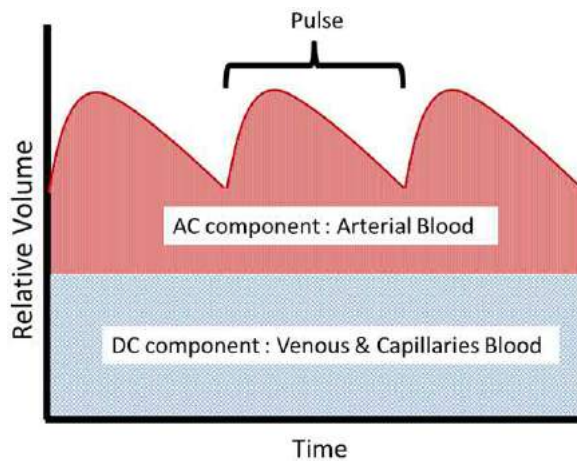


Fig. 7 Relative volume of the non-pulsatile (DC) and pulsatile (AC) tissue compartment during a cardiac cycle.

device can be placed on the tip of a patient's finger or ear lobe, allowing for continuous monitoring. Pulse oximetry takes advantage of red and near-infrared (IR) light that can penetrate tissue and be absorbed by hemoglobin. A red LED-emitting light at 660 nm and a near-IR LED with a wavelength of 940 nm are placed on one side of the device. As the light is transmitted through tissue, oxygenated hemoglobin absorbs IR light more than red light, and deoxygenated hemoglobin absorbs red light more than IR light. A PD located on the opposite side of the tissue detects transmitted light. The ratio of transmitted red and IR light is used to determine oxygen saturation of hemoglobin. Pulse rate can also be measured by the transmitted light. A consistent amount of light is absorbed by the tissue and steady venous flow, resulting in a direct current (DC) component of the signal. Pulsatile arterial blood flow creates an alternating current (AC) signal (Fig. 7). The modulation ratio (R) is calculated by

$$\text{Modulation Ratio } (R) = \frac{A_{\text{red-AC}}/A_{\text{red-DC}}}{A_{\text{IR-AC}}/A_{\text{IR-DC}}} \quad (2)$$

where A is the absorbance. This R is typically inversely proportional to SpO_2 and determines absolute SpO_2 values based on a calibration curve empirically obtained from controlled measurements.

The bilirubinometer is another exemplary device that is routinely used to screen neonatal jaundice in preterm infants. Jaundice is monitored by serum bilirubin levels, typically requiring neonates to endure multiple painful blood drawings. However, because cutaneous bilirubin levels are highly correlated with serum levels, transcutaneous bilirubinometers provide a noninvasive method of determining serum bilirubin levels by measuring the yellowness of the skin. Measurements are made using diffuse reflectance spectroscopy. Placing the device on the neonate's forehead or upper sternum activates a pressure-sensitive probe. A light-generating tube transmits light through subcutaneous tissue. Fiber-optic filaments send resultant scattered light to a spectrophotometric module. Light is then split into different spectra between 460 and 540 nm (different devices vary on the number of wavelengths). The absorbance at different wavelengths is compared to known absorption coefficients for all skin chromophores. The amount of yellowness in the skin is calculated as an indirect measurement of bilirubin values.

Cancer diagnostics

Not many optoelectronic sensors have been introduced to investigate human cancerous tissues because there is not a significant need for miniaturization in clinical diagnostics of human cancers. Over the last three decades, optical spectroscopy has demonstrated its potential for noninvasive detection of cancer in various human organs with a sub-cellular level sensitivity evaluating morphological, physiological biochemical changes upon tumor progression. However, only a few attempts to build optical sensing systems with optoelectronic components for either source or detector have been reported. Most of them were to implement sensor arrays for breast cancer imaging. Recently, the needs for miniaturized sensors have emerged to be compatible with an endoscopic channel (2–5 mm) or even with a hollow needle (< 1 mm) targeting some solid tumors such as pancreas, liver, and lung that are accessible only via a hollow needle. Some studies have delivered a fiber optics through this hollow-needle channel to optically interrogate lung and pancreas. However, the miniaturized optoelectronic sensor fitting in this needle could allow for arrangement of multiple source–detector units through a side-view window, which would increase an optical sampling volume (Lee *et al.*, 2017).

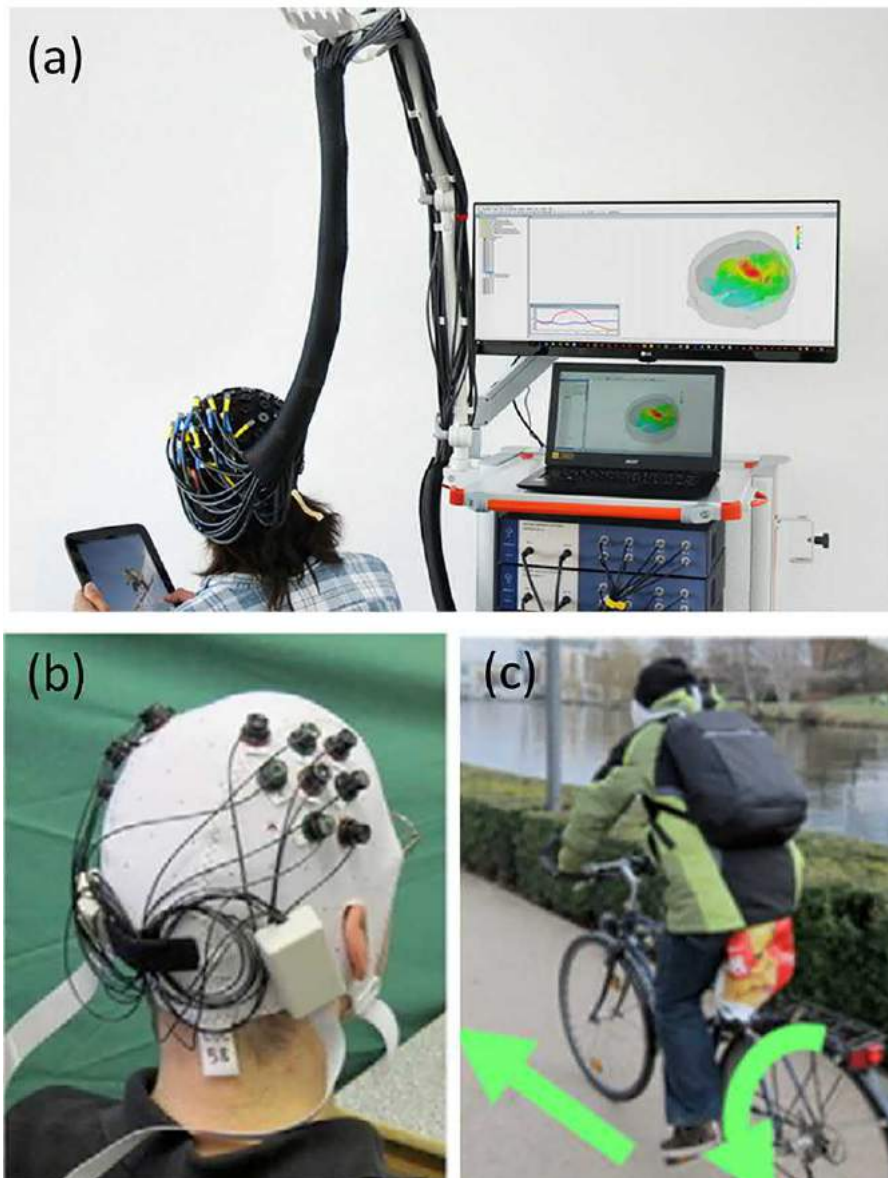


Fig. 8 (a) Commercialized fNIRS system tethered to a control system via fiber bundles, limiting subjects mobility. (b) Miniaturized fNIRS system only with electrical connections and compact electronic box attached to a head gear. (c) Subject wearing the miniaturized fNIRS system can move freely, demonstrating how this type of sensor can expand a kind of possible experiments. (b) and (c) from Piper, S.K., Krueger, A., Koch, S.P., *et al.*, 2014. A wearable multi-channel fNIRS system for brain imaging in freely moving subjects. *NeuroImage* 85, Part 1, 64–71.

Brain research

Since first demonstrated in 1993, functional near-infrared spectroscopy (fNIRS) has been increasingly employed to noninvasively investigate human brain activity by assessing changes in oxygen consumption. Using a deep penetration depth (a few millimeters) of near infrared light, fNIRS is capable of quantifying changes of oxy- and deoxy-hemoglobin concentration in specific brain regions upon external stimulations. The design with multiple pairs of source–detector covers a large area of human brain implementing 2D topological imaging or 3D tomographic imaging that can localize brain activity (Wilson *et al.*, 2016). However, this multichannel system inevitably involves heavy and cumbersome fiber bundles tethered to a remote control system. Most commercial systems use fiber bundles (Fig. 8(a)). The recent wireless fNIRS system based on a miniaturized optoelectronic source–detector pair and signal processing unit has been developed to enable evaluation of brain activity on subjects in a more natural environment (Fig. 8(b) and (c)) (Piper *et al.*, 2014).

Technical advances in optoelectronics/electronics have also opened a new era in functional neuroimaging of animals. Most current optical imaging systems are built with bulky optics and electronic hardware, requiring animals to be restrained and anesthetized, limiting types of experiments and affecting the baseline measurements. In contrast, a miniaturized head-mount optical imaging system enables awake and un-anesthetized animals to be freely moving. Current systems are capable of one photon and two photon imaging using external contrast agents and cerebral blood flow and optical intrinsic imaging via endogenous contrast (Yu *et al.*, 2015).

Summary

Recent technical advances in micro/nano-fabrication techniques, new materials and optoelectronic technologies have enabled fabrication of novel miniaturized optical sensors for in situ tissue diagnostics. In this article, we have discussed the basic construction of the sensors, core optoelectronic source and detector components, sensor fabrication techniques, data processing algorithms, and their clinical applications. The constant innovation in relevant technologies will accelerate development of these miniaturized optical sensors, which will ultimately lead to a paradigm shift in tissue optical diagnostics through noninvasive, objective, wearable, and wireless sensing.

References

- Chan, E.D., Chan, M.M., Chan, M.M., 2013. Pulse oximetry: Understanding its basic principles facilitates appreciation of its limitations. *Respiratory Medicine* 107, 789–799.
- Chandra, M., Scheiman, J., Heidt, D., *et al.*, 2007. Probing pancreatic disease using tissue optical spectroscopy. *Journal of Biomedical Optics* 12, 060501.
- Hu, F., Vishwanath, K., Lo, J., *et al.*, 2013. Rapid determination of oxygen saturation and vascularity for cancer detection. *PLoS One* 8, e82977.
- Kim, J., Gutruf, P., Chiarelli, A.M., *et al.*, 2017. Miniaturized battery-free wireless systems for wearable pulse oximetry. *Advanced Functional Materials* 27, 1604373.
- Lee, S.Y., Lloyd, W.R., Wilson, R.H., *et al.*, 2014. Photon-tissue interaction model for quantitative assessment of biological tissues. *Proceeding of SPIE* 2014, 893528.
- Lee, S.Y., Na, K., Pakela, J.M., *et al.*, 2017. Compact system with handheld microfabricated optoelectronic probe for needle-based tissue sensing applications. *Proceeding of SPIE* 2017.100540Y-100540Y-7.
- Lloyd, W.R., Chen, L.-C., Wilson, R.H., Mycek, M.-A., 2013. Biophotonics: Clinical fluorescence spectroscopy and imaging. In: Maitland, D.J., Moore JR., J.E. (Eds.), *Biomedical Technology and Devices Handbook*, second ed. Boca Raton, FL: CRC Press-Taylor&Francis Group.
- Lloyd, W., Wilson, R.H., Chang, C.-W., Gillispie, G.D., Mycek, M.-A., 2010. Instrumentation to rapidly acquire fluorescence wavelength-time matrices of biological tissues. *Biomedical Optics Express* 1, 574–586.
- Lochner, C.M., Khan, Y., Pierre, A., Arias, A.C., 2014. All-organic optoelectronic sensor for pulse oximetry. *Nature Communications* 5, 5745.
- Mycek, M.-A., Pogue, B.W. (Eds.), 2003. *Handbook of Biomedical Fluorescence*. New York, NY: Marcel Dekker, Inc.
- O'Sullivan, T., Munro, E.A., Parashurama, N., *et al.*, 2010. Implantable semiconductor biosensor for continuous in vivo sensing of far-red fluorescent molecules. *Optics Express* 18, 12513–12525.
- Piper, S.K., Krueger, A., Koch, S.P., *et al.*, 2014. A wearable multi-channel fNIRS system for brain imaging in freely moving subjects. *NeuroImage* 85 (Part 1), 64–71.
- Pitts, J.D., Mycek, M.-A., 2001. Design and development of a rapid acquisition laser-based fluorometer with simultaneous spectral and temporal resolution. *Review of Scientific Instruments* 72, 3061–3072.
- Wilson, R.H., Chandra, M., Chen, L.C., *et al.*, 2010. Photon-tissue interaction model enables quantitative optical analysis of human pancreatic tissues. *Optics Express* 18, 21612–21621.
- Wilson, R.H., Mycek, M.-A., 2011. Models of light propagation in human tissue applied to cancer diagnostics. *Technology in Cancer Research and Treatment* 10, 121–134.
- Wilson, R.H., Vishwanath, K., Mycek, M.-A., 2016. Optical methods for quantitative and label-free sensing in living human tissues: Principles, techniques, and applications. *Advances in Physics: X* 1, 523–543.
- Yu, H., Senarathna, J., Tyler, B.M., Thakor, N.V., Pathak, A.P., 2015. Miniaturized optical neuroimaging in unrestrained animals. *NeuroImage* 113, 397–406.

In Situ Optical Tissue Diagnostics/Laser Speckle-Based Spectroscopy Techniques in Biomedicine

Karthik Vishwanath and Sara Zanfardino, Miami University, Oxford, OH, United States

© 2018 Elsevier Ltd. All rights reserved.

Nomenclature

$\vec{r}$ Vector location of a point of interest in space
 τ Correlation time
 $\Phi(\vec{r})$ Photon fluence rate
 μ_a Absorption coefficient of medium

μ_s Scattering coefficient of medium
 D Optical diffusion coefficient of medium
 $S(\vec{r})$ Input source term
 g Anisotropy coefficient for single scattering of medium

Glossary

Absorption coefficient A wavelength-dependent medium property characterizing the probability of photon absorption in the medium.

Anisotropy coefficient A wavelength-dependent medium property characterizing the mean cosine of the angle between the directions of photon incidence and scattering in the medium.

Coherence The correlation between two (electromagnetic) waves either at different points in space (spatial coherence) or at two different points in time (temporal coherence).

Photon fluence rate The number of photons passing through a specified area per unit of time.

Photon radiance The net photon power passing through a specified area, along a specified direction, per unit of time.

Scattering coefficient A wavelength-dependent medium property characterizing the probability of photon scattering in the medium.

Speckle Small “spots” of bright and dark spots formed when a coherent source is incident on any surface due to constructive and destructive interference at the detector plane.

Temporal autocorrelation function The correlation of a time-dependent function, with itself across the full time-period measured of the original function.

Turbid medium Any medium appears “cloudy” or “murky” due multiple light scattering.

Introduction

Optical spectroscopic methods and techniques have been used for sensing of turbid media such as biological tissues and is an active area of research in the biomedical sciences. The basic prescription in most applications of these methods is to non-invasively derive optical properties of media using experimental measurements in combination with mathematical modeling and analysis to yield quantitative analogues of physiological markers related to human health and disease. Clinical applications over the last two decades have reported using optical spectroscopy for several applications including detection (including surgical margin identifications) of cancer, detection of atherosclerotic plaques, monitoring of burns, wounds and their healing, measuring cerebral functions, and imaging of retinal and dental health.

Optical techniques used for biosensing may be classified as two broad categories – static or dynamic – depending on whether the experimental techniques and methods used are inherently time-integrated, or time-resolved. A large fraction of optical methods that are used to characterize tissue optical properties including diffuse optical spectroscopy and tomography fall under the realm of static techniques – wherein one or more light sources are used to illuminate a medium of interest and a set of measurements (usually at several spatial locations) which are translated to produce a static spatial map of optical properties. It is to be noted that repeated measurements of such static optical techniques as defined here may still be used to study tissue (or biological) dynamics, at time-scales of the acquisition rates of the measurements. Dynamic optical spectroscopic techniques on the other hand make a set of measurements (again at one or more spatial locations) by measuring changes or fluctuations in light intensities (or a shift in the frequency spectrum) across several decades in time and seek to relate these fluctuations to a dynamical varying physical quantity such as motion of particles in the medium.

This article describes a dynamic optical technique – diffuse correlation spectroscopy (DCS) – which combines the theoretical formalism of incoherent photon propagation in a strongly scattering medium, to model experimental measurements of dynamic speckle intensity fluctuations (correlations) measured from the same medium. In DCS, measured speckle intensity fluctuations from an incident monochromatic source is used to compute temporal intensity autocorrelations which in turn are used to extract information about movement of scatterers within the volume of tissue probed by the diffuse optical waves transporting the incident light field to the detector. A closely related optical technique to DCS is Laser Doppler flowmetry which detects the scattered, frequency-shifted spectrum of a monochromatic (coherent) light source after passing through a scattering medium containing moving scatterers. Although these two techniques are closely connected analytically and theoretically as Fourier

transforms of each other, their experimental implementations and analysis are quite distinct. This article will only focus on the description, analysis and applications of DCS for tissue sensing.

Historical Overview and Background

When coherent light is scattered by a material, temporal measurements of the scattered field intensity exhibits fluctuations. Theoretical calculations of such fluctuations of light intensity when scattered by molecules (or other particles) have been studied since beginning of 20th century and include contributions from Einstein, Raleigh, Smoluchowski, and Debye. With the invention of the laser and fast detector-electronics, dynamic light scattering (DLS) became an accessible experimental technique to measure and study a wide variety of dynamic phenomena in complex media such as molecular rotation or diffusion in complex fluids, colloidal particle compositions and biological systems. Historically, DLS theory was developed first for weakly scattering systems (single scattering limit). However, experimentally, the methods of DLS are as applicable to multiply scattering systems as they were to weakly scattering media. DLS was soon applied to study of particle dynamics in various multiply scattering materials and the theoretical modeling of such experimental measurements came to be called diffusing wave spectroscopy (DWS). Experimentally, DCS uses the same experimental theme that is employed in DLS (or DWS) – but instead provides an elegant theoretical encapsulation to analyze the experimental measurements using the now well-established framework of diffusive optical transport in turbid media as sketched in Fig. 1.

Experimentally, all of these techniques employ a long-coherence length laser to illuminate the sample of interest and collect the scattered light intensity from a small volume. Fluctuations in the measured intensity incident on a photodetector and photon-counting system, $I(\vec{r}, t)$ can be obtained. The experimentally calculated intensity unnormalized and normalized correlation functions of these fluctuations are denoted respectively using their standard notations as:

$$G_2(\vec{r}, \tau) = \langle I(\vec{r}, 0) \cdot I(\vec{r}, \tau) \rangle = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T I(\vec{r}, t) I(\vec{r}, t + \tau) dt \quad (1)$$

$$g_2(\vec{r}, \tau) = \frac{\langle I(\vec{r}, 0) \cdot I(\vec{r}, \tau) \rangle}{\langle I(\vec{r}, 0) \rangle^2} \quad (2)$$

The integral in Eq. (1) defines the temporal correlation function of $G_2(\vec{r}, \tau)$ for an infinitely long time-signal $I(\vec{r}, t)$ while the angular brackets indicate ensemble averages. Although it is usually assumed that the system under observation is ergodic and thus ensemble averages are time averages this may not always be true – we discuss this in the section under Experimental Considerations below. In practice, the temporal correlation is usually calculated using a correlator board (or computer) to numerically evaluate the integral in Eq. (1). The smaller dt is, the more accurately measurements can be expected to match theoretical predictions, with the longest possible correlation time being the total duration of continuous data acquisition. Photodetectors with sampling rates in the 10–100 MHz range are typically used to collect DCS data for 1–10 s for in vivo tissue sensing.

Theoretically, the (detected) intensity is related to the scattered complex electric field from a monochromatic coherent source as $I(\vec{r}, t) = E(\vec{r}, t)E^*(\vec{r}, t) = |E(\vec{r}, t)|^2$ and the corresponding correlations of the detected unnormalized and normalized electric fields are defined as per convention by:

$$G_1(\vec{r}, \tau) = \langle E(\vec{r}, 0)E^*(\vec{r}, \tau) \rangle \quad (3)$$

$$g_1(\vec{r}, \tau) = \frac{\langle E(\vec{r}, 0)E^*(\vec{r}, \tau) \rangle}{\langle I(\vec{r}, 0) \rangle} \quad (4)$$

From the above definitions some properties of the field correlation functions can be computed. $G_1(\vec{r}, 0) = \langle |E(\vec{r}, t)|^2 \rangle = \langle I(\vec{r}, 0) \rangle$, giving $g_1(\vec{r}, 0) = 1$. Also, for long correlation times $G_1(\vec{r}, \tau) = \langle E(\vec{r}, t) \rangle^2 = 0$, giving $g_1(\vec{r}, \tau) = 0$. Given that it is the field fluctuations that give rise to the measured intensity fluctuations, theoretical methods are

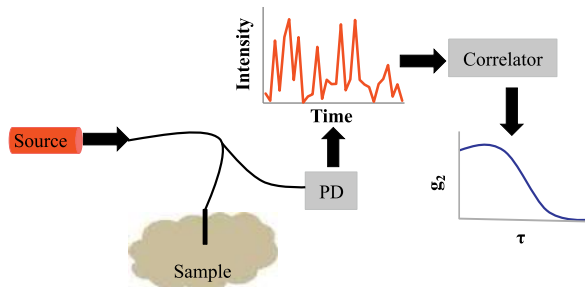


Fig. 1 Schematic representation of the DCS setup and its components. The illumination is provided by a laser source coupled to the sample using fiber optical probes. The detected signal is measured using a photodetector (PD) that can provide continuous (temporal) measurements. The monitored intensity fluctuations are recorded using a correlator device to compute the normalized intensity correlation g_2 .

used to generate models that predict behavior of $G_1(\vec{r}, \tau)$ and $g_1(\vec{r}, \tau)$. If the temporal fluctuations in $E(\vec{r}, t)$ are random and assumed to arise due to uncorrelated motions of several randomly (diffusing or moving) particles, then these fluctuations in $E(\vec{r}, t)$ will be normally distributed. Under this condition, $G_1(\vec{r}, \tau)$ and $G_2(\vec{r}, \tau)$ (and their normalized counterparts) are theoretically related through the Siegert relations:

$$G_2(\vec{r}, \tau) = \langle I(\vec{r}, 0) \rangle^2 + |G_1(\vec{r}, \tau)|^2 \quad (5)$$

$$g_2(\vec{r}, \tau) = 1 + |g_1(\vec{r}, \tau)|^2 \quad (6)$$

In DCS the functional form for $G_1(\vec{r}, \tau)$ is hypothesized to follow the same diffusion transport equation governing photon propagation in a multiply scattering medium, while $g_1(\vec{r}, \tau)$ is modeled using the random, independent single scattering model using the formalism of DLS.

Diffusive Photon Correlation Transport

Photon propagation in a turbid medium can be accurately modeled as a random walk of neutral particles in the medium using optical transport coefficients of the medium – typically specified as absorption, scattering and anisotropy. The radiative transport equation provides a theoretical analog for describing propagation of photon radiance within such media. The mathematical form of the transport equation prevents its use practice since it remains intractable for appropriately modeling experimental (boundary) conditions. The transport equation however reduces to a diffusion type equation if two conditions are met (a) if the medium's absorption is small in comparison to its scattering and (b) if the detected light has been scattered sufficiently and can be thus be considered emanating from an isotropic source. Both these conditions can largely experimentally satisfied – the first by selecting light with appropriate wavelengths, since the absorption and scattering properties of tissue are strongly wavelength dependent (for e.g. the near-infrared (NIR) wavelengths for tissue spectroscopy). The second condition may be satisfied by keeping the source-detector separations large in comparison to the mean scattering transport length of light in the medium of interest.

Under steady-state conditions (i.e. the source is CW or when the detected signal is time-integrated for longer than a few ns, in general) the diffusion equation for a semi-infinite homogenous medium with uniform optical properties the diffusion equation is written as:

$$[D\nabla^2 - \mu_a]\Phi(\vec{r}) = S(\vec{r}) \quad (7)$$

Here, $\Phi(\vec{r})$ represents the photon fluence rate (Wm^{-2}) as the net light energy at all points of interest in the medium, $S(\vec{r})$ describes the input source term, and D the medium's diffusion coefficient given by $D = 3[\mu_a + \mu_s(1 - g)]$, where μ_a is the medium's absorption coefficient, μ_s its scattering coefficient and g the single-scattering anisotropy coefficient. Under the formalism of DCS, the unnormalized field correlation $G_1(\vec{r}, \tau)$ is propagated using the diffusion equation as:

$$\left[D\nabla^2 - \mu_a - \frac{1}{3}\mu_s(1 - g)k_0^2 \langle \Delta r^2(\tau) \rangle \right] G_1(\vec{r}, \tau) = S(\vec{r}) \quad (8)$$

The $\langle \Delta r^2(\tau) \rangle$ in Eq. (8) is the only additional term for the transport of the field correlation in a homogenous turbid medium and represents the mean-square displacement of the diffusing (or moving) particles, which add to the dynamics of the decay in the field correlation. Since closed form expressions of $\Phi(\vec{r})$ for Eq. (7) are readily available for several geometries, they can be used to calculate $G_1(\vec{r}, \tau)$ for measurements made in those geometries by replacing μ_a with $\mu_a - \frac{1}{3}\mu_s(1 - g)k_0^2 \langle \Delta r^2(\tau) \rangle$.

Experimental Considerations

Experimentally, one has to make choices about the illuminating source, the delivery and collection optics to couple to the medium, and the detection system (see Fig. 1). Fig. 2 shows representative examples of experimental measurements obtained using a fiber-based DCS system while for different light sources, different detection fiber sizes and different types of scattering media.

Light Sources

Given that the theoretical formulation for modeling measurements assumes a monochromatic plane wave, DCS sources are usually single mode, stabilized, long-coherence length lasers operating in between 600–800 nm. In order for the experimental data to contain reliable dynamic information about the motion of scatterers, the temporal coherence length must be much longer the average path length of photons diffusing from the source and the detector placed on the tissue. Most applications reported with DCS use sources with coherence lengths longer than several meters.

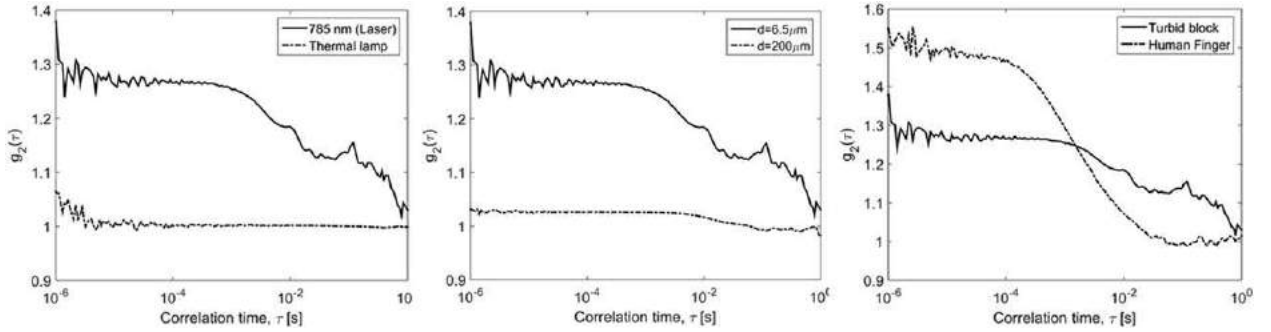


Fig. 2 Measured representative intensity correlation curves (g_2) depicting data obtained for changing incident light source (left), changing detection fiber radius (middle) and different media (right). In all the plots, the solid line represents the autocorrelation detected using a fiber probe (source fiber radius = 200 μm , detector fiber radius = 6.5 μm and center-center distance of 1.5 mm) in reflection geometry on a turbid solid block. The dashed line in the left shows the correlation when the source was changed to a thermal halogen lamp. In the middle figure, the second fiber probe was used that had identical properties as the first probe but with detector radius of 200 μm while using the laser light for illumination. The dashed line in the right figure shows data from the first probe in contact with a human finger under laser illumination. These data indicate how both the intercept on the y-axis as well as the decay curves change across different conditions.

Coupling Optics

Although it is possible to make DCS devices capable of operating in a non-contact mode (i.e. coupling light to and from the medium without physical contact), the most practical method of using DCS for biomedical applications has been to couple light from the source to a medium (tissue site) using a source-fiber and obtaining diffusely scattered measurements using one or more detection fibers. The detection fiber is typically a low-mode fiber (ideally capable of monitoring fluctuations of a single speckle) while the source-fiber is a multi-mode fiber. The common end of this probe sets the source-detector geometry used to probe the sample volume and usually is larger than 5–10 mm for most DCS systems to stay within validity limits of diffusion theory.

Detection Systems

The ability to extract temporal correlations relies which not only requires sampling a fluctuating intensity signal at sampling rates of 10–100 MHz but also to store and process these signals to obtain the required intensity autocorrelation function. Traditional photon detection systems used an amplified photomultiplier tube (PMT) that with high enough response rates, while most recent systems employ detectors are small-area single photon avalanche photodiodes (APD). The digitized output sampled off the photodetector is to a digital correlator – usually a dedicated hardware board that computes required time correlation functions.

Modeling Measured Intensity Correlations

As described above, the experimental observations yield $g_2(\vec{r}, \tau)$ that is the intensity correlation for a measured at the detector situated at location $\vec{r}$ relative to the source and the medium. Thus, the analytical expressions for $G_1(\vec{r}, \tau)$ are converted to calculate $g_1(\vec{r}, \tau)$ which in turn is related to the experimental measurement of $g_2(\vec{r}, \tau)$ as:

$$g_2(\vec{r}, \tau) = 1 + \beta |g_1(\vec{r}, \tau)|^2 \quad (9)$$

Here, β is an experimental factor that is lesser than unity and depends on specific experimental configuration details such as the coherence length, coupled modes, dark noise and detection aperture size and thus is usually obtained as a free parameter during the fitting process. Besides β , experimentally measured data is fit to a specified functional formulation to extract parameters contributing to $\langle \Delta r^2(\tau) \rangle$. For example two commonly used models in tissue spectroscopy include diffusive flow where $\langle \Delta r^2(\tau) \rangle = 6D_B\tau$ where D_B is a diffusion coefficient, or random flow where $\langle \Delta r^2(\tau) \rangle = V^2\tau^2$ and V^2 is the second moment of the moving particle's velocity distribution. It is important to note that analytical fitting process is possible only by having estimates for the optical absorption and scattering coefficients of the medium, at the wavelength of the incident source.

Applications to In Vivo Tissue Sensing

Fiber-based DCS measurements have been obtained across a variety of clinical settings to address problems in monitoring progress of treatments, for monitoring cerebral blood flow during ischemia and for assessment of wound healing. Ultimately, the largest factor of contrast in biological tissues for the decorrelation of the measured intensity autocorrelations is attributed to the motion

of red blood cells in the vascular system. Thus, nearly any health application that requires monitoring of blood flow and its relative changes across perturbations could potentially use DCS a simple, non-invasive clinical adjunct.

DCS has been explored to investigate tissue perfusion since angiogenesis (i.e. the growth of new blood vessels) is now believed to be a marker of tumor growth and progression. Further, since treatments of tumors including chemotherapy and photodynamic therapy would be expected to impact blood flow and its dynamics during such treatments, DCS measurements have been applied to investigate these changes longitudinally in cancers of the prostate, breast and of the head and neck. In many cases DCS has been shown to provide an early indicator of treatment response (and failure). Given that traditional therapy monitoring techniques that include methods such as PET, MRI or CT, which many times are not accessible or amenable for use, DCS may provide a convenient, clinically-adoptable alternative.

DCS has also been used in applications to monitor cerebral blood flow – such monitoring studies were focused particularly on investigating perfusion rates and changes in the brain during ischemic strokes. An ischemic stroke the most common type of hemorrhagic stroke where a blood clot in the body constricts or depletes blood supply to certain regions of the brain. Since such a blockage would result in markedly different perfusion rates, DCS was an appropriate candidate for continuous and non-invasive blood in patients.

Other clinical applications of DCS have explored monitoring patient recovery from a traumatic wound injury, monitoring the blood supply to maxillofacial autographs during early stages of transplant, and monitoring the blood supply to skeletal tissues and muscles during certain high-intensity activities such as exercise. Several of these studies that employ DCS for measurement of tissue perfusion have been validated against clinically used “gold-standards” for flow-monitoring such as CT and MRI.

Preliminary and current data in clinical trials is promising for the field of non-invasive biomedical optical sensing. With an increasing number of studies showing successful applications of dynamic light scattering spectroscopy for in vivo and real-time monitoring of blood flow in tissues, it is likely that DCS-based devices will become more widespread. These application areas will benefit if DCS devices become commercially available as out-of-the box instruments and could become a routine and important tool used for sensing in vascular function or dysfunction in humans.

Further Reading

- Ackerson, B.J., Dougherty, R.L., Reguigui, N.M., Nobbmann, U., 1992. Correlation transfer – Application of radiative-transfer solution methods to photon-correlation problems. *Journal of Thermophysics and Heat Transfer* 6, 577–588.
- Berne, B.J., Pecora, R., 1990. In *dynamic light scattering with applications to chemistry*. New York: Wiley-Interscience.
- Boas, D.A., Yodh, A.G., 1997. Spatially varying dynamical properties of turbid media probed with diffusing temporal light correlation. *Journal of the Optical Society of America* 14, 192–215.
- Bonner, R., Nossal, R., 1981. Model for laser Doppler measurements of blood flow in tissue. *Applied Optics* 20, 2097–2107.
- Briers, J.D., 2001. Laser Doppler, speckle and related techniques for blood perfusion mapping and imaging. *Physiological Measurement* 22, R35–R66.
- Durduran, T., Choe, R., Baker, W.B., Yodh, A.G., 2010. Diffuse optics for tissue monitoring and tomography. *Reports on Progress in Physics* 73 (7), doi:10.1088/0034-4885/73/7/076701.
- Farzam, P., Zirak, P., Binzoni, T., Durduran, T., 2013. Pulsatile and steady-state hemodynamics of the human patella bone by diffuse optical spectroscopy. *Physiological Measurement* 34, 839–857.
- Kim, M.N., Durduran, T., Frangos, S., *et al.*, 2010. Noninvasive measurement of cerebral blood flow and blood oxygenation using near-infrared and diffuse correlation spectroscopies in critically brain-injured adults. *Neurocritical Care* 12, 173–180.
- Pine, D.J., Weitz, D.A., Chaikin, P.M., Herbolzheimer, E., 1988. Diffusing wave spectroscopy. *Physical Review Letters* 60, 1134–1137.
- Yu, G., 2012. Near-infrared diffuse correlation spectroscopy in cancer diagnosis and therapy monitoring. *Journal of Biomedical Optics* 17, 010901.
- Wilson, R.H., Vishwanath, K., Mycek, M.-A., 2016. Optical methods for quantitative and label-free sensing in living human tissues: Principles, techniques, and applications. *Advances in Physics*. 1–21.

Photodynamic Therapy and Photobiomodulation: Can All Diseases be Treated with Light?

Michael R Hamblin, Massachusetts General Hospital, Boston, MA, United States; Harvard Medical School, Boston, MA, United States; and Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction. The Janus Face of Photomedicine

Janus (or Ianus) was the Roman god of beginnings, gates, transitions, time, doorways, passages, and endings. He is usually depicted as having two faces looking in opposite directions, since he looks to the future and also to the past ([Fig. 1](#)). It is thought that the month of January was named for Janus (Ianuarius) because he presided over the ending of one year and the beginning of the next. Janus also presided over the beginning and ending of conflict, and hence over war and peace. The doors of his temple were open in time of war, and closed to mark the onset of peace. The most famous temple to Janus in Rome is called the Ianus Geminus, or “Twin Janus.” When its doors were open, neighboring cities knew that Rome was at war. As a god of transitions, he had functions pertaining to birth and to journeys and exchange.

The two-faced nature of the Roman God, Janus has become inextricably linked with many aspects of biology and medicine. The concept of a process, mechanism, treatment, or drug having both good and bad effects is so widespread and pervasive, that the analogy of Janus has been seized upon for a variety of reasons. Perhaps the best-known application of the Janus name in biology, is Janus kinase (JAK), which represents a family of intracellular, non-receptor tyrosine kinases that transduce cytokine-mediated signals via the JAK-STAT pathway. They were initially named “just another kinase” 1 and 2 (since they were just two of a large number of discoveries in a PCR-based screen of kinases ([Wilks, 1989](#))), but were ultimately published as “Janus kinase”. The name was taken from the two-faced Roman god of beginnings and endings, Janus, because the JAKs possess two near-identical phosphate-transfer domains. One domain mediates the kinase activity, while the other negatively regulates the kinase activity of the first domain.

We intend to show in this chapter that light therapy has many characteristics that can accurately be described as a “two-faced” or “Janus effect”. Depending on whether or not light is combined with a non-toxic photosensitizing dye (photodynamic therapy), and on the overall dose of light (energy density and power density), light is able to kill just about anything that is living such as cancers, microorganisms, blood vessels, parasites, pests, unwanted tissues etc, etc. On the opposite side of the coin (or perhaps on the other side of Janus’ face), light of the correct wavelength (photobiomodulation) and at the right dose (generally accepted to be “low” as in “low level laser therapy”) can have exactly the opposite effect, being able to heal, regenerate, protect, revitalize and restore any kind of dead, damaged, stressed, dying, degenerating cells, tissue, or organ system. Between the bad destroying face of photodynamic therapy, and the good healing face of photobiomodulation, it may well be concluded that indeed it can be said “all diseases can be treated with light”.

Photodynamic Therapy – The Killer

Photodynamic therapy (PDT) is the combination of non-toxic dyes and harmless visible light that combine together in the presence of oxygen to produce highly toxic reactive oxygen species (ROS) that can kill unwanted cells like cancer cells and pathogenic microorganisms ([Hamblin and Mroz, 2008](#)). PDT was discovered over one hundred years ago in 1900 when Oscar Raab (a medical student working with Herman von Tappeiner) on the toxicity of acridine dyes to protozoa. Raab realized that the



Fig. 1 Different depictions of the two-faced Roman god Janus.

variable results he obtained in his experiments using acridine orange and paramecia, could be attributed to variations in the level of ambient sunlight on the days when the experiments were performed. von Tappeiner and a dermatologist named Jesionek published the first report (just 3 years after Raab's initial paper) describing the use of 1% eosin solution painted on the skin, with white light illumination to successfully treat patients with basal cell carcinoma of the skin. In the decade that followed, von Tappeiner's group showed that this "photodynamic" effect required the simultaneous presence of a photosensitizer, light, and oxygen to mediate cellular toxicity. Freiderich Meyer-Betz then performed the first human trial of hematoporphyrin-mediated PDT by injecting himself with 200 mg of the drug and standing in sunlight.

It was however not until 1978 that Thomas Dougherty and his colleagues at Roswell Park Cancer Institute (RPCI) reported the first large-scale trial of PDT in human patients with cancer (Cengel *et al.*, 2016). In total, 113 primary and metastatic skin tumors were treated in 25 patients with diagnoses ranging from malignant melanoma to cutaneous T-cell lymphoma to metastatic breast and endometrial cancers. Many of those tumors were recurrent or had progressed after multiple cytotoxic chemotherapy agents and had failed local ionizing radiotherapy.

Since those early days PDT has continued to progress and has significantly widened the range of diseases and disorders that can be treated (Hamblin and Mroz, 2008), in addition to cancer (Agostinis *et al.*, 2011). The ROS produced during PDT can rapidly and completely kill any type of pathogenic microorganism, whether it be different classes of bacteria, viruses, yeasts, filamentous fungi, parasites (such as *Leishmania*), and also remove unwanted tissue such as atherosclerotic plaque, undesirable blood vessels, and scars.

Mechanisms of PDT

Fig. 2 shows the mechanisms that operate during PDT in a so-called Jablonski diagram. The ground state PS has two electrons with opposite spins (singlet state) in the lowest energy molecular orbital. Following the absorption of light (photons), one of these electrons is boosted into a higher-energy orbital but retains its spin (first excited singlet state). This is a short-lived (nanoseconds) species and can lose its energy by emitting light (fluorescence) or by internal conversion into heat. The fact that most PS are fluorescent has led to the development of sensitive assays to quantify the amount of PS in cells or tissues, and allows in vivo fluorescence imaging in living animals or patients to measure the pharmacokinetics and distribution of the PS. The excited singlet state PS may also undergo the process known as intersystem crossing whereby the spin of the excited electron inverts to form the relatively long-lived (microseconds to milliseconds) excited triplet-state that has both electron spins parallel. The long life of the triplet state is attributed to the relative rarity of other triplet molecules with which it can interact.

The PS excited triplet can undergo three broad kinds of reactions that are usually known as type I and type II (see Jablonski diagram in Fig. 2). Firstly, in a type I reaction, the triplet PS can gain an electron from a neighboring reducing agent. In cells these reducing agents are commonly either nicotinamide adenine dinucleotide (NADH) or NAD-phosphate, reduced form (NADPH). The PS is now a radical anion bearing an additional unpaired electron ($PS^{\bullet-}$). Alternatively two triplet PS molecules can react together involving intermolecular electron transfer to produce a pair consisting of a radical cation and a radical anion. The PS radical anions may further react with oxygen to carry out an electron transfer to produce reactive oxygen species, in particular superoxide anion. The damaging effects of superoxide are relatively mild, however it can cause much more oxidative damage when it reacts with itself to produce hydrogen peroxide and oxygen, in the process called "dismutation". Hydrogen peroxide, can add to an organic (carbon-containing) substrate and result in an oxidizing "chain reaction". This is common in the oxidative damage of fatty acids and other lipids. Hydrogen peroxide is important in the production of the highly reactive hydroxyl radical ($HO^{\bullet}$), by a second one-electron reduction mediated by $PS^{\bullet-}$. Superoxide will additionally react with nitric oxide ($NO^{\bullet}$) (also a radical) to produce peroxynitrite ($ONOO^-$), another highly reactive molecule that can oxidize many functional groups and can also nitrate proteins on tyrosine residues.

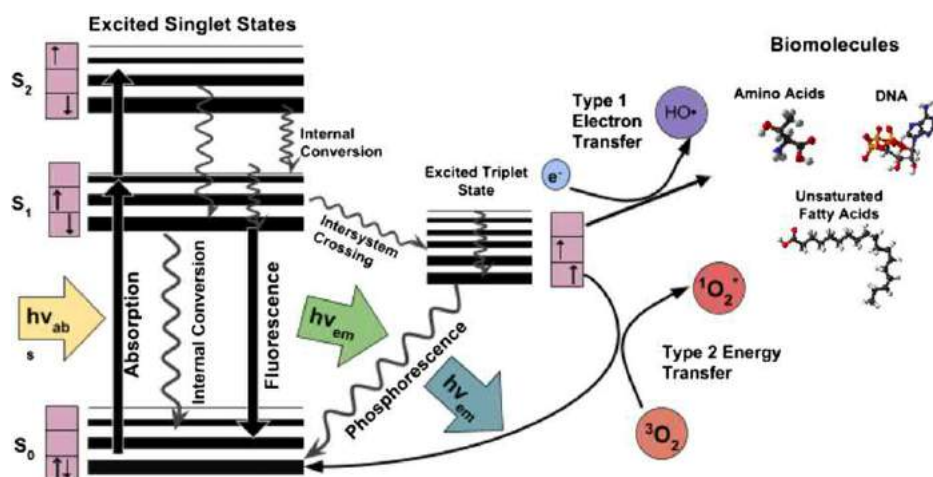


Fig. 2 Jablonski diagram showing photochemical reactions that occur during PDT.

In a type II reaction, the triplet PS can transfer its energy directly to molecular oxygen (itself a triplet in the ground state), to form excited-state singlet oxygen. Type II processes are thought to best preserve the PS molecular structure in a repeatedly photoactivatable state, and in some circumstances a single PS molecule can generate 10,000 molecules of singlet oxygen. The PS can in some circumstances also react with the singlet oxygen it produces in a process known as oxygen-dependent photobleaching.

Singlet oxygen generated during type II photochemical reactions, is believed to be the most important molecule responsible for PDT-induced cellular damage (Henderson and Miller, 1986; Weishaupt *et al.*, 1976). However, because of the high reactivity and short half-life of singlet oxygen only molecules and structures that are proximal to the area of its production (areas of PS localization) are directly affected by PDT. The half-life of singlet oxygen in biological systems is < 40 ns, and, therefore, the radius of the action of singlet oxygen is of the order of 20 nm (Moan and Berg, 1991). In singlet oxygen the spin restriction that reduces reaction rates of triplet oxygen with non-radicals is excluded. Consequently, singlet oxygen has an empty low-energy valence orbital that makes it a potent oxidizing/oxygenating agent compared to its ground-state counterpart. Singlet oxygen interacts with molecules via two essential mechanisms: physical interactions with quenchers or chemical reactions with biomolecules and scavengers. In the first phenomenon singlet oxygen transfers its excitation energy to an acceptor molecule (quencher) promoting it to an excited state while oxygen returns to ground-state. The excited quencher can be cleaved forming new products or react with biomolecules and cause damages. Antioxidant molecules, such as carotenoids, can alternatively dissipate the energy quenched from singlet oxygen in the form of heat, preventing any sort of oxidative damage. Even though, most carotenoids are very hydrophobic substances and exclusively accumulate in cell membranes (Kearns, 1971).

Singlet oxygen mainly reacts with molecules containing thiol groups or double bounds. Hence, it presents more restricted reaction patterns in relation to free radicals produced by type I reactions. Carbon and nitrogen atoms form double bounds sharing two electrons with opposite spin orientation. Note that ground-state oxygen cannot receive those two paired electrons because both low-energy orbitals are already occupied by one unpaired electron. Since singlet oxygen has an empty low-energy orbital, it has no restrictions to react with double bounds (Bland, 1976; Kearns, 1971).

Type I photodynamic reactions start off via donation of one electron from the excited PS ($^3\text{PS}^*$) to O_2 producing superoxide radical ($\text{O}_2^{\bullet-}$) by a 1-electron reduction, and then going on to form hydrogen peroxide by a 2-electron reduction of O_2 that can yet produce more radicals. Radical species are typically paramagnetic because they carry one or more unpaired electrons in valence shell orbitals. This feature makes radicals particularly reactive. They are susceptible to react with several possible biomolecules passing the unpaired electron to another molecule also turning it into radical species or releasing another free radical. Therefore, a single radical formed can initiate chain reactions that lead to damage of several biomolecules. Radical reactions are rather unspecific and can attack lipids, sugars, proteins and nucleic acids forming radical moieties within these biomolecules. Such radical moieties then allow biomolecules to react with each other forming cross-links between them (e.g. lipid-lipid, lipid-protein, protein-DNA, etc.). Either radical addition to biomolecules through hydroxylation and peroxidation or cross-link formation will potentially inhibit biomolecular function. As direct consequences, proteins are denatured, polysaccharides cleaved, nucleic acids are damaged, membranes become less fluid, more porous and may even be disrupted. Proteins can yet be targeted for proteasome degradation.

Lipid peroxidation represents a critical source of cellular damage that often leads to necrosis. Because hydroxyl radicals are extremely reactive non-polar species they can rapidly abstract a hydrogen atom from saturated lipids forming lipid radicals ($\text{L}^{\bullet}$). Secondly, ground-state oxygen molecules can add to the lipid radical forming peroxy radicals that can subsequently react with other lipids that will propagate chain reactions until recombination or dismutation reactions take place. Since cell membrane consists of double layer of phospholipids, closely interacting with each other and with membrane proteins, radical chain reactions can spread through the membrane causing serious damages that ultimately result in necrosis via membrane rupture (Halliwell and Chirico, 1993). To minimize the extent of this type of damage, nearly all cells in nature accumulate antioxidants molecules and enzymes in membranes preventing the propagation of radical reactions (Buettner, 1993; Halliwell and Gutteridge, 2015). All amino acids, including when modified, are susceptible to damage caused by radical reactions. Protein function is highly associated to its structural conformation, that is resultant of the chemical characteristics of each amino acid in the polymeric chain. Of course, some amino acids are more important than others for protein structure and active site, however, the oxidation of any amino acid in the chain can potentially alter protein conformation and consequently its function. In the case of membrane proteins, note that transmembrane domains are constructed with chains of hydrophobic amino acids. If these amino acids are oxidized they become more hydrophilic and may detach from cell membrane to be solvated by water either in the cytosol or the extracellular media (Davies, 2016).

There are an astonishing array of known compounds that can act as photosensitizers in PDT. In the interests of space we can only give a few examples of their chemical structures here. Fig. 3 shows the structure of Photofrin (the first PS to be clinically used for various cancers, and still one of the commonest). Foscan or Temoporfin (*m*-tetrahydroxyphenylchlorin) is another PS that is commonly used in the clinic to treat different types of cancers. Methylene blue is a phenothiazinium dye, which is the most commonly used PS to treat infections.

Antimicrobial Photodynamic Inactivation and PDT for infectious Disease

Rise of antibiotic resistance

Together with vaccines and public health measures to control transmission of communicable diseases, antibiotics have helped to dramatically reduce mortality from infectious disease during the 20th century (Guyer *et al.*, 2000). However, the continuing global challenge of pathogens such as influenza, tuberculosis, and malaria, as well as the emergence of human

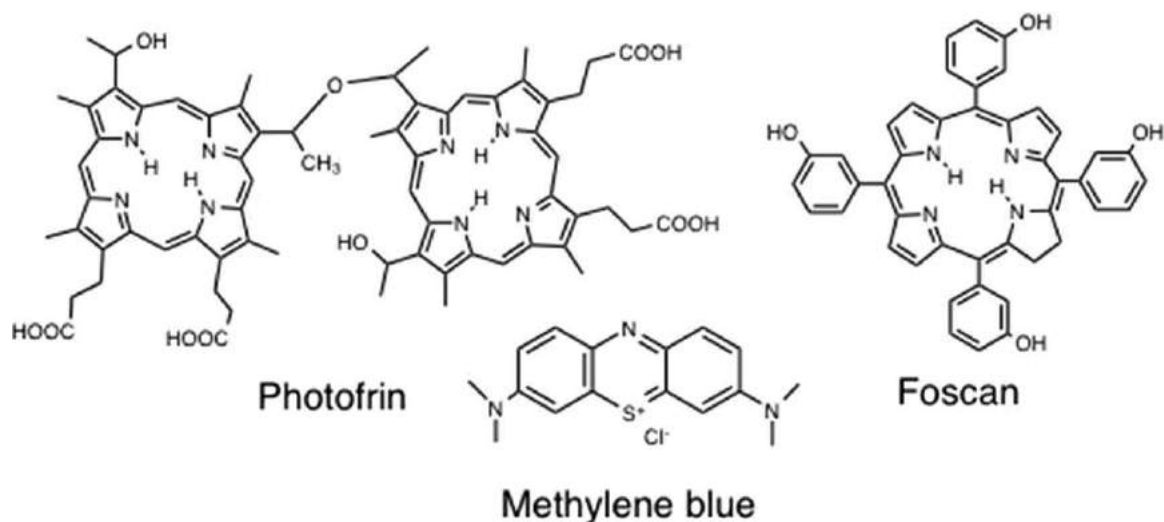


Fig. 3 Chemical structures of three clinically used PS: Photofrin, Foscan, Methylene blue.

immunodeficiency virus, means that infectious disease remains an important problem (Fauci and Morens, 2012). The real value of antibiotics goes beyond simply preventing death and illness due to infection, considering that antibiotics also permit the serious iatrogenic assault on the immune system (immunosuppression) that may occur during cancer treatment or in organ transplantation, where antibiotics have helped to keep surgical complication rates down (Youngblood *et al.*, 2002). Therefore, antibiotics are an extremely valuable resource across the whole spectrum of modern medicine (Smith and Coast, 2013). However, multidrug-resistant (MDR) and pandrug-resistant (PDR) bacterial strains and the related infections they cause, have become emerging threats to public health throughout the world (Kraus, 2008). These infections are associated with an approximately two-fold higher death rate, and with considerably prolonged hospital stays (Munoz-Price *et al.*, 2013). The infections caused by antibiotic resistant microbial strains are often exceptionally hard to treat due to the limited range of therapeutic options (Yoneyama and Katsumata, 2006). Therefore, there is a need for an immediate and all-out search for alternative antimicrobial methods against MDR strains, towards which no resistance likely to develop (Hamblin and Hasan, 2004; Maisch, 2007; Maisch *et al.*, 2011). Recently, Karen Bush *et al.* pointed out that novel non-antibiotic approaches for the prevention and treatment of infectious disease should be considered high-priority international research and development goals (Bush *et al.*, 2011).

In 2015 the O'Neill report garnered much international attention when it delivered the alarming forecast that by 2050 (if nothing were done to stem the growth of MDR bacteria) there would have been 300 million premature deaths that would have cost the world economy \$100 trillion (O'Neill, 2015). One of the most promising and innovative approaches to kill drug-resistant bacteria and treat resistant infections is antimicrobial photodynamic inactivation (aPDI).

History of aPDI

The initial discovery of the PDT/PDI effect was made by in 1900 by a PhD student called Oscar Raab working in the laboratory of Prof Hermann von Tappeiner in Munich, Germany. Raab was studying the toxicity of acridine dyes towards paramecia (a single-celled ciliated protozoan microorganism living in ponds). Raab found that his results were quite variable, but recognized that the apparent toxicity depended on the time of day, and the amount of light coming through the laboratory window, with no toxicity when the experiment was done in the dark. There were sporadic reports of killing of different kinds of microorganisms (bacteria, fungi and viruses) for the next 90 years, but it wasn't until 1990 that major progress was made.

Pathogen classification and PS structure

In the 1990s it was observed that there was a fundamental difference in susceptibility to PDT between Gram (+) and Gram (−) bacteria. It was found that in general neutral or anionic PS molecules are efficiently bound to and photodynamically inactivate Gram (+) bacteria whereas they are bound to a greater or lesser extent only to the outer membrane of Gram (−) bacterial cells but do not inactivate them after illumination (Malik *et al.*, 1992). The high susceptibility of Gram (+) species was explained by their physiology as their cytoplasmic membrane is surrounded by a relatively porous layer of peptidoglycan and lipoteichoic acid that allowed PS to cross. The cell envelope of Gram (−) bacteria consists of an inner cytoplasmic membrane and an outer membrane, which are separated by the peptidoglycan-containing periplasm (Fig. 4). The outer membrane forms a physical and functional barrier between the cell and its environment. In the outer membrane several different proteins are present, some of them function as pores to allow passage of nutrients, whereas others have an enzymatic function or are involved in maintaining the structural integrity of the outer membrane and shape of the bacteria.

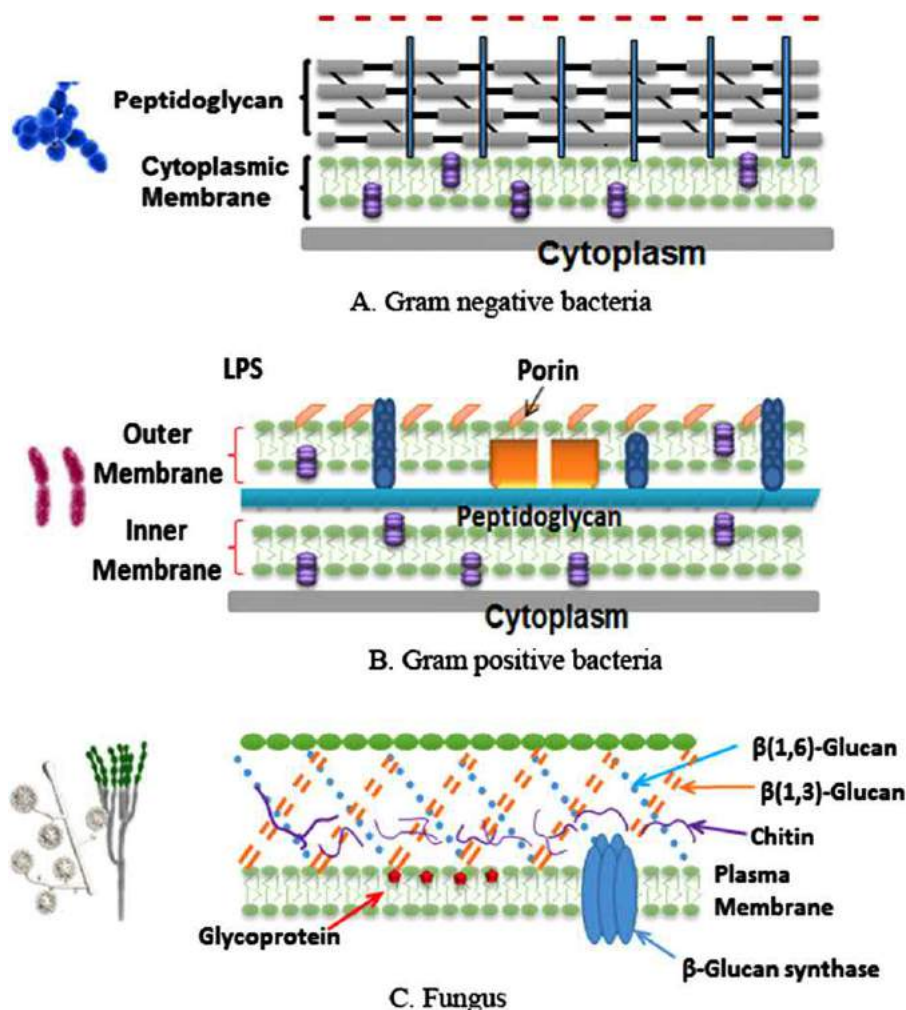


Fig. 4 Cell wall structures of (A) Gram-positive bacteria; (B) Gram-negative bacteria; (C) Fungi.

Then several groups of workers devised approaches that would allow PDI of Gram (–) species. The group in Israel consisting of Malik, Nitzan and co-workers (Nitzan *et al.*, 1992) used the polycationic peptide polymyxin B nonapeptide (PMBN) which increased the permeability of the Gram (–) outer membrane and allowed PS that are normally excluded from the cell to penetrate to a location where the reactive oxygen species generated upon illumination to execute fatal damage. PMBN does not release LPS from the cells but “expands” the outer leaflet of the membrane allowing PS such as deuteroporphyrin (DP) to penetrate and allow PDI of *E. coli* and *P. aeruginosa*. They demonstrated an interaction between PMBN and DP in solution and speculated that this binding assisted the penetration. They used this method to kill a multi-antibiotic resistant strain of *Acinetobacter baumannii* and found that DP seemed to work much better in concert with PMBN than many other PS including porphyrins, phthalocyanines and merocyanine 540 (Nitzan *et al.*, 1998). They also found that the growth medium of the bacteria made a large difference in their susceptibility to PDI, with the high protein brain-heart infusion leading to less killing than the low protein nutrient broth. They found that the type of protein present in the medium made a difference in addition to the concentration. A similar approach was taken by Bertoloni *et al.* (1990) who found that the use of tris- ethylenediamine tetra-acetic acid (EDTA) to release LPS or the induction of competence with calcium chloride sensitized *E. coli* and *Klebsiella pneumoniae* to PDI by haematoporphyrin or zinc phthalocyanine.

A second approach adopted by several groups is to use a PS molecule with an intrinsic positive charge. The groups of Wilson (2004) and Wainwright (1998) have used the phenothiazinium salts such as MB and toluidine blue O to carry out PDI of a large range of bacteria both Gram (+) and Gram (–). The group in Italy led by Jori have used cationic porphyrins (meso-tetra (N-methyl)-4-pyridyl)-porphine tetraiodide and tetra-(4N,N,N-trimethyl-anilinium)-porphine to photoinactivate Gram (–) species such as *Vibrio anguillarum* and *E. coli* (Merchat *et al.*, 1996). They found that washing the loosely bound PS from the cells before illumination decreased the killing and explained this finding by supposing that the first dose of light on PS bound to the outside of the outer membrane causes an initial limited photodamage that then allows further penetration of the PS. The group in Leeds, UK led by Brown has used cationic phthalocyanines for PDI of Gram (–) bacteria (Minnock *et al.*, 1996). They investigated

E. coli DH5a and in particular the mechanism of uptake. They found that incubation with PPC in the dark led to increased sensitivity of the bacteria to hydrophobic but not hydrophilic antibiotics. Incubation with PPC also led to increased uptake of radiolabeled protoporphyrin that was reversed in the presence of up to 50 mM Mg^{++} ions. These observations were consistent with the uptake of PPC proceeding through the self-promoted uptake pathway (see next section).

However there are other reports of PDI of Gram (–) bacteria in which it is clear that the PS does not have to penetrate the bacterium to be effective, or indeed even come into contact with the cells. According to these papers, if singlet oxygen can be generated in sufficient quantities near to the bacterial outer membrane it will be able to diffuse into the cell to inflict damage on vital structures. In one set of studies, the bacteria were separated from the PS by a layer of moist air and singlet oxygen in the gas phase was generated and allowed to diffuse across the gap before contacting the bacteria (Valduga *et al.*, 1993). Gram (–) species were harder to kill than Gram (+), and the intracellular content of carotenoids was found to protect the bacteria from photo-inactivation. In another study, the PS Rose Bengal was covalently bound to small polystyrene beads that were allowed to mix with the bacteria in suspension (Dahl *et al.*, 1989).

How can these two conflicting sets of findings; one pointing out the necessity for PS penetration into Gram (–) bacteria, and the other pointing out that extracellular generated singlet oxygen is effective, be reconciled? One possibility is that indeed singlet oxygen can diffuse into bacteria producing fatal damage if it is produced at the outer surface or in solution in close proximity. The diffusion distance of singlet oxygen in solution has been estimated to be approximately 20–100 nm (Klaper *et al.*, 2016). The failure of some PS that bind to Gram (–) species to produce any killing however, must mean that the reactive species produced on illumination were unable to diffuse inward to sensitive sites. The fact that these PS can lead to efficient PDI of Gram (+) species shows that the molecules have not undergone any conformational change that has rendered the PS no longer photoactive. Hence we hypothesize that PS that operate chiefly via Type I mechanisms need to penetrate the Gram (–) outer membrane, while those that act mainly by Type II PS may not, and that the former may be more efficient.

Virulence

Photodynamic therapy is one of the few antimicrobial therapies, that can not only kill microbial cells, but can also destroy many virulence factors that microbial cells produce in order to facilitate their adherence, invasion, survival, and toxicity in host cells and tissues. These virulence factors are usually biomolecules such as proteins, lipids and carbohydrates, that are just as susceptible to oxidative damage by the ROS produced during PDT, as the microbial cells themselves.

PDT for infectious disease

In recent years the application of PDT for the treatment of localized infections called aPDI has steadily increased in popularity (Fig. 5). This application can superficially appear much simpler than the more well-recognized application of PDT for cancer treatment (see below). The complexity of dealing with the vascular effects of PDT is absent, because the PS is almost always topically applied into/onto the infected area of tissue. Moreover, there have not as yet been many reports about activation of the immune system being triggered after carrying out aPDI for an infectious disease (see later).

So it is generally assumed that a PS with cationic charges, which is more or less selective for microbial cells, is injected or otherwise topically applied into the infected area, where it binds to the microbial cells after a short-drug light interval, in order to

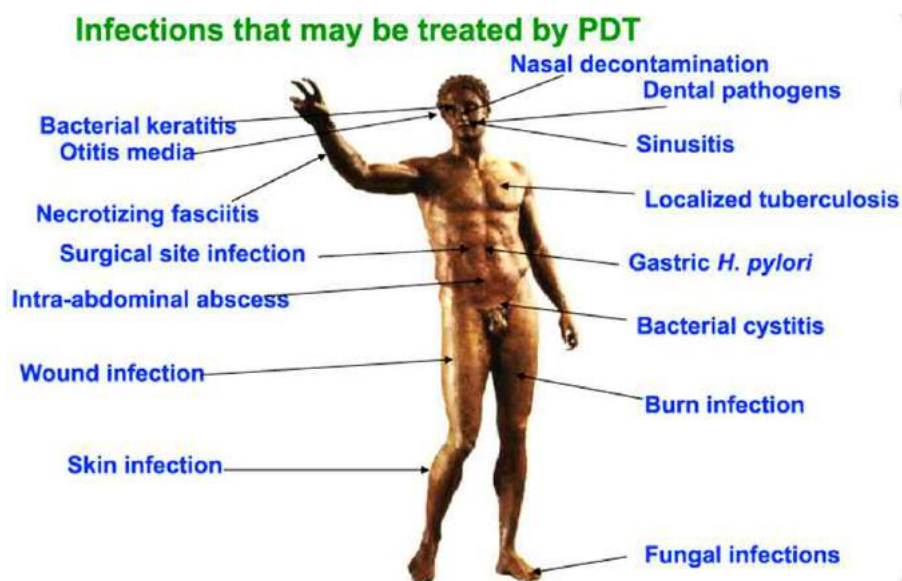


Fig. 5 Localized infections that could possibly be treated by PDT.

reduce the uptake of the PS by the surrounding host cells, and increase selectivity. However what is not often realized is the big difference between tumors and infections. In tumors all the surrounding tissue is considered to be undesirable, diseased and in need of removal. This is completely different to the case of an infection. An infection is defined as greater than 10(5) CFU of microbial cells per gram of tissue. The mass of 10(5) CFU of bacterial cells is approximately 1 µg. So the ratio between the mass of the bacterial tissue and the mass of the host tissue is in the region of 1 to 1,000,000 (Hamblin and Dai, 2010). So in sharp contrast to the case, where in tumors the goal is to destroy 100% of the surrounding tissue, in infections the goal is to destroy only 0.0001% of the surrounding tissue (just the bacterial cells). For this to be anyway near feasible the selectivity of the PS for bacteria has got to be exquisite so that 999,999 PS molecules out of every million bind to their target bacteria and not to the host cells. As can be imagined this is rather unlikely. However the situation is not as dire as the foregoing has made it out to be. The reason for this is that a short drug light interval is used for aPDI and even if the selectivity for bacteria over host cells is somewhat less than 1 million to 1, then the poly-cationic nature of the PS means that it is fairly slow to be taken up by host mammalian cells. So if light is delivered after say 30 s from topical delivery of the PS, it can be reasonably save to assume that not many of the surrounding host cells would be killed.

There have not been many reports of damage to host tissue being caused by aPDI in animal models. One study (Gad *et al.*, 2004) did use a model of a subcutaneous abscess caused by Gram-positive bioluminescent *Staphylococcus aureus* injected into the mouse thigh muscle. A small volume of PS solution was injected into the infected area and light was shone as a spot onto the surface. Two different PS were compared, the polycationic conjugate, poly-L-lysine chlorin(e6) (pL-ce6), and the anionic PS, free chlorin(e6). Both PS were equally effective in destroying the bacteria as judged by non-invasive bioluminescence imaging. However the non-bacterial targeted free chlorin(e6) caused more damage to the surrounding mouse thigh muscle as measured by a leg function test. The bacterial-targeted pL-ce6 was able to kill the bacteria without damaging the thigh muscle.

There has only been one report of activation of the immune system after PDT, when used against an infection in an animal model (Tanaka *et al.*, 2012). This involved injection of bioluminescent *S. aureus* into the mouse knee to produce a form of septic arthritis. When methylene blue was injected into the knee and 660 nm light delivered to the surface, the bacteria were killed in a similar manner to the experiment described above. However when the MB PDT was carried out before the bacteria were injected into the knee, rather than afterwards, an interesting observation was made. PDT carried out 24 h (but not 2 h) before infection protected the mice from developing a severe knee infection when the bacteria were subsequently injected, as shown by the fact that the infection resolved earlier (Tanaka *et al.*, 2012). This effect was dependent on neutrophils in the mouse knee as demonstrated by use of several blocking antibodies and inhibitors. Although there was no immediate influx of neutrophils into the mouse knee immediately after PDT, the neutrophils did appear to be “primed” to react quicker to attack the bacteria, once these were injected into the knee joint. Therefore, PDT of the mouse knee could be envisaged as a ‘non-specific vaccination procedure’, which could in principle be used in cases of orthopedic surgery, which might be considered high-risk for infection.

As mentioned above, for many years PDT was studied as a cancer therapy by designing PS that could be administered either systemically (by intravenous injection) or applied topically (e.g. aminolevulinic acid), and some time later the tumor would be irradiated with light (either as a surface spot or by insertion of interstitial fiber optics) (Agostinis *et al.*, 2011). However starting in the 1990s it was realized that PDT could also exert a powerful antimicrobial effect, if PS could be designed that could selectively bind to microbial cells, while not binding to host mammalian cells (Malik *et al.*, 1992). The best way of achieving this goal of antimicrobial photodynamic inactivation (aPDI) was to ensure that the PS had a pronounced cationic charge, as it was realized that microbial cells in general have a more pronounced negative charge compared to mammalian cells and cationic PS will bind selectively. Moreover the binding of the PS to the microbial cells is relatively rapid, while uptake of the cationic PS by mammalian cells is slow, thus giving good selectivity when a short drug-light interval (few minutes) is employed (Dai *et al.*, 2009). The advantages of aPDI as a potential clinical antimicrobial therapy were bolstered when it was realized that aPDI works equally well regardless of the antibiotic resistance status of the microbial cells (Vera *et al.*, 2012) and moreover, that aPDI has (so far) not been shown to produce resistance in bacteria (Maisch, 2015) even after 20 successive cycles of partial killing followed by regrowth (Giuliani *et al.*, 2010). Another advantage of aPDI is that the PS is applied topically or locally into the infected area. Many chronic infections involve a build up of microbial biofilms, into which it is now well-recognized, that systemically administered antibiotics fail to penetrate. However, aPDI has been shown to kill biofilm-grown cells both in vitro and in vivo (de Melo *et al.*, 2013). This anti-biofilm application has found particular application in dental infections such as periodontitis (Kikuchi *et al.*, 2015) and peri-implantitis (Vohra *et al.*, 2014). Moreover, infections in burns or damaged tissue suffer from a compromised blood supply, so systemically administered antibiotics fail to reach the site of infection in sufficient concentrations. The killing of microbial cells with aPDI is rapid (seconds) while the action of antibiotics can take hours or days, giving a potential advantage against fast-spreading infections such as necrotizing fasciitis. Moreover the broad-spectrum nature of aPDI means that treatment can be instituted before the infectious agents have been identified (Ferreyra *et al.*, 2016). Although many infections can occur deep inside the body, it is now possible to deliver both PS and light to almost any anatomical region, via endoscopes and narrow-diameter interstitially-inserted needles and fiber optics (Gad *et al.*, 2004).

PDT for Cancer

Cancer is still the most often-discussed indication for PDT and many papers have been published discussing modes of cell death in vitro (Fig. 6), and investigating pharmacokinetics and tumor response in vivo.

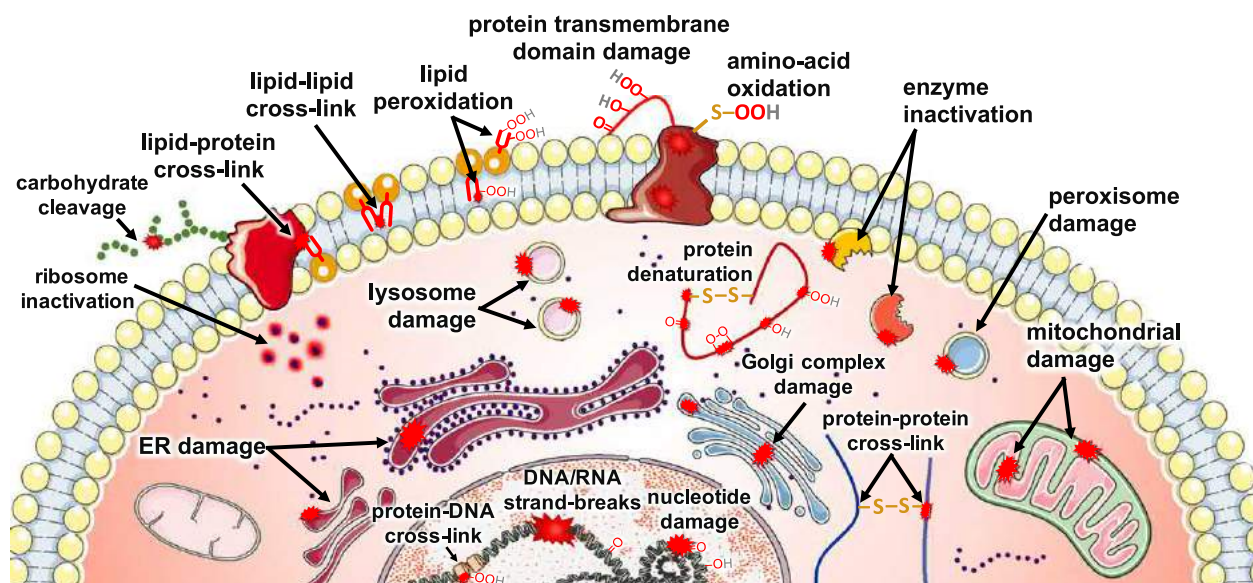


Fig. 6 PDT causes damage to biomolecules and intracellular organelles.

Cellular mechanisms of PDT

PDT and apoptosis

Apoptosis is a very complex, multi-step, multi-pathway cell-death program that is genetically encoded in every cell of the body (Pettigrew and Cotter, 2009). It can be initiated either through the activation of death receptors or the mitochondrial release of cytochrome c (Rustin, 2002). Both events eventually lead to activation of caspase cascades. Caspases are cysteine aspartic acid proteases and some members of the family are 'executioner caspases' such as caspase-3, -6 and -7 (Gougeon and Kroemer, 2003; Perfettini and Kroemer, 2003). The active executioner caspases cleave cellular substrates, which leads to characteristic biochemical and morphological changes observed in dying cells (Rathmell and Thompson, 1999). Chromatin condensation and nuclear shrinkage; cleavage of the inhibitor of the DNase CAD (caspase activated deoxyribonuclease) causes DNA fragmentation. Cleavage of cytoskeletal proteins leads to cell fragmentation and formation of apoptotic bodies (Savill and Fadok, 2000). The apoptotic process is tightly controlled by various proteins (Cotter, 2009; Okada and Mak, 2004). It is well known that resistance of tumor cells to apoptosis might be an essential feature of cancer development thus, modulation of the key elements of apoptosis signaling may directly influence tumor-cell death induced by various forms of cancer therapy (Igney and Krammer, 2002a,b).

PDT is thought to induce photodamage to Bcl-2 and related antiapoptotic proteins and activate the proapoptotic members of the family (Oleinick *et al.*, 2002). Alteration in expression of members of Bcl-2 family proteins following PDT has been reported in various cell lines and tumors. During apoptosis, the pro-apoptotic proteins Bax and Bak oligomerize in the mitochondrial outer membrane and probably disturb its integrity, leading to release of pro-apoptotic signaling molecules like cytochrome c, which then allows the activation of caspase 9 (Cory and Adams, 2002).

Accumulating evidence suggests that PDT is an efficient inducer of apoptosis in many cancer cell lines, whereas cells that lack the ability to efficiently undergo apoptosis are not protected from the lethal effects of PDT. Bax was not found to be essential for Photofrin-PDT induced apoptosis in human lung adenocarcinoma cells (ASTC-a-1) (Wu *et al.*, 2011). In the absence of Bax/ Bak, autophagy appears to be predominantly responsible for cell death following PDT with different photosensitizers (Buytaert *et al.*, 2006a,b; Kessel, 2006).

PDT produces a level of oxidative stress that can result in activation and translocation of the transcription factor NF- κ B to the nucleus, as originally shown for treatment of L1210 murine leukemia cells with Photofrin and light (Ryter and Gomer, 1993). NF- κ B provides an anti-apoptotic signal or cell survival pathway for cells exposed to PDT (Coupienne *et al.*, 2010).

PDT and necrosis

Necrotic cell death has been described as a violent and rapid form of degeneration affecting large fractions of cell populations, characterized by cytoplasmic swelling, destruction of organelles and disruption of the plasma membrane, leading to the release of intracellular contents and consequent inflammation (Danial and Korsmeyer, 2004).

Necrosis has been referred to as accidental cell death, caused by physical or chemical damage and has generally been considered an unprogrammed process (Kitsis and Molkentin, 2010). It is characterized by formation of a pyknotic nucleus, cytoplasmic swelling, and progressive disintegration of cytoplasmic membranes, all of which lead to cellular fragmentation and release of material into the extracellular compartment. In necrosis, decomposition is principally mediated by proteolytic activity, but the precise identities of proteases and their substrates are poorly defined (Proskuryakov and Gabai, 2010).

Studying the factors and parameters that cause cellular necrosis after PDT is not as easy as studying those factors which lead to apoptosis. The crucial factors in determining the type of cell death, e.g. apoptosis or necrosis following PDT are: the cell type, the presence of an intact set of apoptosis machinery, the subcellular localization of the PS, the light dose applied to activate it locally, and the oxygen partial pressure (Buytaert *et al.*, 2007). One factor that can be agreed upon by all commentators is that high dose PDT (either a high photosensitizer concentration or a high light fluence or both) tends to cause cell death by necrosis, while PDT administered at lower doses tend to predispose cells towards apoptotic cell death.

It is possible that high doses of PDT can photochemically inactivate essential enzymes and other components of the apoptotic cascade such as caspases. For instance, Lavie and coworkers (Lavie *et al.*, 1999) used the perylenequinones (hypericin and dimethyl tetrahydroxylanthrone) and found high dose PDT inhibited apoptosis by interfering with lamin phosphorylation, or by photodynamic cross-linking of lamins.

Xue *et al.* (2001) compared Pc4-mediated PDT of MCF7 cells that lack caspase 3 with the same cell line with caspase 3 transfected back in. They found apoptotic indicators only in the caspase expressing cells which also showed more loss of viability by an assay involving reduction of a tetrazolium dye; however both cell lines showed an equal degree of cytotoxicity by a clonogenic assay.

Dahle, Steele and Moan reported (Dahle *et al.*, 1999) that the mode of cell death induced by PDT depended on cell density. They seeded Madison Darby canine kidney II cells (MDCK II) at two different densities and incubated them with meso-tetra(4-sulfonatophenyl)porphine (TPPS4) for 18 h, washed and irradiated with blue light. Four hours later the cells were studied by fluorescence microscopy. With <55% total cell death the apoptotic fraction was significantly higher for cells in confluent monolayers than for cells growing in microcolonies at equitoxic doses. Confluent cells were 2.9 times more sensitive than cells in microcolonies partly due to a 1.5 times higher uptake of TPPS4 in monolayer cells. The difference in mode of cell death for the different cell densities was not related to any observable difference in subcellular localization pattern of TPPS4 at equitoxic doses of PDT.

PDT and autophagy

Autophagy is a catabolic cellular mechanism that allows the cell to maintain a balance between the synthesis, degradation, and recycling of cellular products (Mathew *et al.*, 2007). A variety of autophagic processes exist, all of which involve the lysosomal degradation of the cellular organelles and proteins. The most well-known mechanism proceeds in the following manner: A double membrane structure called autophagosome surrounds the target region, creating a vesicle that separates its contents from the rest of the cytoplasm. This vesicle is then transported and fused to the lysosome, forming a structure called the autophagolysosome, the contents of which are subsequently degraded by lysosomal hydrolases (Kroemer and Jaattela, 2005). Besides facilitating the disposal of unwanted proteins, organelles, and invading microorganisms, autophagy also allows a cell to reallocate its nutrients from unnecessary processes to life-essential ones in times of starvation or stress. The term “autophagy” was coined in 1963 at the Ciba Foundation Symposium on Lysosomes, where it was used to describe the presence of membrane vesicles that contain disintegrating organelles and cytoplasm components. Evidence soon confirmed that autophagy is an adaptive, energy-generating process. Molecular control of autophagy came into focus in the late 1990s. Signaling pathways governing this process became better understood after the identification of the target of rapamycin kinase (TOR), which controls cell growth and protein synthesis (Bjornsti and Houghton, 2004).

It is still unclear exactly how autophagy affects the outcome of PDT (Kessel, 2006; Kessel *et al.*, 2006; Reiners *et al.*, 2010). In general, mammalian cells use autophagy as a defense against ROS mediated damage by clearing the cell of damaged organelles (Scherz-Shouval *et al.*, 2007). Depending on the type of ROS and degree of oxidative injury, PDT may stimulate autophagy (Sasnauskiene *et al.*, 2009) that either acts in a cytoprotective manner or induces autophagic cell death (Scherz-Shouval and Elazar, 2007). Autophagy may play a role in PDT induced apoptosis, but the two processes can also occur independently of one another (Kessel and Oleinick, 2009). A study done on murine leukemia L1210 cells found that a wave of autophagy occurs right before apoptosis (Kessel and Arroyo, 2007). It was also found that prevention of autophagy by silencing the Agt7 gene allowed photo-killing to occur at lower light doses. This observation is consistent with the theory that autophagy is a defense mechanism against PDT induced ROS (Kessel and Reiners, 2007). The situation is different for tumor cells that do not have the ability to undergo apoptosis through a deficiency in Bax or Bac, which regulate the apoptotic pathway. In these cells, PDT induced autophagy stimulates a necrotic, caspase-independent cell death (Buytaert *et al.*, 2006a). Suppression of autophagy in apoptosis-deficient cells resulted in the inhibition of cell death during PDT. In general, the induction of autophagy in PDT treated cells occurs independently of an apoptotic outcome. While autophagy seems to play a pro-survival role in tumor cells that are capable of apoptosis, it has been shown to promote death in cells that are apoptosis-deficient.

PS pharmacokinetics

When PS preparations are injected into the bloodstream they tend to bind to various serum proteins and this binding can dramatically affect their pharmacokinetics and biodistribution (Boyle and Dolphin, 1996). Different PS can have very different pharmacokinetics and this can directly affect the illumination parameters and particularly the pharmacokinetics will determine the drug-light interval (Agostinis *et al.*, 2011). Intravenously injected PS undergo a gradual transition from being bound to serum proteins, then becoming bound to endothelial cells, then they are bound to the adventitia of the vessels, then bound either to the extracellular matrix or to the cells within the tumor, and eventually they will be cleared from the tumor tissue by removal by lymphatic vessels or by blood vessels, and they will be excreted either by the kidneys or the liver. The effect of PDT on the tumor largely depends at which

stage of this continuous process light is delivered. The anti-tumor effects of PDT are divided into three main mechanisms. Direct tumor cell death by apoptosis or necrosis can occur, if the PS has been allowed to be taken up by tumor cells. Powerful anti-vascular effects can lead to thrombosis and hemorrhage in tumor blood vessels that subsequently leads to shut-down of tumor blood vessels, which kills the tumor by deprivation of oxygen and nutrients. Finally the release of damage associated molecular patterns (DAMPs) leads to acute inflammation and the release of cytokines (Garg *et al.*, 2011). The DAMPs, cytokines and stress response proteins induced in the tumor by PDT work together and lead to an influx of leukocytes that can both contribute to tumor destruction as well as to stimulate the immune system to recognize and destroy tumor cells even at distant locations.

When PS are injected into the bloodstream a certain sequence of events commences that can take very different lengths of time to reach completion, for PS with different chemical structures. Firstly, depending on the delivery solvent or vehicle that is used for the PS injection, the PS must come to equilibrium with the components of the circulating blood. This can involve the PS disaggregating from itself being released from its delivery vehicle and binding instead to various protein components of serum (see later). In addition the blood has many circulating cells (erythrocytes and leukocytes) that would in principle be available to bind injected PS (Maugain *et al.*, 2004). Secondly, the circulating PS must bind to the walls of the blood vessels, and it is thought that the nature of the various blood vessels in the tumor and in surrounding normal tissues that vary significantly in size, speed of blood flow, and physiological characteristics such as leakiness and tortuosity in tumors and various normal organs, governs to a great extent where the PS localizes or accumulates. Thirdly, the PS will penetrate through the wall of the blood vessel at a rate that probably depends on how strong the initial binding of the PS to the intimal surface was. PS that bind strongly to the blood vessel wall will take a longer time to cross the entire wall of the blood vessel, and those that have an initial weaker binding will pass through more quickly. Fourthly, after extravasation, the PS will diffuse throughout the parenchyma of the organ or tumor to which it has been delivered. If the organ happens to be the liver or other metabolically active organ, the PS may be subject to chemical changes by metabolic enzymes, but this is thought to be rather unlikely for most tetrapyrrole-based PS in common clinical use. Fifthly, the PS will eventually be eliminated from the tissue. Although this elimination has not been much studied, if the tissue containing the PS is a tumor or other non-metabolic organ, elimination will probably be by lymphatic drainage and the PS will then return into the general blood circulation via the thoracic duct. Sixthly, the PS will be excreted from the body. For the majority of PS in clinical use, excretion is from the liver into the bile and thence to the intestine where it is lost via fecal elimination. However entero-hepatic re-circulation, which is the term used for the re-absorption of the PS from the small intestine, back into the blood stream, has been reported for some PS.

Measurement of pharmacokinetics requires sequential measurements of PS concentration in tissue or other biological substance or fluid in order to construct temporal profiles that can be compared for different PS. This is frequently done for blood where samples can be readily removed in small amounts at frequent intervals and analysis then yields values of serum half-lives (first or second order) in minutes, hours or days. Alternatively, analysis of PS concentrations in urine or fecal samples can give terminal elimination half-lives. All these methods require a method of quantitatively analyzing PS concentrations in samples of biological material. The fact that the majority of PS used for PDT are also fluorescent, has led to the development of assays involving homogenization or dissolving the tissue or biological fluid and measurement of fluorescence in solution after centrifugation or extraction steps (Bellnier *et al.*, 1997). Calibration curves can be prepared with mixtures of known amounts of PS and weighed amounts of biological material. Care must be taken when these experiments are carried out on small rodents such as mice and rats because these animals accumulate a naturally red-fluorescent compound derived from chlorophyll in their diet that ends up in the skin and the digestive tract amongst other organs (Gudgin Dickson *et al.*, 1995). This can be avoided by feeding the animals a chlorophyll-free diet for enough time before the experiment to allow the interfering fluorescent compound to be eliminated (Allison *et al.*, 1994). In some cases the PS is quantified by a traditional analytical chemistry technique such as high performance liquid chromatography (Almeida-Lopes *et al.*, 2001). This method has the advantage of detecting possible metabolism products that have been formed from the originally pure PS after injection. Finally in some cases of experimental animal models, the PS has been presynthesized in the laboratory to include a radioisotope label (for instance carbon-14 or tritium) or a chelator can be attached that can bind a metal radioisotope and biological samples can then be quantified using a scintillation counter (Bellnier *et al.*, 1989; Little *et al.*, 1984; Schuitmaker *et al.*, 1993). An alternative approach to carrying out analysis of samples of biological material removed sequentially, is to use a continuous optical monitoring method to determine PS concentration in tissue based on fluorescence. In this approach a non-invasive or minimally invasive approach with fiber-based fluorescence detection systems can be used to measure the variation over time of a fluorescence signal linked to the PS of interest (Frisoli *et al.*, 1993). In some small animal models the skin can be removed from a subcutaneous tumor so the fiber that collects fluorescence can be in contact with the underlying tumor as well as the adjacent skin (Sheng *et al.*, 2004).

PS biodistribution

In experimental animals it has sometimes been possible to establish the overall pattern of distribution in the different organs after intravenous injection of the PS. A study (Bellnier *et al.*, 1989) using radioisotope labeled Photofrin in mice with subcutaneous fibrosarcoma tumors that were sacrificed 24h later found highest concentrations in liver, adrenal glands, and urinary bladder. Next was pancreas, kidney, spleen, and these organs were higher than stomach, bone, lung, and heart. Muscle concentration was low and brain was lowest. Only skeletal muscle, brain, and skin located distant to the tumor had peak concentrations lower than tumor tissue; the skin overlying the tumors showed concentrations not significantly different from tumor. The higher molecular weight components of Photofrin were partially retained in the liver and spleen at 75 days after injection.

In most cases where organ biodistribution data have been reported for small animals such as mice and rats, the highest concentration of PS has been found in the liver (Bertoloni *et al.*, 1990; Bhargava *et al.*, 2012). Chan *et al.* studied a set of sulfonated phthalocyanines found that the accumulation in liver was inversely proportional to the degree of sulfonation and hence to the overall lipophilicity of the PS molecules (Bhat *et al.*, 2005). The liver is known to have a highly permeable blood supply with fenestrated endothelium that contains pores that allow molecules to pass easily out of the blood vessels. This porous vasculature is required by the role of the liver in detoxifying the body of extraneous molecules (particularly organic molecules such as PS). These molecules are then excreted into the bile via the gall bladder in a similar fashion to the production of bile acids from cholesterol. The bile is routed into the duodenum where it is possible for unchanged PS to be absorbed into the circulation a second time from the small intestine (a process known as entero-hepatic recycling) as well as being excreted via the feces. It is known that soon after injection, large amounts of PS can accumulate in the lungs. Lung accumulation has been reported (Brun *et al.*, 2004) as the mode of dose-limiting dark toxicity when the PS injected doses were increased eventually leading to lung hemorrhage and acute interstitial pneumonia after injection of a high dose of 80 mg/kg of Pc4 into mice. The explanation is probably that large injected doses of PS can aggregate in the blood vessels and the particles so formed can easily collect in the fine capillary network of the lungs.

The amount of PS that accumulates in the spleen seems to vary widely with the structure of the PS (overall charge and hydrophobicity). Woodburn *et al.* (1992) studied a range of porphyrins with varying octanol/water partition coefficients and found the accumulation in the spleen varied the most compared to other organs. Egorin *et al.* (1999) compared organ distribution of Pc4 administered by IV injection of the same dose (10 mg/kg) dissolved in several solvents and found that the accumulation in the spleen was the most variable.

The kidney and bladder frequently accumulate high amounts of PS, despite the clearance route being almost exclusively via liver and bile (Richter *et al.*, 1990) rather than the kidneys and urine. The digestive organs (stomach, large and small intestines) seem to take up middling amounts of PS (i.e. less than liver but more than muscle) (Schuitemaker *et al.*, 1993). The skin has not been found to accumulate high amounts of PS in experimental animals, which might be considered somewhat surprising in view of the fact, that skin photosensitivity due to PS accumulation in patients' skin is the main cause of PDT-related side-effects. The lowest concentrations of PS are generally found in organs such as heart, skeletal muscle, bone, eye and brain. These are organs that are known to have a relatively impermeable blood supply. The blood-brain barrier serves to exclude PS from the parenchyma of the brain, while the fact that the blood-brain barrier is frequently breached by the growth of brain tumors is probably responsible for the very large (over 100) tumor-to-normal brain ratios reported for some PS (Boyle and Dolphin, 1996).

Since the original observation of the tumor localizing ability of hematoporphyrin derivative (HPD), many workers have investigated the underlying mechanism of this effect whereby PS preferentially localize in tumors, and in other specific organs and anatomical locations (Hamblin and Newman, 1994; Jori, 1989; Larroque *et al.*, 1996). The precise definition of "tumor to normal tissue ratio" has also been subject to debate. Some investigators working with subcutaneous tumors in experimental animals use the ratio between the tumor and peritumoral muscle or skin, while others use the ratio between the tumor to distant muscle and skin. Furthermore, there has been much effort made to determine which molecular features in the chemical structures of the PS are optimal for maximizing the selectivity for the tumor over normal tissue and organs. This has proved quite complicated because the pharmacokinetics of different PS can vary dramatically. For instance, one PS can have its best tumor to normal tissue ratio at a relatively early time point after administration such as 3 h, while for another PS this time point can be as long as 7 days after injection. One of the notable properties of these tetrapyrrole molecules that are commonly employed as PS, that may be relevant to their tumor localizing ability, is their tendency to bind strongly to serum proteins and also to bind to each other (aggregation). This high tendency to bind means that most PS when injected into the bloodstream behave as macromolecules rather than as small molecules, either because they are more or less firmly bound to large protein molecules or because they have formed intermolecular aggregates of similar size. Many workers have reported (Kessel and Poretz, 2000; Kongshaug *et al.*, 1989; Maziere *et al.*, 1990) on the distribution of PS between the various classes of serum proteins when mixed with serum (either human or animal) *in vitro*. These serum proteins are usually divided into four classes: (a) albumin and other heavy proteins; (b) high density lipoprotein (HDL); (c) low density lipoprotein (LDL); and (d) very low density lipoprotein (VLDL). However even this study has been complicated by the fact that the most lipophilic PS are insoluble in aqueous media and need to be delivered in a solvent mixture which may alter the serum protein distribution of the PS. It has been argued that PS that preferentially bind to LDL are better tumor localizers (see below) (Jori and Reddi, 1993) but this is by no means always the case (Korbelik, 1992).

It is useful to make a distinction between selective accumulation and selective retention. The tumor localizing ability of the PS with the faster pharmacokinetics is probably due to selective accumulation in the tumor, while the localization of PS with slower acting pharmacokinetics is more likely due to selective retention. In the selective accumulation model it is thought that the increased vascular permeability to macromolecules typical of tumor neovasculature is chiefly responsible for the preferential extravasation of the PS. These quick-acting PS frequently bind to albumin, which is of ideal size and Stokes radius to pass through the "pores" in the endothelium of the tumor micro-vessels (Yuan *et al.*, 1993). The selective retention model where PS can be retained in tumors, while they are eliminated from surrounding normal tissue and organs, has been the subject of much speculation. As mentioned above, a popular theory maintains that the binding of the PS to LDL is of major importance (Jori and Reddi, 1993). According to this theory, it is proposed that cancer cells over-express the LDL (apoB/E) receptor. Up-regulation of the expression of LDL receptors is one strategy that rapidly growing malignant cells have developed to obtain sufficient cholesterol that the tumor cells need for the biosynthesis of lipids required due to the rapid turnover of their cellular membranes. There is experimental evidence both for and against this LDL receptor theory (Allison *et al.*, 1994; Korbelik, 1992). Other theories have been proposed to account for the selective retention of PS in tumor tissue. One is that tumors are characterized by poorly

developed lymphatic drainage, and that macromolecules, which extravasate from the hyperpermeable tumor neovasculature, are retained in the extravascular space (Buytaert *et al.*, 2006b). This theory has been termed the well known “enhanced permeability and retention” (EPR) effect (Maeda *et al.*, 2016). Another theory is attributed to macrophages, which are phagocytic cells which infiltrate solid tumors to varying extents (Blanco *et al.*, 2015). These tumor-associated macrophages (TAMs) have been shown to accumulate up to thirteen times the amount of some PS compared to the cancer cells themselves (Buytaert *et al.*, 2007). The explanation for this preferential uptake by TAMs has been proposed to be due to either the phagocytosis of aggregates of PS (Canti *et al.*, 1994), or to the preferential uptake by TAMs of lipoproteins which have been subtly altered by the binding of porphyrins (Blanco *et al.*, 2015). Another theory proposes that the low pH commonly found in tumors (due to the Warburg effect in which the metabolism changes from oxidative phosphorylation to glycolysis, and produces large amounts of lactic acid) has the effect of trapping some of the anionic PS, which are normally ionized at normal physiological pH. These PS then become neutrally charged at low pH and hence more lipophilic, when they encounter the acidic pH in the tumor environment (Pottier and Kennedy, 1990). Yet another proposed mechanism is the selective accumulation of PS in tumors is related to the propensity of tumors to accumulate lipid droplets which may contain lipophilic molecules (Freitas, 1990). A final theory concerns the tendency of some tumors to have very high populations of tumor-associated macrophages, that may engulf particles of aggregated PS (Korbelik and Kros, 1995), or else take up PS that have bound to LDL, and then altered its behavior to be recognized by scavenger receptors (Hamblin and Newman, 1994).

Direct destruction of tumors by PDT

Many of the studies relating to this topic have been concerned with attempts to distinguish between direct tumor cell killing *in vivo* as described in the previous chapter, and indirect tumor cell killing caused by the vascular effects of PDT (see later).

Some studies have investigated an experimental model in which tumors in mice that had been treated with a potentially curative dose of PDT *in vivo* were removed from the mice, and then dissociated into a suspension of individual tumor cells. These tumor cells were still able to form colonies in a clonogenic assay if the tumors were removed immediately from the host after PDT (Henderson and Fingar, 1989; Henderson *et al.*, 1985). However when these tumors were left in the host animal for increasing lengths of time, a progressive loss in clonogenicity of the dissociated tumor cells was observed. These times corresponded to the emergence of PDT-induced tumor hypoxia as determined radiologically. These studies suggested a central role for vascular damage in governing the tumor response to PDT in mouse models.

Anti-vascular effect of PDT

Photodynamic alteration of tissue microcirculation was first reported in 1963 (Castellani *et al.*, 1963). A study by Star *et al.* (1986) utilized a window chamber in rats bearing an implanted mammary tumor, to make direct observations of the microcirculation of the tumor and the adjacent normal tissue before, during, and at various times after PDT mediated by HPD. An initial slowing down of blood flow and vasoconstriction of the tumor vessels was followed by heterogeneous responses including eventual complete cessation of blood flow, hemorrhage, and in some larger vessels, the formation of platelet aggregates. Other workers have histologically examined tissues removed from animal models that had been treated with PDT such as subcutaneous urothelial tumors and normal rat jejunum and reported a range of vascular responses including disruption of blood flow, breakdown of the blood brain barrier in the normal brain of mice, and endothelial cell damage in subcutaneous tumors and normal tissue (Bhuvaneswari *et al.*, 2009; Tseng *et al.*, 1988). Many reports cited above directly implicate the endothelium as a primary target for PDT *in vivo*; this stimulated research into the relative sensitivity of endothelial cells to PDT and the responses of endothelial cells that could initiate the various phenomena at the vessel level. Gomer *et al.* (1988) showed that bovine endothelial cells were significantly more sensitive to Photofrin-mediated PDT than smooth muscle cells or fibroblasts from the same species. This increased sensitivity to PDT, measured by clonogenic assay, was not a result of increased Photofrin accumulation. Sensitivity to HPD-mediated PDT of bovine aorta endothelial cells and human colon adenocarcinoma cells was investigated by West *et al.* (1990). Exponentially growing endothelial cells were significantly more sensitive than tumor cells with a similar rate of proliferation, and the difference in sensitivity was accompanied by greater PS accumulation in the endothelial cells.

Intravital microscopy of the rat cremaster muscle showed that Photofrin-PDT caused vessel constriction with an increase in vessel permeability and leukocyte adhesion was observed (Fingar *et al.*, 1992). The cyclooxygenase inhibitor indomethacin was able to inhibit the vessel constriction and reduce the increase in vascular permeability, implicating the involvement of the metabolites of cyclooxygenase called eicosanoids (such as thromboxane or prostacyclin). An increase in vascular permeability, thrombus formation and blood flow reduction were observed in animals treated with verteporfin-PDT in a dose- and time-dependent manner (Debefe *et al.*, 2010; Fingar *et al.*, 1999; He *et al.*, 2008). Blood vessels with lower flow rates were more sensitive to verteporfin-PDT-induced vascular shutdown than vessels with a higher flow rate. In agreement with light and electron microscopic studies, intravital fluorescence microscopic imaging demonstrated in live animals that verteporfin-PDT decreased endothelial cell viability and caused platelet aggregation (Khurana *et al.*, 2008). Rapid shutdown of tumor blood flow by formation of blood clots was also found after PDT with Tookad and MV6401, which was also confirmed by histological assessment (Dolmans *et al.*, 2002; Madar-Balakirski *et al.*, 2010). Many optical imaging studies have led to the conclusion that PDT is essentially a vascular disrupting therapy.

Immune stimulation effect of PDT: Innate immunity

PDT frequently provokes a strong acute inflammatory reaction in the targeted site, and this can often be observed as a fairly rapid induction of localized edema (Dougherty *et al.*, 1998). This reaction is a consequence of PDT-induced oxidative stress. Thus, PDT can be ranked among cancer therapies (including cryotherapy, hyperthermia, and focused ultrasound ablation) that all produce a chemical or physical destruction of tumor tissue that is perceived by the host as localized acute trauma. This tissue damage prompts the host to launch protective actions designed by evolution to react against threats to tissue integrity and to try to restore homeostasis at the affected site (Korbelik, 2006). The acute inflammatory response is the principal process engaged in this context. Its main task is containing the disruption of homeostasis, ensuring removal of damaged cells, and to protect against possible invasion of pathogenic microorganisms that may gain entry to the body as a result of a traumatic breach in external barriers. The inflammatory response also contributes to the local healing process with restoration of normal tissue function. The inflammation elicited by PDT is orchestrated by the innate immune system (Korbelik, 2006). The principal cells involved in this process consist of neutrophils, monocyte/macrophages, dendritic cells and mast cells. The recognition systems of the innate immune system are based upon a wide array of pattern-recognition receptors (PRRs), which are responsible for detecting the presence of tissue damage revealed to its sensors as the appearance of “altered-self” (Korbelik, 2006). PDT appears particularly effective in generating rapidly an abundance of alarm/danger signals, also called damage-associated molecular patterns (DAMPs) at the treated site that can be detected by the innate immune cells (Korbelik, 2006). Many DAMPs are nuclear or cytosolic proteins, or even small molecules that are normally safely contained inside healthy cells. When released outside from damaged or necrotic cells, or exposed on the surface of the cell following damage, they can encounter an oxidizing environment, which results in their changes in their conformation or chemical structure (Garg *et al.*, 2015). DNA can be released outside the nucleus from necrotic cells, and becomes a DAMP (Magna and Pisetsky, 2016). Other DAMPs (Garg *et al.*, 2015) include heat-shock proteins, high mobility group box protein 1 (HMGB1), hyaluronan fragments, extracellular ATP, uric acid, heparin sulfate and S100 molecules (members of a family of calcium modulated proteins).

Dendritic cells (DCs) are professional antigen presenting cells that have been described as the “orchestrators” of the immune response (Bhargava *et al.*, 2012). DCs express many receptors that specifically recognize DAMPs. Since PDT of tumor cells causes both cell death and cell stress (Dougherty *et al.*, 1998; Henderson and Gollnick, 2003; Oleinick and Evans, 1998) it is hypothesized that the activation of DCs by PDT-treated cells is the result of recognition of DAMPs released/secreted/exposed by PDT from dying cells (Gollnick *et al.*, 2006; Korbelik *et al.*, 2005, 2007). HSP70 is a well-characterized DAMP that interacts with the danger signal receptors, TLRs (Toll-like receptors) 2 and 4 (Vabulas *et al.*, 2002) and is induced by PDT (Gomer *et al.*, 1996). The level of expression of HSP70 in PDT-treated tumor cells appears to correlate with an ability to stimulate DC maturation (Gollnick *et al.*, 2004) and initiation of inflammation (Korbelik *et al.*, 2005; Stott and Korbelik, 2007).

The onset of PDT-induced inflammation is marked by dramatic changes in the tumor vasculature that becomes permeable to serum proteins and expresses adhesion molecules that circulating inflammatory cells can recognize and to which they can adhere (Korbelik, 2006). This process is responsible for the well-known “rolling” behavior of neutrophils and other leukocytes. The inflammatory process is predominantly initiated by signals originating from photooxidative damage produced in perivascular regions of the tumor with chemotactic gradients reaching out from the damaged endothelium. The inflammatory cells, initially consisting of neutrophils, which are then followed by mast cells and monocytes/macrophages, rapidly and massively invade tumors after PDT (Fig. 7) (Dougherty *et al.*, 1998; Krosi *et al.*, 1995). The primary task of these infiltrating cells is to neutralize the source of DAMPs by removing debris containing injured and dead cells and oxidatively damaged tissue components. Damage and dysfunction of PDT-treated tumor vasculature frequently results in vascular occlusion that serves to “seal off” the damaged tumor tissue until it is removed by invading phagocytic cells (Korbelik, 2006). Depletion of neutrophils or inhibition of their activity after PDT was shown to diminish the therapeutic effect (de Vree *et al.*, 1997; Korbelik and Cecic, 1999, 2003; Kousis *et al.*, 2007). Among the cytokines involved in the regulation of the inflammatory process, the most important ones for the PDT response are IL-1 β and IL-6 (Gollnick *et al.*, 2003; Sun *et al.*, 2002). Blocking the function of various adhesion molecules was also shown to inhibit the PDT response (Gollnick *et al.*, 2003; Sun *et al.*, 2002). On the other hand, blocking anti-inflammatory cytokines such as IL-10 and TGF- β can significantly improve the cure rates after PDT (Korbelik, 2006).

Fig. 7 illustrates the stimulation of immune response after PDT.

Immune stimulation effect of PDT: Adaptive immunity

A wide range of pre-clinical animal studies and even some clinical studies have demonstrated that PDT can influence the adaptive immune response in disparate ways. Some PDT regimens result in potentiation of adaptive immunity, while others lead to immunosuppression. The precise reasons why PDT can lead to potentiation vs. suppression are unclear; however it appears as though the effect of PDT on the immune system is dependent upon the treatment regimen, as well as the area treated and the precise PS type (Hunt and Levy, 1998; Kousis *et al.*, 2007). PDT induced immune suppression is largely confined to cutaneous and transdermal PDT regimens involving large surface areas (Hunt and Levy, 1998; Yusuf *et al.*, 2008).

Broadly speaking PDT efficacy appears to be dependent upon the induction of adaptive anti-tumor immunity. Long-term tumor response and cures are diminished or absent in immunocompromised mice such as nude mice or SCID mice (Korbelik and Cecic, 1999; Korbelik *et al.*, 1996). Reconstitution of these animals by injecting bone marrow or T cells isolated from immunocompetent (normal) mice can result in increased PDT efficacy. Canti *et al.* (1994) were the first to show PDT-induced immune stimulation, demonstrating that cells isolated from tumor-draining lymph nodes of PDT-treated mice were able to confer tumor

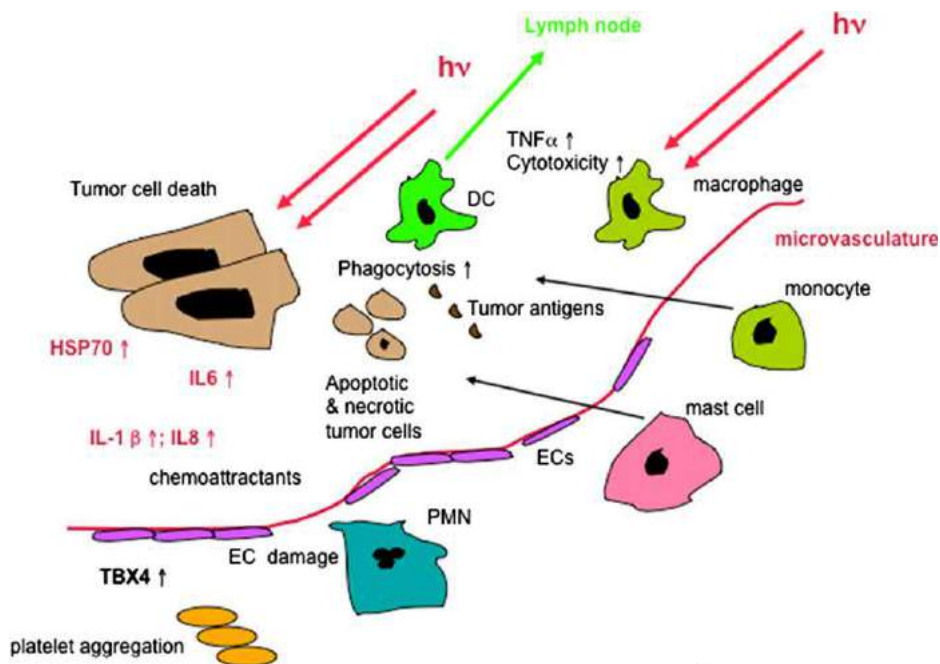


Fig. 7 Stimulation of immune response after PDT of tumors.

resistance when injected into naïve mice. Subsequent studies demonstrated that PDT directed against murine tumors resulted in the generation of immune memory such as rejection of a tumor rechallenge in mice that had obtained a long-term cure after PDT (Korbelik and Dougherty, 1999).

In order for the adaptive immune response produced by PDT to be highly effective, it appears that the tumor cell line employed should express what has been called a “tumor rejection antigen”. This has been demonstrated by testing PDT of syngeneic mouse tumors that have been engineered to express a model tumor antigen in immunocompetent mice. One of these model antigens was the bacterial protein beta-galactosidase (b-gal) (Mroz *et al.*, 2010). Although the b-gal tumors grew as rapidly as their wild-type counterparts, there was 100% long-term cures after BPD-mediated PDT with the b-gal tumors but no long-term cures with the wild-type tumors. The strength of the anti-tumor immune response was demonstrated by the fact that if mice had two b-gal tumors (one in each leg) and only one was treated with PDT, the contralateral untreated tumor underwent spontaneous regression in 70% of the mice. A similar result was seen in mice bearing the mastocytoma P815 that is well known to express the naturally-occurring tumor rejection antigen called P1A (Mroz *et al.*, 2013). The CD8 cytotoxic T-cells can recognize specific peptide epitopes derived from the amino acid sequence of the tumor antigen.

MHC-I is critical for allowing the CD8⁺ T cells to recognize the tumor antigen, and tumors that lack MHC-I are resistant to cell-mediated anti-tumor immune reactions (Maeurer *et al.*, 1996). Since MHC1 expression and expression of tumor associated antigens (such as P1A) are both susceptible to down-regulation by epigenetic silencing, an epigenetic reversal agent (the demethylating compound, 5-aza-2'-deoxycytidine) could potentiate the induction of anti-tumor immunity and increase cures in several mouse tumor models (Wachowska *et al.*, 2014).

The mechanism whereby PDT enhances anti-tumor immunity has been examined for the past several decades. PDT activates both humoral and cell-mediated anti-tumor immunity, although the importance of the humoral response (antibodies) is at present unclear. PDT efficacy in mice and humans is reduced in the absence of CD8⁺ T cell activation and/or tumor infiltration (Abdel-Hady *et al.*, 2001; Kabingu *et al.*, 2007; Korbelik and Cecic, 1999). Therefore most mechanistic studies have focused on the means by which PDT potentiates CD8⁺ T cell activation. It is clear that induction of anti-tumor immunity following PDT is dependent upon induction of inflammation (Henderson *et al.*, 2004). PDT-induced acute local and systemic inflammation is postulated to culminate in the maturation and activation of dendritic cells (DCs). Mature DCs are critical for activation of tumor specific CD8⁺ T cells and induction of anti-tumor immunity (Reis e Sousa, 2004). DCs are activated in response to PDT (Gollnick *et al.*, 2003) and migrate to tumor draining lymph nodes where they present the antigens they have taken up from damaged tumor cells to naïve T cells in the so-called “immune synapse” (Benvenuti, 2016). The T-cells in the lymph node become activated and proliferate dramatically and a “flood” of tumor-specific CTLs is released into the blood stream where they can search for other tumors cells to attack and destroy (Gollnick *et al.*, 2003; Sur *et al.*, 2008). Generation of CD8⁺ effector and memory T cells is frequently, but not always dependent upon the presence and activation of CD4⁺ T cells (Castellino and Germain, 2006). PDT induced anti-tumor immunity may (Korbelik and Cecic, 1999) or may not depend on CD4⁺ T cells (Kabingu *et al.*, 2007) and may be augmented by natural killer (NK) cells (Kabingu *et al.*, 2007).

PDT-mediated enhancement of anti-tumor immunity is believed to be due, at least in part, to stimulation of DCs by dead and dying tumor cells, suggesting that in vitro PDT-treated tumor cells may act as effective anti-tumor vaccines (Gollnick *et al.*, 2002).

This hypothesis has been proven by several studies using a wide variety of different PS and tumor models in both preventative and therapeutic settings (Gollnick *et al.*, 2002; Korbelyik and Cecic, 2003; Korbelyik *et al.*, 2009; Korbelyik and Sun, 2006).

Mechanistic studies showed that incubation of immature DCs with PDT-treated tumor cells leads to enhanced DC maturation, activation and increased ability to stimulate T cells (Gollnick *et al.*, 2002; Jalili *et al.*, 2004). Furthermore, opsonization of PDT-treated tumor cells by complement proteins increases the efficacy of PDT-generated vaccines (Korbelyik and Sun, 2006).

Induction of anti-tumor immune response after PDT

How does an effective anti-tumor immune response triggered after PDT manifest itself? This depends on whether pre-clinical animal tumor models (most often mice or rats) are being studied, or whether clinical studies are being considered. In an ideal scenario the immune stimulation would be strong enough for an animal or a patient with metastatic cancer who had the primary tumor treated with PDT, and then to have the distant untreated metastases regress over time as the host T-cells recognize specific tumor antigens and then attack and destroy their cognate target cells.

The implications of PDT-induced anti-tumor immunity and efficacious PDT-generated vaccines are significant and provide an exciting possibility for using PDT in the treatment of metastatic disease and as an adjuvant in combination with other cancer modalities. Several pre-clinical studies demonstrated that PDT is able to control the growth of tumors present outside the treatment field (Gomer *et al.*, 1987; Kabingu *et al.*, 2007) although others have failed to demonstrate control of distant disease following PDT (van Duijnhoven *et al.*, 2003).

This ideal scenario was demonstrated (Mroz *et al.*, 2010) in a mouse model where a BALB/c syngeneic colon carcinoma called CT26 was implanted subcutaneously in the leg and treated by a vascular PDT regimen mediated by liposomal verteporfin and 690 nm light. Although an apparently complete local response was achieved by application of PDT, the mice all suffered a local recurrence of the tumor in the leg and were sacrificed within a few weeks. By contrast, when the CT26 tumor was engineered to express a model tumor antigen (β -galactosidase) to form a tumor cell line called CT26. CL25, the tumors grew in the mice the same as CT26 wild-type. However CT26. CL25 tumors gave 100% permanent cures after PDT, and the cured mice were resistant to a rechallenge with CT26. CL25 tumor cells (but not to CT26 cells). When mice with two bilateral CT26. CL25 tumors (1 in each leg) had only the right-leg tumor treated with PDT, the distant untreated left-leg tumor spontaneously regressed over a period of several weeks. In about 70% the regression of both tumors was permanent, while in the remaining 30% of mice where the untreated CT26. CL25 tumor was observed to recur, it was found that the recurrent tumor has lost expression of the β -galactosidase antigen. Loss of antigen expression is a recognized route by which highly plastic tumors can evade the host immune response.

The efficacy of clinical PDT also appears to depend on induction of anti-tumor immunity. Patients with vulval intraepithelial neoplasia (VIN) who did not respond to ALA-PDT were more likely to have tumors that lacked major histocompatibility complex class I molecules (MHC-I) than patients who responded to ALA-PDT (Abdel-Hady *et al.*, 2001). VIN patients who responded to PDT had increased CD8⁺ T cell infiltration into the treated tumors as compared to non-responders. In other clinical studies patients who were treated with PDT for actinic keratoses and Bowen's disease were divided into immunocompetent and immunosuppressed sub-groups. Both sub-groups had similar initial response rates to PDT, but immunosuppressed patients had less long-term responses and were more likely to suffer appearance of new lesions (Dragieva *et al.*, 2004). In a recent clinical study, PDT of multifocal angiosarcoma of the head and neck resulted in increased immune cell infiltration into distant untreated tumors that was accompanied by tumor regression (Thong *et al.*, 2007, 2008). PDT was also shown to be an effective surgical adjuvant in non-small-cell lung cancer patients with pleural spread (Friedberg *et al.*, 2004). Recent reports have shown that clinical anti-tumor PDT also increases anti-tumor immunity. PDT of basal cell carcinoma (BCC) increased immune cell reactivity against a BCC-associated antigen (Kabingu *et al.*, 2009).

Photobiomodulation – The Healer

Light alone when it is delivered at a relatively low irradiance and fluence is called low level light (laser) therapy (LLLT) and has almost the complete opposite effect to PDT. The internationally accepted terminology for LLLT has recently been changed to "photobiomodulation" (PBM) therapy, due to widespread uncertainties about just exactly what parameters constituted "low level" (Anders *et al.*, 2015). LLLT (or PBM) was discovered in a rather serendipitous manner by Endre Mester in the 1960s. The discovery of the laser in 1960 by Theodore Maiman, had been recognized as a landmark in modern physics, and garnered an enormous amount of international attention (Maiman, 1960). In 1961 Leon Goldman started to experiment with the effect that these newly discovered laser beams had on the skin (Goldman *et al.*, 1963), and asked whether they could be used to perform "bloodless surgery". In the early 1960s Paul McGuff in Boston made medical history by using a laser beam to vaporize human cancer cells that had been transplanted into a hamster (McGuff *et al.*, 1965). In 1965 Mester started laser research in Hungary and tried to repeat McGuff's experiments by implanting tumor cells beneath the skin of laboratory rats and exposing them to the beam from a customized ruby laser. However the tumor cells were not destroyed by doses of what was presumed to be high-power laser energy, but instead, the skin incisions that had been made to implant the cancer cells appeared to heal faster in laser-treated animals, compared to incisions of control animals that were not treated with light (Mester *et al.*, 1968b). Moreover the regrowth of hair on the depilated rat skin, was observed to be faster after exposure to his ruby laser (Mester *et al.*, 1968c). After being initially puzzled by these contradictory findings, he realized that the output power of his custom-designed ruby laser was much weaker than he originally thought it to be, and instead of being photo-ablative against the tumor tissue, the low-power light stimulated the skin to

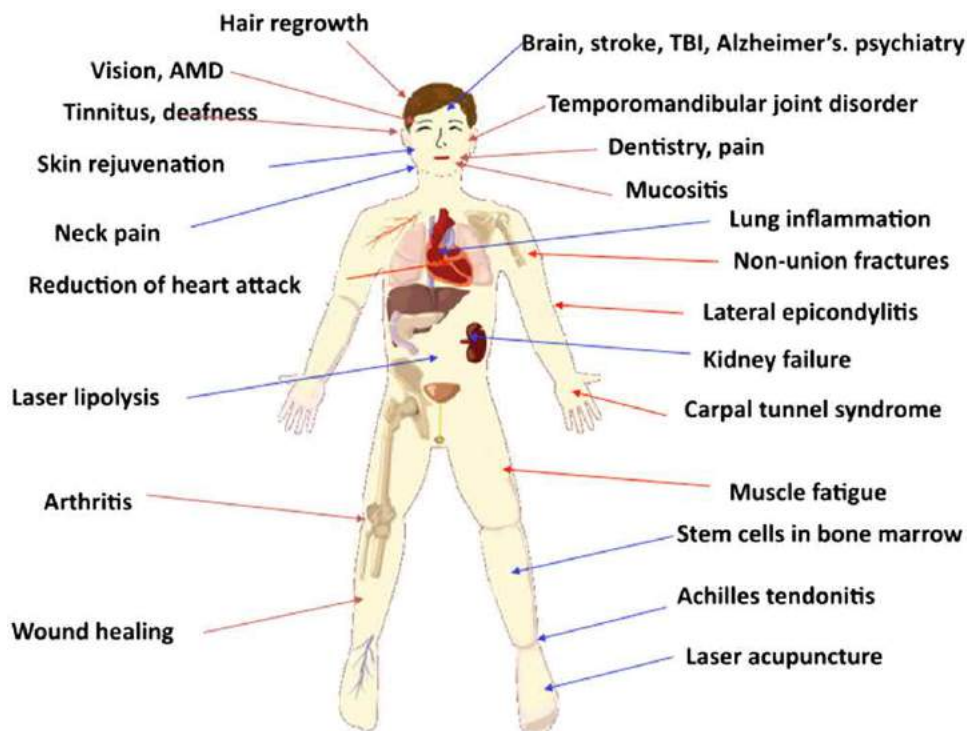


Fig. 8 Wide range of medical indications susceptible to treatment by PBM.

heal faster and caused the hair to regrow. Since those early days the range of medical indications that can be treated by LLLT or PBM has continued to expand almost exponentially as shown in [Fig. 8](#).

Mechanisms of Action

The first law of photobiology lays down that in order for a photon to have a biological effect it has to be absorbed by a specific photoacceptor molecule also known as a “chromophore”. The leading hypothesis is that light is absorbed by chromophores in the mitochondria such as cytochrome c oxidase leading to production of more ATP and activation of a host of cell signaling pathways leading to gene transcription (Chen *et al.*, 2009). A secondary hypothesis is that longer wavelength light (in the infrared region) is absorbed by water in sensitive cellular locations such as ion channels (see [Fig. 9](#) for a graphical illustration). Instead of killing cells (as is the case with PDT) PBM stimulates cellular proliferation and prevents cells from dying (Huang *et al.*, 2009). The medical applications of PBM range from serious diseases such as stroke, heart attack, spinal cord injury, traumatic brain injury to less serious indications such as wound healing, reduction of pain and inflammation and depression (Hashmi *et al.*, 2010).

Cytochrome C oxidase

Cytochrome C oxidase (Cox) is the terminal enzyme of the electron transport chain, mediating the electron transfer from cytochrome c to molecular oxygen. Several lines of evidence show that Cox acts as a photoacceptor and transducer of photosignals in the red and near-infrared regions of the light spectrum (Anneroth *et al.*, 1988). It seems that PBM increases the availability of electrons for the reduction of molecular oxygen in the catalytic center of Cox, increasing the mitochondrial membrane potential (MMP) and the levels of ATP, cAMP and ROS as well (Barcia, 2012).

PBM increases the activity of complexes I, II, III, IV and succinate dehydrogenase in the electron transfer chain. Cox is known as complex IV and, as mentioned before, appears to be the primary photoacceptor. This assumption is supported by the increased oxygen consumption during low-level light irradiation (the majority of the oxygen consumption of a cell occurs at complex IV in the mitochondria), and by the fact that sodium azide, a Cox inhibitor, prevents the beneficial effect of PBM. Besides ATP and cAMP, nitric oxide (NO) level is increased, either by release from metal complexes in Cox (Cox has two heme and two copper centers) or by up-regulation of the activity of Cox as a nitrite reductase (Barolet *et al.*, 2009).

Retrograde mitochondrial signaling

One of the most accepted mechanisms for light-cell interaction was proposed by Karu (Barrett and Gonzalez-Lima, 2013), referring to the retrograde mitochondrial signaling that occurs with light activation in the visible and infrared range. According to Karu, the first step is the absorption of a photon with energy $h(\nu)$ by the chromophore Cox. This interaction increases

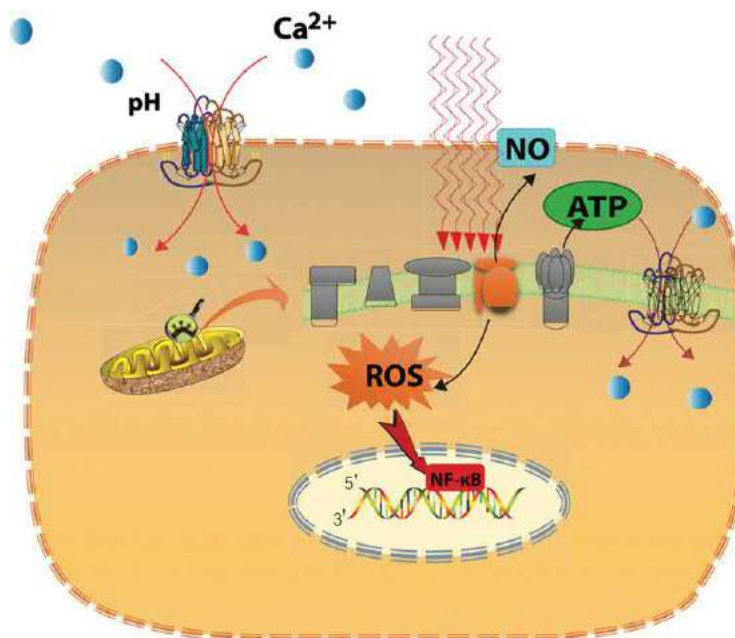


Fig. 9 Main mechanisms of action of PBM at a cellular level.

mitochondrial membrane potential ($\Delta\Psi_m$), causing an increase in the synthesis of ATP and changes in the concentrations of reactive oxygen species (ROS), Ca^{2+} and NO. Furthermore, there is a communication between mitochondria and the nucleus, driven by changes in the mitochondria ultrastructure, i.e. changes in the fission-fusion homeostasis in a dynamic mitochondrial network. The alteration in the mitochondrial ultrastructure induces changes in ATP synthesis, in the intracellular redox potential, in the pH and in cyclic adenosine monophosphate (cAMP) levels. Activator protein-1 (AP1) and NF- κ B have their activities altered by changes in membrane permeability and ion flux at the cell membrane. Some complementary routes were also suggested by Karu, such as the direct upregulation of some genes (Bellnier *et al.*, 1997).

Light/heat sensitive ion channels

The most well-known ion channels that can be directly gated by light are the channelrhodopsins (ChRs), which are seven-transmembrane-domain proteins that can be naturally found in algae providing them with light perception. Once activated by light, these cation channels open and depolarize the membrane. They are currently being applied in neuroscientific research in the new discipline of optogenetics (Bellnier *et al.*, 1989).

However members of another broad group of ion-channels are now known to be more or less light sensitive (Benvenuti, 2016). These channels are called “transient receptor potential” (TRP) channels as they were first discovered in a *Drosophila* mutant (Benvenuti, 2016). TRP channels are responsible for vision in insects. There are now known to be at least 50 different TRP isoforms distributed amongst seven subfamilies (Bernstein, 2005). These are called the TRPC (‘Canonical’) subfamily, the TRPV (‘Vanilloid’), the TRPM (‘Melastatin’), the TRPP (‘Polycystin’), the TRPML (‘Mucolipin’), the TRPA (‘Ankyrin’) and the TRPN (‘NOMPC’) subfamilies. A wide range of stimuli modulate the activity of different TRP such as light, heat, cold, sound, noxious chemicals, mechanical forces, hormones, neurotransmitters, spices, and voltage. TRP are calcium channels modulated by phosphoinositides (Bertoloni *et al.*, 1990).

The evidence that light mediated activation of TRP is responsible for some of the mechanisms of action of PBM is somewhat sparse at present, but is slowly mounting.

Mast cells are known to accumulate at the site of skin wounds, and there is some degree of evidence suggesting that these cells play a role in the biological effects of laser irradiation on promoting wound healing. Yang and co-workers demonstrated that after laser irradiation (532 nm), the intracellular $[\text{Ca}^{2+}]$ was increased and, as a consequence, there was a release of histamine. If the TRPV4 inhibitor, ruthenium red, was used, the histamine release was blocked, indicating the central role of these channels in promoting histamine-dependent wound healing after laser irradiation (Bhargava *et al.*, 2012).

It seems that TRPV1 ion channels are involved in the degranulation of mast cells and laser-induced mast cell activation. It was demonstrated that capsaicin, temperatures above 42°C and acidic pH could induce the expression of TRPV1 in oocytes, and these ion channels can be activated by green light (532 nm) in a power-dependent manner, although blue and red light were not able to activate them (Bhat *et al.*, 2005). Infrared light (2780 nm) attenuates TRPV1 activation by capsaicin in cultured neurons, decreasing the generation of pain stimuli. TRPV4 is also attenuated by laser light, but the antinociceptive effect was less intense, therefore the antinociception in this model is mainly dependent on TRPV1 inhibition (Bhuvaneswari

et al., 2009) The stimulation of neurons with pulsed infrared light (1875 nm) is able to generate laser-evoked neuronal voltage variations and, in this case, TRPV4 channels were demonstrated to be the primary effectors of the chain reaction activated by the laser (Bisht *et al.*, 1994).

Flavins

Rachel Lubart showed that flavins were able to absorb light in the 400–500 nm wavelength range (Eichler *et al.*, 2005). This led to the production of ROS inside cells. Since ROS appear to be produced by different wavelengths of light that target different chromophores, it is probably a general mechanism involved in PBM effects (Lavi *et al.*, 2010).

Direct cell-free light-mediated effects on molecules

There have been some scattered reports that light can have effects on some important molecules in cell free systems (in addition to the established effect on Cox). The latent form of transforming growth factor beta has been reported to be activated by light exposure (Bisht *et al.*, 1999). Superoxide dismutase (Cu,Zn) from bovine erythrocytes that had been inactivated by exposure to pH 5.9 was reactivated by exposure to He-Ne laser light (632.7 nm) (Bjordan *et al.*, 2003). The same treatment also reactivated the heme-containing catalase. Amat *et al.* showed that irradiation of ATP in solution by 655 nm or 830 nm light appeared to produce changes in its enzyme reactivity, fluorescence and Mg²⁺ + binding capacity (Bjornsti and Houghton, 2004). However other workers were unable to repeat this somewhat surprising result (Blanco *et al.*, 2015).

PBM for Healing of Wounds and Tissue Damage

The first demonstration the clinical effectiveness of PBM was in non-healing wounds in patients in the early 1970s (Mester *et al.*, 1972). Since then many authors have published studies in both animal models and clinical trials showing positive effects. Our study in excisional wounds in mice showed that certain wavelengths (635 nm and 810 nm) were better than others and moreover there was a biphasic dose response with an optimum fluence of 2 J/cm² (Demidova-Rice *et al.*, 2007).

Wound healing was one of the first applications of PBM, when HeNe lasers were used by Mester *et al.* to treat skin ulcers (Mester *et al.*, 1971, 1972, 1976, 1968b). PBM is believed to affect all three phases of wound healing (Thomas *et al.*, 1995): the inflammatory phase, in which immune cells migrate to the wound, the proliferative phase, which results in increased production of fibroblasts and macrophages, and the remodeling phase, in which collagen deposition occurs at the wound site and the extracellular matrix is rebuilt.

PBM is believed to promote wound healing by inducing the local release of cytokines, chemokines, and other biological response modifiers that reduce the time required for wound closure, and increase the mean breaking strength of the wound (Bisht *et al.*, 1999; Hawkins *et al.*, 2005; Meyers, 1990). Proponents of PBM speculate that this result is achieved by increasing the production and activity of fibroblasts and macrophages, improving the mobility of leukocytes, promoting collagen formation, and inducing neovascularization (Hawkins and Abrahamse, 2005; Loevschall and Arenholt-Bindslev, 1994; Medrado *et al.*, 2003; Noble *et al.*, 1992; Reddy *et al.*, 2001; Skinner *et al.*, 1996).

However, there is a lack of convincing clinical studies that either prove or disprove the efficacy of PBM in wound healing. The results that are currently available are conflicting and do not lead to any clear conclusions. For example, Abergel *et al.* found that the 632.8 nm HeNe laser did not have any effect on the cellular proliferation of fibroblasts, while the 904 nm GaAs laser actually lowered fibroblasts proliferation (Abergel *et al.*, 1987a). In contrast, other studies noted an increase in proliferation of human fibroblasts exposed to 904 nm GaAs lasers (Pereira *et al.*, 2002), rat myofibroblasts exposed to 670 nm GaAs lasers (Medrado *et al.*, 2003), and gingival fibroblasts exposed to diode lasers (670, 692, 780, and 786 nm) (Almeida-Lopes *et al.*, 2001). *In vivo* studies in both animal and human models show similar discrepancies. A study by Kana *et al.* claimed that treatment of open wounds in rats with HeNe and argon lasers resulted in faster wound closure (Kana *et al.*, 1981). Bisht *et al.* found a similar increase in granulation tissue and collagen expression in rats using the same treatment as Kana (Bisht *et al.*, 1994). However, Anneroth *et al.* failed to observe any beneficial effects after laser treatment in a comparable rat model (Anneroth *et al.*, 1988). In human studies, Schindl *et al.* reported that application of a HeNe laser was beneficial in promoting wound healing in 3 patients (Schindl *et al.*, 2000), whereas Lundeberg *et al.* found no statistically significant difference between leg ulcer patients treated with an HeNe laser and those treated with a placebo (Lundeberg and Malm, 1991).

The scarcity of well-designed clinical trials makes it difficult to assess the impact of PBM on wound healing. Our task is further complicated by the difficulty in comparing studies, due to the large number of factors involved. In addition to the multiple parameters that must be adjusted in order to apply PBM, such as the wavelength and power of the light, the effectiveness of the treatment also depends on many factors such as the location and nature of the wound, and the physiologic state of the patient. For example, impaired wound healing is one of the major chronic complications of diabetes (Goodson and Hunt, 1979; Raskin *et al.*, 1975), and is thought to result from various factors, including decreased collagen production and impaired functionality of fibroblasts, leukocytes, and endothelial cells (Goodson and Hunt, 1979; Spanheimer *et al.*, 1988). It has therefore been hypothesized that PBM could have beneficial effects in stimulating wound healing in diabetic patients (Schindl *et al.*, 1998, 2002; Yu *et al.*, 1997). Thus, in order to obtain a convincing verdict on the impact of PBM on wound healing, we will require several large, randomized, placebo controlled, and double blind trials that compare the effects of PBM on wounds that are as similar as possible. A greater understanding of the cellular and biochemical mechanisms of PBM would

also be useful in assessing these studies, as it would enable us to pinpoint exactly what criteria to use in determining the effectiveness of the therapy.

PBM for Pain and Inflammation

PBM can be used with great benefit in numerous inflammatory disorders such as arthritis, carpal tunnel syndrome, and tendinitis. We showed that 810-nm laser was very effective in a rat model of inflammatory arthritis caused by intra-articular injection of zymosan (Castano *et al.*, 2007). A sufficiently long illumination time was more important than total fluence or the irradiance in determining the benefit of the therapy.

There appears to be more firm evidence to support the success of PBM in alleviating pain and treating chronic joint disorders, than in healing wounds. A review of 16 randomized clinical trials including a total of 820 patients found that PBM reduces acute neck pain immediately after treatment, and up to 22 weeks after completion of treatment in patients with chronic neck pain (Chow *et al.*, 2009). PBM has also been shown to relieve pain due to cervical dentinal hypersensitivity (Sandford and Walsh, 1994), or from periodontal pain during orthodontic tooth movement (Wahl and Bastanier, 1991). A study of 88 randomized controlled trials indicated that PBM can significantly reduce pain and improve health in chronic joint disorders such as osteoarthritis, patellofemoral pain syndrome, and mechanical spine disorders (Bjordan *et al.*, 2003). However, the authors of the study urge caution in interpreting the results due to the wide range of patients, treatments, and trial designs involved.

PBM for Cosmetic and Esthetic Applications

Skin rejuvenation

PBM is a novel treatment option for non-thermal and non-ablative skin rejuvenation which has been shown to be effective for improving skin conditions such as wrinkles and skin laxity (Barolet *et al.*, 2009; Bhat *et al.*, 2005; Dierickx and Anderson, 2005; Russell *et al.*, 2005; Weiss *et al.*, 2004a,b). A wide range of different light sources have been used to deliver light for these treatments, particularly to the face, and some are shown in Fig. 10. PBM is a treatment modality that has been shown to provide increased rates of skin rejuvenation and wound healing with great efficacy, while also reducing post-operative pain, edema and several types of inflammation, making it highly desirable modality (Kim and Calderhead, 2011; R.G. *et al.*, 2008). Early studies by Abergel *et al.* and Yu *et al.* reported an increase in production of pro-collagen, collagen, basic fibroblast growth factors (bFGF) and proliferation of fibroblasts after exposure to low-energy laser irradiation in vitro and in vivo animal models (Abergel *et al.*, 1987b; Yu *et al.*, 1994). Use of light sources with wavelengths of 633 nm/830 nm is most common in cases of clinical application involving wound healing and skin rejuvenation. PBM is now used for the treatment of even chronic non-healing wounds through restoration of imbalances in collagenesis/collagenase and allows for rapid and enhanced wound healing in general (Kim and Calderhead, 2011). Lee *et al.* conducted a study to investigate the histological and ultrastructural alterations that followed a series of light treatments utilizing light emitting diodes (LEDs) with parameters of 830 nm, 55 mW/cm², 66 J/cm² and 633 nm, 105 mW/cm², 126 J/cm². They observed alteration in the levels of MMPs and tissue inhibitors of metalloproteinases (TIMPs)



Fig. 10 Selection of light sources used for PBMT of skin conditions.

(Lee *et al.*, 2007). The study also showed increased mRNA levels of interleukin-1 beta (IL-1 β), tumor necrosis factor alpha (TNF- α), intercellular adhesion molecule 1 (ICAM-1), and connexin 43 (C $\times$ 43) following LED phototherapy whereas IL-6 levels were decreased (Lee *et al.*, 2007). Subsequently the study also demonstrated a well-marked increase in the amount of collagen in the post-treatment specimens (Lee *et al.*, 2007). In fractional laser resurfacing it is thought that the deliberate development of microscopic photothermally-mediated wounds is responsible for the recruitment of pro-inflammatory cytokines IL-1 β and TNF- α *in order cause wound repair*. The generation of such a wound healing cascade thus contributes to new collagen synthesis (Lee *et al.*, 2007). PBM may also induce this wound healing process through athermal and atraumatic induction of a subclinical 'quasi-wound', even without any actual wounding created by thermal damage which can possibly cause complications as in some other laser treatments (Lee *et al.*, 2007). MMP activities are known to be inhibited by TIMPs, suggesting the possibility of other mechanisms for increased collagen synthesis through induction of TIMPs.

Collectively viewing these findings, they are suggestive of the idea that an increased production of IL-1 β and TNF- α might be responsible for induction of MMP activity as an early response to light treatment, which might possibly contribute to the removal of photodamaged collagen fragments in order to facilitate new collagen biosynthesis. Furthermore, as a consequence of the therapy there may be increased concentrations of TIMPs that most likely play a role in the protection of the newly synthesized collagen, from proteolytic degradation by MMPs (Lee *et al.*, 2007). Subsequently, heightened expression of C $\times$ 43 may possibly enhance cell-cell communication between dermal components, especially between fibroblasts, allowing for greater synchrony between cellular responses, following the effects of PBM in order to promote synthesis of new collagen in a greater area including even the regions that did not receive light irradiation (Lee *et al.*, 2007).

A clinical study conducted by Weiss *et al.* demonstrated the benefits of PBM over traditional thermal-based rejuvenation modalities. A group of 300 patients were administered PBM (590 nm, 0.10 J/cm²) alone, and another group of 600 patients received a combination of PBM with a thermal-based photorejuvenation procedure. Of the patients who received just the light treatment, 90% reported an observed softening of skin textures as well as a reduction in skin coarseness and fine lines that ranged from small alterations to significant changes (Weiss *et al.*, 2005b). It was observed that patients who received a form of PBM (n=152) reported a noticeable reduction in post-treatment erythema and an overall impression of increased efficacy versus patients that received treatment through a thermal photorejuvenation laser or light source lacking any sort of PBM photomodulation (Kucuk *et al.*, 2010; Weiss *et al.*, 2005b). Reduction in post-treatment erythema can most likely be attributed to the anti-inflammatory effects of PBM (Barolet *et al.*, 2009). Utilizing different pulse sequence parameters, a multicenter clinical trial was conducted, wherein 90 patients received 8 PBM treatments over 4 weeks (Geronemus *et al.*, 2003; McDaniel *et al.*, 2003; Weiss *et al.*, 2004a,b, 2005a). The study found good overall results with more than 90% of patients improving by at least one Fitzpatrick photoaging category and 65% of the patients displaying global improvement in facial texture, fine lines, background erythema and pigmentation with results peaking at 4 to 6 months following completion of the 8 treatments. A noticeable increase in papillary dermal collagen and reductions in MMP-1 were generally observed. A study conducted by Barolet *et al.* also proved to be consistent with the aforementioned studies. The study used a 3-D model of tissue-engineered Human Reconstructed Skin (HRS) to investigate the potential of PBM (660 nm, 50 mW/cm, 4 J/cm²) in collagen and MMP-1 modulation. The results showed up-regulation of collagen and down-regulation MMP-1 *in vitro* (Barolet *et al.*, 2009). A split-face, single-blinded clinical study was then carried out to assess the results of this light treatment on skin texture and appearance of individuals with aged/photoaged skin (Barolet *et al.*, 2009). Profilometry quantification (to measure wrinkles) demonstrated that more than 90% of individuals had a reduction in rhytid depth and surface roughness, and, 87% of the individuals reported that they have experienced a reduction in the Fitzpatrick wrinkling severity score following 12 PBM treatments (Barolet *et al.*, 2009).

PBM for hair regrowth

As mentioned above in the late 1960s, Endre Mester, a Hungarian physician, conducted a series of experiments to investigate the anti-cancer effect of lasers using a low-powered ruby laser (694 nm) on mice and much to his surprise, not only did laser exposure not cure cancer instead it seemed to enhance hair growth in the shaved area of the mice that had been set up for the original experiment (Mester *et al.*, 1968a). Quite recently, lasers have garnered much attention due to their remarkable ability to cause selective hair removal, however in some cases it was observed that lasers in fact produced undesired hair growth effects such as increase in hair density, color or coarseness or a combination of any of these, in areas of the skin that were treated for hair removal (Moreno-Arias *et al.*, 2002a,b; Vlachos and Kontoes, 2002; Wikramanayake *et al.*, 2012a). This phenomenon is known as 'paradoxical hypertrichosis' and its incidence varies from 0.6% to 10% (Wikramanayake *et al.*, 2012a). A group of researchers also observed that low-powered laser irradiation of small vellus hairs cause them to transform into larger terminal hairs, and thus termed this process as "terminalization" of vellus hair follicles (Bernstein, 2005; Bouzari and Firooz, 2006). Different mechanisms have been proposed by dermatologists (who did not really believe in PBM) in an attempt to explain these effects of lasers on hair removal. In one particular study this ability was attributed to the fact that polycystic ovarian syndrome was present in 5 out of 49 females undergoing intense pulsed laser (IPL) laser treatment for facial hirsutism (Moreno-Arias *et al.*, 2002a). Another study suggested that despite the heat generating effects of laser treatments, the heat produced was not sufficient enough to induce total hair follicle thermolysis but, it may be sufficient to stimulate follicular stem cell proliferation and differentiation by increasing levels of heat shock proteins (HSPs) such as HSP27 which influence regulation of cell growth and differentiation (Wikramanayake *et al.*, 2012a).

In 2007, the FDA approved PBM as a possible treatment modality for hair loss (Wikramanayake *et al.*, 2012a). It is believed that non-thermal PBM can stimulate reentry of anagen hair follicles into the telogen stage, bringing about greater rates of proliferation

in active anagen follicles, can prevent development of premature catagen stage follicles and extend the duration of the anagen phase (Leavitt *et al.*, 2009; Wikramanayake *et al.*, 2012a).

In one study conducted by Yamazaki and colleagues, irradiated the backs of Sprague Dawley rats using a linearly polarized IR source, where an up-regulation of hepatocyte growth factor (HGF) and HGF activator expression was discovered (Miura *et al.*, 1999). Another study reported increases in temperature of the skin and improved blood flow around the stellate ganglion area, following treatment with PBM (Wajima *et al.*, 1996). The exact mechanism for the action of minoxidil in treating hair loss is not entirely understood, but it is known that minoxidil does contain nitric oxide (NO) which is an important cellular signaling molecule and vasodilator (Proctor, 1989) that is influential to a variety of physiological and pathological processes (Hou *et al.*, 1999). Furthermore, NO is a regulator of the opening of ATP-dependent potassium (K^+) channels and is thus responsible for the hyperpolarization of cell membranes (Rossi *et al.*, 2012). Also, it has been suggested that ATP sensitive K^+ channels of the mitochondria and elevated levels of NO might be involved in the mechanism of action of PBM (Karu, 2008; Karu *et al.*, 2005; Tuby *et al.*, 2006) in areas of the brain and heart (Karu, 2008; Karu *et al.*, 2004; Y.D. Ignatov, 2005). Thus given the dependency of both minoxidil and PBM on the aforementioned factors there is possibly some mechanistic overlap between the two modalities. A study conducted by Weiss *et al.* demonstrated that PBM was able to modulate 5- α reductase, the enzyme responsible for the conversion of testosterone to DHT, as well as alter the genetic expression of vascular endothelial growth factor (VEGF), which plays an influential role in hair follicle growth and thus PBM is able to stimulate hair growth (Castex-Rizzi *et al.*, 2002; Weiss *et al.*, 2005a,b; Yano and Detmar, 2001). Furthermore, it has been demonstrated that PBM may stimulate hair growth through modulation of inflammatory processes and immunological responses (Meneguzzo *et al.*, 2012). A study conducted by Wikramanayake *et al.* on C3H/HeJ AA mouse models supported this assumption, wherein the mice were exposed to a laser-comb and it was observed that the treatment led to an increase in the quantity of hair follicles where the majority of the follicles in anagen phase were seen to have decreased inflammatory infiltrates (Wikramanayake *et al.*, 2012a). Taking into account the disruptive effect that inflammatory infiltrates have on hair follicles along with the notion that several cytokines such as interferon gamma (IFN- γ), IL-1 α and β , TNF- α , MHC and Fas-antigen and macrophage migration inhibitory factor are all involved in cyclic hair growth as well as the pathogenesis of AA, PBM may be able to play significant role in the treatment of AA due to its modulating effects on inflammation (Wikramanayake *et al.*, 2012a).

A clinical study was carried out to investigate the effect of PBM on treatment of alopecia areata consisting of a sample size of 15 patients (6 men, 9 women) utilizing Super Lizer TM, a medical instrument operating on polarized linear light with a high output (1.8 W) of NIR radiation (600–1600 nm) possessing sufficient penetration depth to reach deep subcutaneous tissue (Yamazaki *et al.*, 2003). The patients received a 3 min laser treatment on the scalp either once a week or once every 2 weeks and were administered additional carpronium chloride 5% twice daily to all lesions (Yamazaki *et al.*, 2003). Supplemental oral antihistamines, cepharranthin and glycyrrhizin (extracts of medicinal Chinese herbs) were prescribed as well (Yamazaki *et al.*, 2003). The results of the study showed that 47% of the patients experienced hair growth 1.6 months earlier on areas irradiated with laser when compared to the areas that were not irradiated (Yamazaki *et al.*, 2003).

In another study 24 male androgenetic alopecia (AGA) patients were evaluated via global photography and phototrichogram using 655 nm red light and 780 nm IR light once a day for a period of 10 min (Kim *et al.*, 2007). Following 14 weeks of treatment significant increases in hair density and anagen/telogen ratio were observed at both the vertex and occiput, with 83% patients reporting that the treatment resulted in satisfactory results (Kim *et al.*, 2007).

Satino *et al.* conducted a study to investigate the efficacy of PBM on hair growth and tensile strength involving 28 male and 7 female AGA patients (Satino, 2003). Each patient was given a 655 nm HairMax LaserComb® to use at home for a period of 6 months, applying it for five to ten minutes per day on alternate days (Satino, 2003). Regarding the tensile strength of the hair, the results showed improvements in hair growth in all treated areas for both male and female sexes, however in the case of males the greatest improvements were observed in the vertex area whereas for females, the best improvements were seen in the temporal area (Satino, 2003). With regards to hair count, again all treated areas of both sexes showed improvement, but the vertex area showed the greatest improvement for the male patients (Satino, 2003). Leavitt *et al.* conducted a double-blind, sham device-controlled, multi-center, randomized 26 week trial where they tested the same device on 110 male AGA patients (Leavitt *et al.*, 2009). The patients were made to use the device three times a week for fifteen minutes for a total period of 26 weeks (Leavitt *et al.*, 2009). Noticeable increases in mean terminal density of hair were reported in the treatment group when compared to the sham treatment group (Leavitt *et al.*, 2009). Also, subjective assessment of the patients over the 26 week period reported prominent improvements in overall hair regrowth, decreased rate of hair loss, thicker feeling hair, scalp health and hair shine (Leavitt *et al.*, 2009).

Around 65% of the patients receiving chemotherapy for cancer develop alopecia which can have detrimental effects on psychological health of the patient (Trueb, 2009). It has been proposed that PBM could serve as treatment modality to stimulate and promote hair growth in cases of chemotherapy induced alopecia. In one study, a rat model was given varying regimens of chemotherapy in conjunction with PBM administered with a device possessing components (laser unit and switch, lacking comb or handle) of the Hair Max LaserComb (Wikramanayake *et al.*, 2012b). In all rats that were given laser treatment, hair regrowth occurred at a faster rate when compared to the sham treatment group. Additionally, PBM did not hinder the efficacy of the chemotherapeutic procedures (Wikramanayake *et al.*, 2012b).

PBM for fat reduction and cellulite treatment

In 2000 Neira *et al.* demonstrated the use of low level laser as a new alternative to liposuction, and successfully utilized it with doses that did not produce any detectable increases in tissue temperature or cause any noticeable macroscopic alterations in the tissue

structure (Neira *et al.*, 2000, 2002). Prior investigations concerned with the effects of LLLT on wound healing, pain relief and edema prevention paved the way for this therapeutic application (GD *et al.*, 1991; PR, 1989). The development of LLLT as a therapeutic modality to augment liposuction while avoiding macroscopic tissue alterations were based on determination of optimal parameters such as wavelength and power output for use (J.L., 2000). Evidence suggests that the best wavelengths suitable for biomodulation range between 630 and 640 nm (FAH. and XY., 1996; Frohlich, 1970, 1975; Fàa, 1968; Sroka *et al.*, 1997; van Breugel and Bar, 1992). Nierra *et al.* made several intriguing observations regarding the effects of LLLT on adipocytes. They utilized low level diode laser (635 nm) and a maximal power of 10 mW with energy values ranging from 1.2 to 3.6 J/cm² (Neira *et al.*, 2002). Using scanning electron microscopy (SEM) and transmission electron microscopy (TEM) it was demonstrated that adipocyte plasma membranes exhibited transitory pore formation as result of irradiation. It was formulated that this enabled the release of intracellular lipids from the adipocytes and thus supplemented the liposuction as it was expected to reduce the time taken for the procedure, allowed for extraction of greater volumes of fat and overall, reduced the energy expenditure of the surgeon.

Although the findings associated with LLLT enjoyed much praise and enthusiasm, an extensively study conducted put these findings surrounding LLLT into question (Brown *et al.*, 2004). In their study, cultured human preadipocytes did not show any differences when compared to non-irradiated cells after 60 min of irradiation using an LLLT source (635 nm and fluence 1 J/cm²) (Brown *et al.*, 2004). Furthermore, histological examination of lipoaspirates, in a porcine model exposed to LLLT for 30 min and human lipoaspirates, failed to demonstrate transitory pores when analyzed using SEM (Brown *et al.*, 2004). Additional data, raised questions regarding the ability of red light (635 nm) to effectively penetrate below the skin, into the sub-dermal tissues (Kolari and Airaksinen, 1993). Peter Foddor supportively stated: "One could postulate that the presence of the black dots on scanning electron microscopy images on the surface of fat cells reported by Neira *et al.* could represent an artifact." (Brown *et al.*, 2004). Since the data reported by Brown *et al.* from 2004, there have been several publications reporting the efficacy of LLLT.

In the original paper published by Neira *et al.* the fat liberating effects of LLLT on adipocytes were attributed to its ability to induce transitory micropores which were visualized with the help of SEM (Neira *et al.*, 2002). Furthermore, it was postulated that this stimulated the release of intracellular lipids from the adipocytes. Based on this, it was formulated that up to 99% of the fat stored within the adipocytes could be released and subsequently removed with the help of LLLT (635nm, 10 mW intensity, 6 min irradiation time) (Neira *et al.*, 2002). Re-cultured adipocytes exhibited a tendency to attain their native cellular conformation which was further confirmed by Caruso-Davis *et al.* utilizing a live-dead assay to assess the viability of these adipocytes following irradiation (Caruso-Davis *et al.*, 2011). Increase in ROS following LLLT has been proposed to bring about lipid peroxidation within the cell membrane which may cause damage and may present as the transitory pores. This may cause temporary damage that presents as transitory micropores. (Chen *et al.*, 2011; Geiger *et al.*, 1995; Karu, 2008; Tafur and Mills, 2008). However, when Brown *et al.* attempted to replicate Neira *et al.*'s findings (Neira *et al.*, 2002), they failed to visualize any transitory micropores via SEM (Brown *et al.*, 2004). No further SEM studies have documented these pores, but many publications have reported findings that indirectly support the transitory micropore formation theory. Another proposed mechanism that explains the release of intracellular lipids from adipocytes, suggests the activation of the compliment cascade which is responsible for induction of adipocyte apoptosis and subsequent release of intracellular lipid components (Caruso-Davis *et al.*, 2011). To test the feasibility of this theory Caruso-Davis *et al.* exposed differentiated human adipocytes to plasma and exposed one group of cells to laser, while the control group received no laser intervention (Caruso-Davis *et al.*, 2011). No differences were noticed with regard to complement system components in either group and it was concluded that the LLLT did not act through activation of the complement cascade (Caruso-Davis *et al.*, 2011).

PBM for Vision and Hearing Problems

PBM has been used to treat disorders of sensory systems such as vision and hearing. Light can easily be shone into the eyes, and its known ability to heal and preserve cells and restore function means that PBM works well in the retina (Eells *et al.*, 2016; Geneva, 2016). It should be remembered that the retina is one of the most metabolically active tissue in the body, with a high demand for oxygen and glucose (Sanchez-Chavez *et al.*, 2012). Most studies of PBM for ophthalmology appear to have used red light (more specifically 670 nm from LEDs). PBM has been reported to be beneficial (in patients as well as in animal models) for age-related macular degeneration (Ivandic and Ivandic, 2008), diabetic retinopathy (Tang *et al.*, 2013, 2014), retinopathy of prematurity (Natoli *et al.*, 2013), light-induced photoreceptor degeneration (Albarracin *et al.*, 2011; Qu *et al.*, 2010), amblyopia (Ivandic and Ivandic, 2012), retinitis pigmentosa (Ivandic and Ivandic, 2014), and blindness caused by methanol toxicity (Eells *et al.*, 2003).

In a similar approach to shining light in the eyes for problems with vision, investigators have studied the effect of shining light in the ears to treat disorders of hearing. In this case however there is no ideal clear optical path for the light to get to the intended target cells in the inner ear. PBM has been studied for noise-induced hearing loss (Lee *et al.*, 2016; Tamura *et al.*, 2015), chronic tinnitus and sensorineural hearing loss (Tauber *et al.*, 2003), antibiotic-induced ototoxicity (Rhee *et al.*, 2013) and Meniere's disease (Teggi *et al.*, 2008).

PBM for Stroke

There are a plethora of plausible mechanisms (shown in Fig. 11) that suggest the beneficial effects of PBM on the brain will be pronounced, and this prediction has largely been proved to be true. Perhaps the most well-investigated application of PBM to the

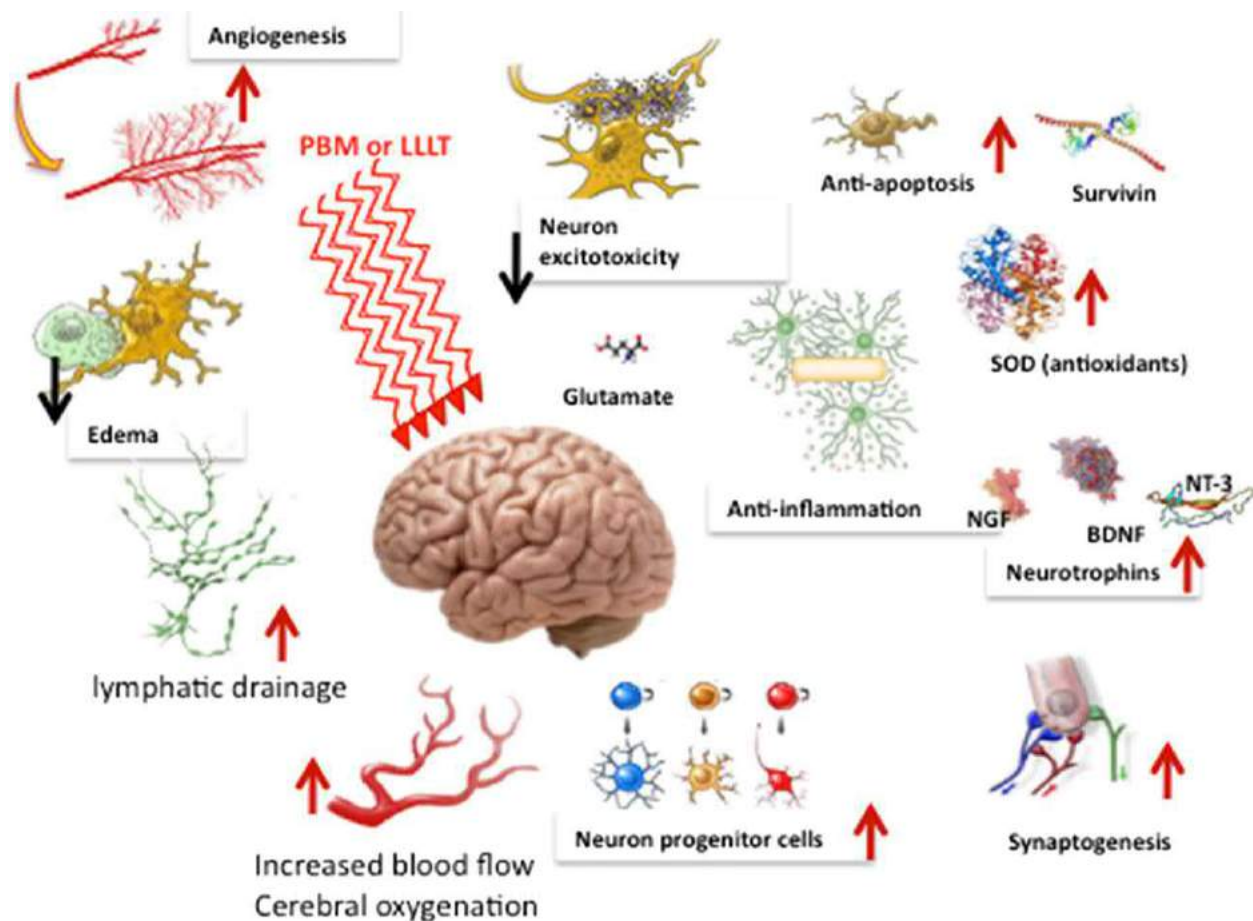


Fig. 11 Many different mechanisms may contribute to the beneficial effects of PBMT on brain disorders.

brain, lies in its possible use as a treatment for acute stroke (Zhao *et al.*, 2015). Animal models such as rats and rabbits, were first used as laboratory models. These animals had experimental strokes induced by a variety of methods, and were then treated with light (usually 810 nm laser) within 24h of stroke onset (Peplow, 2015). In these studies intervention by tPBM within 24 h had meaningful beneficial effects. For the rat models, stroke was induced by middle cerebral artery occlusion (MCAO) via an insertion of a filament into the carotid artery or via craniotomy (Oron *et al.*, 2006). Stroke induction in the “rabbit small clot embolic model” (RSCEM) was by injection of a preparation of small blood clots (made from blood taken from a second donor rabbit) into a catheter placed in the right internal carotid artery.

The promising results from pre-clinical animal studies led to a series of three clinical trials called being conducted for treatment of acute stroke in humans. These were called the “Neurothera Effectiveness and Safety Trials” (NEST-1 (Lampl *et al.*, 2007), NEST-2 (Huisa *et al.*, 2013), and NEST-3 (Zivin *et al.*, 2014)) using an 810 nm laser applied to the shaved head within 24 h of patients suffering an ischemic stroke. The first study, NEST-1, enrolled 120 patients between the ages of 40 to 85 years of age with a diagnosis of ischemic stroke involving a neurological deficit that could be measured. The purpose of this first clinical trial was to demonstrate the safety and effectiveness of laser therapy for stroke within 24h (Lampl *et al.*). tPBM significantly improved outcome in human stroke patients, when applied at ~18 h post-stroke, over the entire surface of the head (20 points in the 10/20 EEG system) regardless of stroke (Lampl *et al.*, 2007). Only one laser treatment was administered, and 5 days later, there was significantly greater improvement in the Real- but not in the Sham-treated group ($p < .05$, NIH Stroke Severity Scale). This significantly greater improvement was still present at 90 days post-stroke, where 70% of the patients treated with Real-PBM had a successful outcome, while only 51% of Sham-controls did. The second clinical trial, NEST-2, enrolled 660 patients, aged 40 to 90, who were randomly assigned to one of two groups (331 to PBM, 327 to sham) (Zivin *et al.*). Beneficial results ($p < .04$) were found for the moderate and moderate-severe (but not for the severe) stroke patients, who received the Real laser protocol (Zivin *et al.*, 2009). These results suggested that the overall severity of the individual stroke should be taken into consideration in future studies, and very severe patients are unlikely to recover with any kind of treatment. The last clinical trial, NEST-3, was planned for 1000 patients enrolled. Patients in this study were not to receive tissue plasminogen activator, but the study was prematurely terminated by the DSMB for futility (an expected lack of statistical significance) (Zivin *et al.*). NEST1 was considered successful, even though as a phase 1 trial, it was not designed to show efficacy. NEST2 was partially successful when the patients were stratified, to exclude

very severe strokes or strokes deep within the brain (Huisa *et al.*, 2013). There has been considerable discussion in the scientific literature on precisely why the NEST-3 trial failed (Lapchak and Boitano, 2016). Many commentators have wondered how could tPBM work so well in the first trial, in a sub-group in the second trial, and fail in the third trial. Lapchak's opinion is that the much thicker skull of humans compared to that of the other animals discussed above (mouse, rat and rabbit), meant that therapeutically effective amounts of light were unlikely to reach the brain (Lapchak and Boitano, 2016). Moreover the time between the occurrence of a stroke and initiation of the PBMT may be an important factor. There are reports in the literature that neuroprotection must be administered as soon as possible after a stroke (Lapchak, 2013a,b). Furthermore, stroke trials in particular should adhere to the RIGOR (rigorous research) guidelines and STAIR (stroke therapy academic industry roundtable) criteria (Lapchak *et al.*, 2013). Other contributory causes to the failure of NEST3 may have been included the decision to use only one single tPBM treatment, instead of a series of treatments. Moreover, the optimum brain areas to be treated in acute stroke remain to be determined. It is possible that certain areas of the brain that have sustained ischemic damage should be preferentially illuminated and not others.

Somewhat surprisingly, there have not as yet been many trials of PBM for rehabilitation of chronic stroke patients with only the occasional report to date. Naeser reported in an abstract the use of tPBM to treat chronic aphasia in post-stroke patients (Naeser *et al.*, 2012). Boonswang *et al.* (2012) reported a single patient case in which PBM was used in conjunction with physical therapy to rehabilitate chronic stroke damage. However the findings that PBM can stimulate synaptogenesis in mice with TBI, does suggest that tPBM may have particular benefits in rehabilitation of stroke patients. Norman Doidge, in Toronto, Canada has described the use of PBM as a component of a neuroplasticity approach to rehabilitate chronic stroke patients (Doidge, 2015).

PBM for Traumatic Brain Injury (TBI)

There have been a number of studies looking at the effects of PBM in animal models of TBI. Oron's group was the first (Oron *et al.*, 2007) to demonstrate that a single exposure of the mouse head to a NIR laser (808 nm) a few hours after creation of a TBI lesion could improve neurological performance and reduce the size of the brain lesion. A weight-drop device was used to induce a closed-head injury in the mice. An 808 nm diode laser with two energy densities (1.2–2.4 J/cm² over 2 min of irradiation with 10 and 20 mW/cm²) was delivered to the head 4 h after TBI was induced. Neurobehavioral function was assessed by the neurological severity score (NSS). There were no significant difference in NSS between the power densities (10 vs 20 mW/cm²) or significant differentiation between the control and laser treated group at early time points (24 and 48 h) post TBI. However, there was a significant improvement (27% lower NSS score) in the PBM group at times of 5 days to 4 weeks. The laser treated group also showed a smaller loss of cortical tissue than the sham group (Oron *et al.*).

Hamblin's laboratory then went on (in a series of papers (Oron *et al.*, 2007)) to show that 810nm laser (and 660nm laser) could benefit experimental TBI both in a closed head weight drop model (Wu *et al.*, 2012a), and also in controlled cortical impact model in mice (Ando *et al.*, 2011). Wu *et al.* (Wu *et al.*) explored the effect that varying the laser wavelengths of PBM had on closed-head TBI in mice. Mice were randomly assigned to PBM treated group or to sham group as a control. Closed-head injury (CHI) was induced via a weight drop apparatus. To analyze the severity of the TBI, the neurological severity score (NSS) was measured and recorded. The injured mice were then treated with varying wavelengths of laser (665, 730, 810 or 980 nm) at an energy level of 36 J/cm² at 4 h directed onto the scalp. The 665 nm and 810 nm groups showed significant improvement in NSS when compared to the control group at day 5 to day 28. Conversely, the 730 and 980 nm groups did not show a significant improvement in NSS and these wavelengths did not produce similar beneficial effects as in the 665 nm and 810 nm PBM groups (Wu *et al.*). The tissue chromophore cytochrome c oxidase (CCO) is proposed to be responsible for the underlying mechanism that produces the many PBM effects that are the byproduct of PBM. CCO has absorption bands around 665 nm and 810 nm while it has low absorption bands at the wavelength of 730 nm (Karu *et al.*). It should be noted that this particular study found that the 980 nm did not produce the same positive effects as the 665 nm and 810 nm wavelengths did; nevertheless previous studies did find that the 980 nm wavelength was an active one for PBM. Wu *et al.* proposed these dissimilar results may be due to the variance in the energy level, irradiance etc. between the other studies and this particular study (Wu *et al.*).

Ando *et al.* (2011) used the 810 nm wavelength laser parameters from the previous study and varied the pulse modes of the laser in a mouse model of TBI. These modes consisted of either pulsed wave at 10 Hz or at 100 (50% duty cycle) or continuous wave laser. For the mice, TBI was induced with a controlled cortical impact device via open craniotomy. A single treatment with an 810 nm Ga-Al-As diode laser with a power density of 50 mW/m² and an energy density of 36 J/cm² was given via tPBM to the closed head in mice for a duration of 12 min at 4 h post CCI. At 48 h to 28 days post TBI, all laser treated groups had significant decreases in the measured neurological severity score (NSS) when compared to the control. Although all laser treated groups had similar NSS improvement rates up to day 7, the PW 10 Hz group began to show greater improvement beyond this point. At day 28, the forced swim test for depression and anxiety was used and showed a significant decrease in the immobility time for the PW 10 Hz group. In the tail suspension test, which measures depression and anxiety, there was also a significant decrease in the immobility time at day 28, and this time also at day 1, in the PW 10 Hz group.

Studies using immunofluorescence of mouse brains showed that tPBM increased neuroprogenitor cells in the dentate gyrus (DG) and subventricular zone at 7 days after the treatment (Xuan *et al.*, 2013). The neurotrophin called brain derived neurotrophic factor (BDNF) was also increased in the DG and SVZ at 7 days, while the marker (synapsin-1) for synaptogenesis and neuroplasticity was increased in the cortex at 28 days but not in the DG, SVZ or at 7 days (Xuan *et al.*, 2015). Learning and memory as measured by the

Morris water maze was also improved by tPBM (Xuan *et al.*, 2014). Whalen's laboratory (Khuman *et al.*, 2012) and Whelan's laboratory (Quirk *et al.*, 2012) also successfully demonstrated therapeutic benefits of tPBM for TBI in mice and rats respectively.

Zhang *et al.* (2014) showed that secondary brain injury occurred to a worse degree in mice that had been genetically engineered to lack "Immediate Early Response" gene X-1 (IEX-1) when exposed to a gentle head impact (this injury is thought to closely resemble mild TBI in humans). Exposing IEX-1 knockout mice to PBM 4 h post injury, suppressed proinflammatory cytokine expression of interleukin (IL)-1 β and IL-6, but upregulated TNF- α . The lack of IEX-1 decreased ATP production, but exposing the injured brain to PBM elevated ATP production back to near normal levels.

Dong *et al.* (2015) even further improved the beneficial effects of PBM on TBI in mice, by combining the treatment with metabolic substrates such as pyruvate and/or lactate. The goal was to even further improve mitochondrial function. This combinatorial treatment was able to reverse memory and learning deficits in TBI mice back to normal levels, as well as leaving the hippocampal region completely protected from tissue loss; a stark contrast to that found in control TBI mice that exhibited severe tissue loss from secondary brain injury.

Margaret Naeser and collaborators have tested PBM in human subjects who had suffered TBI in the past (Naeser and Hamblin, 2015). Many sufferers from severe or even moderate TBI, have very long lasting and even life-changing sequelae (headaches, cognitive impairment, and difficulty sleeping) that prevent them working or living any kind of normal life. These individuals may have been high achievers before the accident that caused damage to their brain (McClure, 2011). Initially Naeser published a report (Naeser *et al.*, 2011) describing two cases she treated with PBM applied to the forehead twice a week. A 500 mW continuous wave LED source (mixture of 660 nm red and 830 nm NIR LEDs) with a power density of 22.2 mW/cm² (area of 22.48 cm²), was applied to the forehead for a typical duration of 10 min (13.3 J/cm²). In the first case study the patient reported that she could concentrate on tasks for a longer period of time (the time able to work at a computer increased from 30 min to 3 h). She had a better ability to remember what she read, decreased sensitivity when receiving haircuts in the spots where PBM was applied, and improved mathematical skills after undergoing PBM. The second patient had statistically significant improvements compared to prior neuropsychological tests after 9 months of treatment. The patient had a 2 standard deviation (SD) increase on tests of inhibition and inhibition accuracy (9th percentile to 63rd percentile) on the Stroop test for executive function and a 1 SD increase on the Wechsler Memory scale test for the logical memory test (83rd percentile to 99th percentile) (Naeser *et al.*, 2010).

Naeser *et al.* then went on to report a case series of a further eleven patients (Naeser *et al.*, 2010). This was an open protocol study that examined whether scalp application of red and near infrared (NIR) light could improve cognition in patients with chronic, mild traumatic brain injury (mTBI). This study had 11 participants ranging in age from 26 to 62 (6 males, 5 females) who suffered from persistent cognitive dysfunction after mTBI. The participants' injuries were caused by motor vehicle accidents, sports related events and for one participant, an improvised explosive device (IED) blast. tPBM consisted of 18 sessions (Monday, Wednesday, and Friday for 6 weeks) and commenced anywhere from 10 months to 8 years post-TBI. A total of 11 LED clusters (5.25 cm in diameter, 500 mW, 22.2 mW/cm², 13 J/cm²) were applied for about 10 min per session (5 or 6 LED placements per set, Set A and then Set B, in each session). Neuropsychological testing was performed pre-LED application and 1 week, 1 month and 2 months after the final treatment. Naeser and colleagues found that there was a significant positive linear trend observed for the Stroop Test for executive function, in trial 2 inhibition ($p=0.004$); Stroop, trial 4 inhibition switching ($p=0.003$); California Verbal Learning Test (CVLT)-II, total trials 1–5 ($p=0.003$); CVLT-II, long delay free recall ($p=0.006$). Improved sleep and fewer post-traumatic stress disorder (PTSD) symptoms, if present beforehand, were observed after treatment. Participants and family members also reported better social function and a better ability to perform interpersonal and occupational activities. Although these results were significant, further placebo-controlled studies will be needed to ensure the reliability of this these data (Naeser *et al.*, 2014).

Henderson and Morries (2015) used a high-power NIR laser (10–15 W at 810 and 980 nm) applied to the head to treat a patient with moderate TBI. The patient received 20 NIR applications over a 2 month period. They carried out anatomical magnetic resonance imaging (MRI) and perfusion single-photon emission computed tomography (SPECT). The patient showed decreased depression, anxiety, headache, and insomnia, whereas cognition and quality of life improved, accompanied by changes in the SPECT imaging.

PBM for Neurodegenerative Diseases

PBM has also been considered as a candidate for treating degenerative brain disorders such as Alzheimer's disease (AD), and Parkinson's disease (PD) (Moges *et al.*, 2009; Zhang *et al.*, 2008). Although only preliminary studies have so far been carried out, these are encouraging indications that merit further investigation.

Alzheimer's disease

There was a convincing study (De Taboada *et al.*, 2011) carried out in an A β PP transgenic mouse of AD. tPBM (810 nm laser) was administered at different doses 3 times/week for 6 months starting at 3 months of age. The numbers of A β plaques were significantly reduced in the brain with administration of tPBM in a dose-dependent fashion. tPBM mitigated the behavioral effects seen with advanced amyloid deposition and reduced the expression of inflammatory markers in the transgenic mice. In addition, TLT showed an increase in ATP levels, mitochondrial function, and c-fos expression suggesting there was an overall improvement in neurological function.

There has been a group of investigators in Northern England who have used a helmet containing 1072 nm LEDs to treat AD, but somewhat surprisingly no peer-reviewed publications have described this approach (see Relevant Website section). However a small pilot study (19 patients) that took the form of a randomized placebo-controlled trial investigated the effect of the Vielight Neuro system (a combination of tPBM and intranasal PBM) on patients with dementia and mild cognitive impairment (Saltmarche *et al.*, 2016). This was a controlled single blind pilot study in humans to investigate the effects of PBM on memory and cognition. 19 participants with impaired memory/cognition were randomized into active and sham treatments over 12 weeks with a 4-week no-treatment follow-up period. They were assessed with MMSE and ADAS-cog scales. The protocol involved in-clinic use of a combined transcranial-intranasal PBM device; and at-home use of an intranasal-only PBM device and participants/caregivers noted daily experiences in a journal. Active participants with moderate to severe impairment (MMSE scores 5–24) showed significant improvements (5-points MMSE score) after 12 weeks. They also reported better sleep, fewer angry outbursts and decreased anxiety and wandering. Declines were noted during the 4 week no-treatment follow-up period. Participants with mild impairment to normal (MMSE scores of 25 to 30) in both the active and sham sub-groups showed improvements. No related adverse events were reported.

An interesting paper from Russia (Maksimovich, 2015) described the use of intravascular PBM to treat 89 patients with AD who received PBM (46 patients) or standard treatment with memantine and rivastigmine (43 patients). The PBM consisted of threading a fiber-optic through a catheter in the femoral artery and advancing it to the distal site of the anterior and middle cerebral arteries and delivering 20 mW of red laser for 20–40 min. The PBM group had improvement in cerebral microcirculation leading to permanent (from 1–7 years) reduction in dementia and cognitive recovery.

Parkinson's disease

The majority of studies on PBM for Parkinson's disease have been in animal models and have come from the laboratory of John Mitrofanis in Australia (Johnstone *et al.*, 2015). Two basic models of Parkinson's disease were used. The first employed administration of the small molecule (MPTP or 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine) to mice (Shaw *et al.*, 2010). MPTP was discovered as an impurity in an illegal recreational drug to cause Parkinson's like symptoms (loss of substantia nigra cells) in young people who had taken this drug (Barcia, 2012). Mice were treated with tPBM (670-nm LED, 40 mW/cm², 3.6 J/cm²) 15 min after each MPTP injection repeated 4 times over 30 h. There were significantly more (35%–45%) dopaminergic cells in the brains of the tPBM treated mice (Shaw *et al.*, 2010). A subsequent study showed similar results in a chronic mouse model of MPTP-induced Parkinson's disease (Peoples *et al.*, 2012). They repeated their studies in another mouse model of Parkinson's disease, the tau transgenic mouse strain (K3) that has a progressive degeneration of dopaminergic cells in the substantia nigra pars compacta (SNc) (Purushothuman *et al.*, 2013). They went on to test a surgically implanted intracranial fiber designed to deliver either 670 nm LED (0.16 mW) or 670 nm laser (67 mW) into the lateral ventricle of the brain in MPTP-treated mice (Moro *et al.*, 2014). Both low power LED and high power laser were effective in preserving SNc cells, but the laser was considered to be unsuitable for long term use (6 days) due to excessive heat production. As mentioned above, these authors also reported a protective effect of abscopal light exposure (head shielded) in this mouse model (Johnstone *et al.*, 2014). Recently this group has tested their implanted fiber approach in a model of Parkinson's disease in adult Macaque monkeys treated with MPTP (El Massri *et al.*, 2016). Clinical evaluation of Parkinson's symptoms (posture, general activity, bradykinesia, and facial expression) in the monkeys were improved at low doses of light (24 J or 35 J) compared to high doses (125 J) (Moro *et al.*, 2016).

The only clinical report of PBM for Parkinson's disease in humans was an abstract presented in 2010 (Maloney *et al.*, 2010). Eight patients between 18 to 80 years with late stage PD participated in a non-controlled, non-randomized study. Participants received tPBM treatments of the head designed to deliver light to the brain stem, bilateral occipital, parietal, temporal and frontal lobes, and treatment along the sagittal suture. A Visual Analog Scale (VAS), was used to record the severity of their symptoms of balance, gait, freezing, cognitive function, rolling in bed, and difficulties with speech pre-procedure and at study endpoint with 10 being most severe and 0 as no symptom. Compared with baseline, all participants demonstrated a numerical improvement in the VAS from baseline to study endpoint. A statistically significant reduction in VAS rating for gait and cognitive function was observed with average mean change of -1.87 ($p < 0.05$) for gait and a mean reduction of -2.22 ($p < 0.05$) for cognitive function. Further, freezing and difficulty with speech ratings were significantly lower (mean reduction of 1.28 ($p < 0.05$) for freezing and 2.22 ($p < 0.05$) for difficulty with speech).

PBM for Psychiatric Disorders

A common and well-accepted animal model of depression is called "chronic mild stress" (Willner, 1997). After exposure to a series of chronic unpredictable mild stressors, animals develop symptoms seen in human depression, such as anhedonia (loss of the capacity to experience pleasure, a core symptom of major depressive disorder), weight loss or slower weight gain, decrease in locomotor activity, and sleep disorders (Anisman and Matheson, 2005). Wu *et al.* used Wistar rats to show that after 5 weeks of chronic stress, application of tPBM 3 times a week for 3 weeks (810 nm laser, 100 Hz with 20% duty cycle, 120 J/cm²) gave significant improvement in the forced swimming test (FST) (Wu *et al.*, 2012b). In a similar study Salehpour *et al.* (2016) compared the effects of two different lasers (630 nm at 89 mW/cm², and 810 nm at 562 mW/cm², both pulsed at 10 Hz, 50% duty cycle). The 810 nm laser proved better than the 630 nm laser in the FST, in the elevated plus maze and also reduced blood cortisol levels.

The first clinical study in depression and anxiety was published by Schiffer *et al.* in 2009 (Schiffer *et al.*, 2009). They used a fairly small area 1 W 810 nm LED array applied to the forehead in patients with major depression and anxiety. They found improvements in the Hamilton depression rating scale (HAM-D), and the Hamilton anxiety rating scale (HAM-A), 2 weeks after a single treatment. They also found increases in frontal pole regional cerebral blood flow (rCBF) during the light delivery using a commercial NIR spectroscopy device. Cassano and co-workers (Cassano *et al.*, 2015) used tPBM with an 810 nm laser (700 mW/cm² and a fluence of 84 J/cm² delivered per session for 6 sessions) in patients with major depression. Baseline mean HAM-D17 scores decreased from 19.8 ± 4.4 (SD) to 13 ± 5.35 (SD) after treatment ($P=0.004$).

Enhancement of Performance

PBM has been studied in two broad areas of performance enhancement, namely athletic performance and cognitive performance.

PBM for muscle and athletic performance

PBM has been used for prevention of muscle damage after exercise, including delayed onset muscle soreness, and improvement in muscle performance, demonstrated by increases in muscle torque, power or work; workload capacity; fatigue resistance; functional or sport activity, and exercise recovery.

The use of PBM to prevent muscle damage has been investigated in animal models, irradiating skeletal muscles before a bout of intense exercise (called in this chapter “muscular pre-conditioning”) and by following the muscle damage by measuring the creatine kinase (CK) levels in the bloodstream. The first study that used photobiomodulation by PBM to prevent muscle damage was carried out by Lopes-Martins *et al.* (2006) in rats. These authors investigated the effects of different doses of 655 nm laser (0.5 J/cm²; 1.0 J/cm² and 2.5 J/cm²) to prevent muscle fatigue and muscle damage (CK) induced by neuromuscular electrical stimulation. This study reported a PBM dose response to decrease CK activity.

Another experimental study used exercise-trained rats on an inclined treadmill and measured inhibition of inflammation, reduction of CK activity and reduction of oxidative stress. They also found increases in defense against oxidative stress (increased activity of superoxide dismutase) 24 h and 48 h after exercise (Gladwin and Shiva, 2009). These previous studies and other similar reports (de Almeida *et al.*, 2011; Leal *et al.*, 2010a,b; Liu *et al.*, 2009; Lopes-Martins *et al.*, 2006; Santos *et al.*, 2014; Sussai *et al.*, 2010) were important to establish the scientific basis to conduct clinical trials for the prevention of muscle damage by PBM.

Regarding clinical trials of improved muscle performance and exercise recovery, the first studies published were on delayed onset muscle soreness (Douris *et al.*, 2006; Vinck *et al.*, 2006) and resistance to muscle fatigue (Gorgey *et al.*, 2008; Leal *et al.*, 2008) during one or a few bouts of exercise; and improved muscle strength (Ferraresi *et al.*, 2011) and resistance to muscle fatigue (Vieira *et al.*, 2012) after exercise training programs combined with PBM. PBM has been tested in real life athletic competition such as in a team of volleyball players (Ferraresi *et al.*, 2015). A randomized double-blind placebo-controlled trial delivered PBM using LEDs (25 clusters of 4 infrared LEDs 850 nm, 130 mW and 25 clusters of 4 red LEDs 630 nm; 80 mW). PBM was applied over the *quadriceps femoris* muscles, hamstrings, and *triceps surae* of volleyball players before official matches. A professional male volleyball team (12 athletes) was randomized to receive one of four different total doses over each muscle group in a double-blind protocol. PBM at a dose of 210 J or higher avoided significant increases in creatine kinase (a marker of muscle damage) that were found in the placebo group.

PBM for enhancement of cognitive performance

The second area of performance enhancement by PBM (that is not yet as advanced as athletic performance) is the possibility of improving cognitive function in normal people or healthy animals using PBM. The first report was in middle aged (12 months) CD1 female mice (Michalikova *et al.*, 2008). Exposure of the mice to 1072 nm LED arrays led to improved performance in a 3D maze compared to sham treated age-matched controls. Francisco Gonzalez-Lima at the University of Texas Austin, has worked in this area for some time (Gonzalez-Lima and Barrett, 2014). Working in rats they showed that transcranial PBM (9 mW/cm² with 660 nm LED array) induced a dose-dependent increase in oxygen consumption of 5% after 1 J/cm² and 16% after 5 J/cm² (Rojas *et al.*, 2012). They also found that tPBM reduced fear renewal and prevented the reemergence of extinguished conditioned fear responses (Rojas *et al.*, 2012).

In normal human volunteers Gonzalez-Lima used transcranial PBM (1064 nm laser, 60 J/cm² at 250 mW/cm²) delivered to the forehead in a placebo-controlled, randomized study, to influence cognitive tasks related to the prefrontal cortex, including a psychomotor vigilance task (PVT), a delayed match-to-sample (DMS) memory task, and the positive and negative affect schedule (PANAS-X) to show improved mood (Barrett and Gonzalez-Lima, 2013). Subsequent studies in normal humans showed that tPBM with 1064 nm laser could improve performance in the Wisconsin Card Sorting Task (considered the gold standard test for executive function) (Blanco *et al.*, 2015). They also showed that tPBM to the right forehead (but not the left forehead) had better effects on improving attention bias modification (ABM) in humans with depression (Disner *et al.*, 2016).

A study by Salgado *et al.* used transcranial LED PBM on cerebral blood flow in healthy elderly women analyzed by transcranial Doppler ultrasound (TCD) of the right and left middle cerebral artery and basilar artery. Twenty-five non-institutionalized elderly women (mean age 72 years old), with cognitive status > 24, were assessed using TCD before and after transcranial LED therapy. tPBM (627 nm, 70 mW/cm², 10 J/cm²) was performed at four points of the frontal and parietal region for 30 s each twice a week for 4 weeks. There was a significant increase in the systolic and diastolic velocity of the left middle cerebral artery (25 and 30%,

respectively) and the basilar artery (up to 17 and 25%), as well as a decrease in the pulsatility index and resistance index values of the three cerebral arteries analyzed (Salgado *et al.*, 2015).

Conclusions

The goal of this chapter has been to argue the case that different aspects of photomedicine can exert two complementary, but completely opposite therapeutic effects. PDT can kill undesirable cells such as cancer cells and pathogenic microorganisms and remove unwanted tissue, while PBM can heal, restore and protect tissues that have been damaged or are at risk. Ironically both of these therapeutic approaches often use remarkably similar light parameters, namely wavelengths of red light (between 600 and 700 nm) at power densities and fluences that are also surprisingly comparable. However PBM is more versatile than PDT, which has an upper wavelength limit of about 800 nm. PBM on the other hand works very well with NIR light above 800 nm. In fact 810 nm is probably the single most popular wavelength for PBM. Several commentators have pointed out that the mechanisms of action of PDT and PBM have aspects in common; specifically the production of ROS (Lubart and Friedmann, 2011). However the amount of ROS produced is very different. In PDT the type of ROS (predominantly singlet oxygen and hydroxyl radicals produced in large quantities) produced by exogenously added PS are highly damaging to cells. These ROS are produced by photochemical reactions occurring between the administered PS and the light. By contrast the much lower amounts of ROS produced in PBM, arise from the action of light on the mitochondria inside the cells containing intrinsic chromophores, and are much less damaging, their duration is short, and they tend to stimulate protective responses and trigger signaling pathways.

We can illustrate the dual simultaneous use of PDT and PBM by describing an application of photomedicine in dentistry. Periodontitis is caused by pathogenic bacteria that accumulate in the dental pocket (between the tooth and the gums). An inflammatory response to these bacteria leads to production of matrix metalloproteinase enzymes that degrade the periodontal ligament eventually leading to tooth loss. PDT is used to treat periodontitis by injecting a small amount of a solution of methylene blue into the dental pocket and inserting a narrow fiber optic to deliver red light to excite the dye and kill the bacteria. However the red light spreads throughout the gingival tissue (containing no MB) where it acts to reduce inflammation and stimulate healing of the periodontal ligament. This dual action mechanism may explain why PDT is so effective for this condition.

References

- Abdel-Hady, E.S., Martin-Hirsch, P., Duggan-Keen, M., *et al.*, 2001. Immunological and viral factors associated with the response of vulval intraepithelial neoplasia to photodynamic therapy. *Cancer Res.* 61 (1), 192–196.
- Abergel, R.P., Lyons, R.F., Castel, J.C., Dwyer, R.M., Uitto, J., 1987a. Biostimulation of wound healing by lasers: Experimental approaches in animal models and in fibroblast cultures. *J. Dermatol. Surg. Oncol.* 13, 127–133.
- Abergel, R.P., Lyons, R.F., Castel, J.C., Dwyer, R.M., Uitto, J., 1987b. Biostimulation of wound healing by lasers: Experimental approaches in animal models and in fibroblast cultures. *J. Dermatol. Surg. Oncol.* 13 (2), 127–133.
- Agostinis, P., Berg, K., Cengel, K.A., *et al.*, 2011. Photodynamic therapy of cancer: An update. *CA Cancer J. Clin.* 61 (4), 250–281.
- Albarracin, R., Eells, J., Valter, K., 2011. Photobiomodulation protects the retina from light-induced photoreceptor degeneration. *Invest. Ophthalmol. Vis. Sci.* 52 (6), 3582–3592.
- Allison, B.A., Pritchard, P.H., Levy, J.G., 1994. Evidence for low-density lipoprotein receptor-mediated uptake of benzoporphyrin derivative. *Br. J. Cancer* 69 (5), 833–839.
- Almeida-Lopes, L., Rigau, J., Zangaro, R.A., *et al.*, 2001. Comparison of the low level laser therapy effects on cultured human gingival fibroblasts proliferation using different irradiance and same fluence. *Lasers Surg. Med.* 29, 179–184.
- Anders, J.J., Lanzafame, R.J., Arany, P.R., 2015. Low-level light/laser therapy versus photobiomodulation therapy. *Photomed. Laser Surg.* 33 (4), 183–184.
- Ando, T., Xuan, W., Xu, T., *et al.*, 2011. Comparison of therapeutic effects between pulsed and continuous wave 810-nm wavelength laser irradiation for traumatic brain injury in mice. *PLoS One* 6 (10), e26212.
- Anisman, H., Matheson, K., 2005. Stress, depression, and anhedonia: Caveats concerning animal models. *Neurosci. Biobehav. Rev.* 29 (4-5), 525–546.
- Anneroth, G., Hall, G., Ryden, H., Zetterqvist, L., 1988. The effect of low-energy infra-red laser radiation on wound healing in rats. *Br. J. Oral Maxillofac. Surg.* 26, 12–17.
- Barcia, C., 2012. Who else was intoxicated with MPTP in Santa Clara? *Parkinsonism Relat. Disord.* 18 (9), 1005–1006.
- Barolet, D., Roberge, C.J., Auger, F.A., Boucher, A., Germain, L., 2009. Regulation of skin collagen metabolism in vitro using a pulsed 660 nm LED light source: Clinical correlation with a single-blinded study. *J. Invest. Dermatol.* 129 (12), 2751–2759.
- Barrett, D.W., Gonzalez-Lima, F., 2013. Transcranial infrared laser stimulation produces beneficial cognitive and emotional effects in humans. *Neuroscience* 230, 13–23.
- Baxter, G.D., Bell, A.J., Allen, J.M., Ravey, J., 1991. Low level laser therapy: Current clinical practice in Northern Ireland. *Physiotherapy* 77 (3), 171–178.
- Bellnier, D.A., Greco, W.R., Parsons, J.C., *et al.*, 1997. An assay for the quantitation of Photofrin in tissues and fluids. *Photochem. Photobiol.* 66 (2), 237–244.
- Bellnier, D.A., Ho, Y.K., Pandey, R.K., Missert, J.R., Dougherty, T.J., 1989. Distribution and elimination of Photofrin II in mice. *Photochem. Photobiol.* 50 (2), 221–228.
- Bervenuti, F., 2016. The dendritic cell synapse: A life dedicated to T cell activation. *Front. Immunol.* 7, 70.
- Bernstein, E.F., 2005. Hair growth induced by diode laser treatment. *Dermatol. Surg.* 31 (5), 584–586.
- Bertoloni, G., Rossi, F., Valduga, G., Jori, G., van Lier, J., 1990. Photosensitizing activity of water- and lipid-soluble phthalocyanines on *Escherichia coli*. *FEMS Microbiol. Lett.* 59 (1-2), 149–155.
- Bhargava, A., Mishra, D., Banerjee, S., Mishra, P.K., 2012. Dendritic cell engineering for tumor immunotherapy: From biology to clinical translation. *Immunotherapy* 4 (7), 703–718.
- Bhat, J., Birch, J., Whitehurst, C., Lanigan, S.W., 2005. A single-blinded randomised controlled study to determine the efficacy of Omnilux Revive facial treatment in skin rejuvenation. *Lasers Med. Sci.* 20 (1), 6–10.
- Bhuvaneshwari, R., Gan, Y.Y., Soo, K.C., Olivo, M., 2009. The effect of photodynamic therapy on tumor angiogenesis. *Cell. Mol. Life Sci.* 66 (14), 2275–2283.
- Bisht, D., Gupta, S.C., Mishra, V., *et al.*, 1994. Effect of low intensity laser radiation on healing of open skin wounds in rats. *Indian J. Med. Res.* 100, 43–46.
- Bisht, D., Mehroratra, R., Singh, P.A., Atri, S.C., Kumar, A., 1999. Effect of helium-neon laser on wound healing. *Indian J. Exp. Biol.* 37, 187–189.

- Bjoridal, J.M., Couppe, C., Chow, R.T., Tuner, J., Ljunggren, E.A., 2003. A systematic review of low level laser therapy with location-specific doses for pain from chronic joint disorders. *Aust. J. Physiother.* 49, 107–116.
- Bjornsti, M.A., Houghton, P.J., 2004. The TOR pathway: A target for cancer therapy. *Nat. Rev. Cancer* 4 (5), 335–348.
- Blanco, N.J., Maddox, W.T., Gonzalez-Lima, F., 2015. Improving executive function using transcranial infrared laser stimulation. *J. Neuropsychol.*
- Bland, J., 1976. Biochemical effects of excited state molecular oxygen. *J. Chem. Educ.* 53 (5), 5.
- Boonswang, N.A., Chicchi, M., Lukachek, A., Curtiss, D., 2012. A new treatment protocol using photobiomodulation and muscle/bone/joint recovery techniques having a dramatic effect on a stroke patient's recovery: a new weapon for clinicians. *BMJ Case Rep.* 2012.
- Bouzari, E., Firooz, A.R., 2006. Lasers may induce terminal hair growth. *Dermatol. Surg.* 32 (3), 460.
- Boyle, R.W., Dolphin, D., 1996. Structure and biodistribution relationships of photodynamic sensitizers. *Photochem. Photobiol.* 64 (3), 469–485.
- Brown, S.A., Rohrich, R.J., Kenkel, J., *et al.*, 2004. Effect of low-level laser therapy on abdominal adipocytes before lipoplasty procedures. *Plast Reconstr Surg* 113 (6), 1796–1804. (discussion 1805-6).
- Brun, P.H., DeGroot, J.L., Dickson, E.F., Farahani, M., Pottier, R.H., 2004. Determination of the in vivo pharmacokinetics of palladium-bacteriopheophorbide (WST09) in EMT6 tumour-bearing Balb/c mice using graphite furnace atomic absorption spectroscopy. *Photochem. Photobiol. Sci.* 3 (11-12), 1006–1010.
- Buettner, G.R., 1993. The pecking order of free radicals and antioxidants: lipid peroxidation, alpha-tocopherol, and ascorbate. *Arch. Biochem. Biophys.* 300 (2), 535–543.
- Bush, K., Courvalin, P., Dantas, G., *et al.*, 2011. Tackling antibiotic resistance. *Nat. Rev. Microbiol.* 9 (12), 894–896.
- Buytaert, E., Callewaert, G., Hendrickx, N., *et al.*, 2006a. Role of endoplasmic reticulum depletion and multidomain proapoptotic BAX and BAK proteins in shaping cell death after hypericin-mediated photodynamic therapy. *FASEB J* 20 (6), 756–758.
- Buytaert, E., Callewaert, G., Vandenhede, J.R., Agostinis, P., 2006b. Deficiency in apoptotic effectors Bax and Bak reveals an autophagic cell death pathway initiated by photodamage to the endoplasmic reticulum. *Autophagy* 2 (3), 238–240.
- Buytaert, E., Dewaele, M., Agostinis, P., 2007. Molecular effectors of multiple cell death pathways initiated by photodynamic therapy. *Biochim. Biophys. Acta* 1776 (1), 86–107.
- Canti, G.L., Lattuada, D., Nicolin, A., *et al.*, 1994. Immunopharmacology studies on photosensitizers used in photodynamic therapy. *Proc. SPIE* 2078, 268–275.
- Caruso-Davis, M.K., Guillot, T.S., Podichetty, V.K., *et al.*, 2011. Efficacy of low-level laser therapy for body contouring and spot fat reduction. *Obes. Surg.* 21 (6), 722–729.
- Cassano, P., Cusin, C., Mischoulon, D., *et al.*, 2015. Near-infrared transcranial radiation for major depressive disorder: Proof of concept study. *Psychiatry J.* 2015, 352979.
- Castano, A.P., Dai, T., Yaroslavsky, I., *et al.*, 2007. Low-level laser therapy for zymosan-induced arthritis in rats: Importance of illumination time. *Lasers Surg. Med.* 39 (6), 543–550.
- Castellani, A., Pace, G.P., Concioli, M., 1963. Photodynamic effect of haematoporphyrin on blood microcirculation. *J. Pathol. Bacteriol.* 86, 99–102.
- Castellino, F., Germain, R.N., 2006. Cooperation between CD4 + and CD8 + T cells: When, where, and how. *Annu. Rev. Immunol.* 24, 519–540.
- Castex-Rizzi, N., Lachgar, S., Charveron, M., Gall, Y., 2002. Implication of VEGF, steroid hormones and neuropeptides in hair follicle cell responses. *Ann. Dermatol. Venerol.* 129 (5 Pt 2), 783–786.
- Cengel, K.A., Simone 2nd, C.B., Glatstein, E., 2016. PDT: What's past is prologue. *Cancer Res* 76 (9), 2497–2499.
- Chen, A.C., Arany, P.R., Huang, Y.Y., *et al.*, 2011. Low-level laser therapy activates NF- κ B via generation of reactive oxygen species in mouse embryonic fibroblasts. *PLoS One* 6 (7), e22453.
- Chow, R.T., Johnson, M.I., Lopes-Martins, R.A., Bjoridal, J.M., 2009. Efficacy of low-level laser therapy in the management of neck pain: A systematic review and meta-analysis of randomised placebo or active-treatment controlled trials. *Lancet* 374, 1897–1908.
- Cory, S., Adams, J.M., 2002. The Bcl2 family: regulators of the cellular life-or-death switch. *Nat. Rev. Cancer* 2 (9), 647–656.
- Cotter, T.G., 2009. Apoptosis and cancer: The genesis of a research field. *Nat. Rev. Cancer* 9 (7), 501–507.
- Coupienne, I., Piette, J., Bontems, S., 2010. How to monitor NF-kappaB activation after photodynamic therapy. *Methods Mol. Biol.* 635, 79–95.
- Dahle, J., Steen, H.B., Moan, J., 1999. The mode of cell death induced by photodynamic treatment depends on cell density. *Photochem. Photobiol.* 70 (3), 363–367.
- Dahl, T.A., Midden, W.R., Hartman, P.E., 1989. Comparison of killing of gram-negative and gram-positive bacteria by pure singlet oxygen. *J. Bacteriol.* 171 (4), 2188–2194.
- Dai, T., Huang, Y.Y., Hamblin, M.R., 2009. Photodynamic therapy for localized infections-state of the art. *Photodiagnosis Photodyn. Ther.* 6 (3-4), 170–188.
- Daniel, N.N., Korsmeyer, S.J., 2004. Cell death: critical control points. *Cell* 116 (2), 205–219.
- Davies, M.J., 2016. Protein oxidation and peroxidation. *Biochem. J.* 473 (7), 805–825.
- de Almeida, P., Lopes-Martins, R.A., Tomazoni, S.S., *et al.*, 2011. Low-level laser therapy improves skeletal muscle performance, decreases skeletal muscle damage and modulates mRNA expression of COX-1 and COX-2 in a dose-dependent manner. *Photochem. Photobiol.* 87 (5), 1159–1163.
- Debeve, E., Cheng, C., Schaefer, S.C., *et al.*, 2010. Photodynamic therapy induces selective extravasation of macromolecules: insights using intravital microscopy. *J. Photochem. Photobiol. B* 98 (1), 69–76.
- de Melo, W.C., Avci, P., de Oliveira, M.N., *et al.*, 2013. Photodynamic inactivation of biofilm: taking a lightly colored approach to stubborn infection. *Expert Rev. Anti Infect. Ther.* 11 (7), 669–693.
- Demidova-Rice, T.N., Salomatina, E.V., Yaroslavsky, A.N., Herman, I.M., Hamblin, M.R., 2007. Low-level light stimulates excisional wound healing in mice. *Lasers Surg. Med.* 39 (9), 706–715.
- De Taboada, L., Yu, J., El-Amouri, S., *et al.*, 2011. Transcranial laser therapy attenuates amyloid-beta peptide neuropathology in amyloid-beta protein precursor transgenic mice. *J. Alzheimers Dis.* 23 (3), 521–535.
- de Vree, W.J., Essers, M.C., Koster, J.F., Sluiter, W., 1997. Role of interleukin 1 and granulocyte colony-stimulating factor in photofrin-based photodynamic therapy of rat rhabdomyosarcoma tumors. *Cancer Res.* 57 (13), 2555–2558.
- Dierickx, C.C., Anderson, R.R., 2005. Visible light treatment of photoaging. *Dermatol. Ther.* 18 (3), 191–208.
- Disner, S.G., Beever, C.G., Gonzalez-Lima, F., 2016. Transcranial laser stimulation as neuroenhancement for attention bias modification in adults with elevated depression symptoms. *Brain Stimul.*
- Doidge, N., 2015. *The Brain's Way of Healing: remarkable Discoveries and Recoveries from the Frontiers of Neuroplasticity*. New York, NY: Viking Press.
- Dolmans, D.E., Kadambi, A., Hill, J.S., *et al.*, 2002. Vascular accumulation of a novel photosensitizer, MV6401, causes selective thrombosis in tumor vessels after photodynamic therapy. *Cancer Res.* 62 (7), 2151–2156.
- Dong, T., Zhang, Q., Hamblin, M.R., Wu, M.X., 2015. Low-level light in combination with metabolic modulators for effective therapy of injured brain. *J. Cereb. Blood Flow Metabolism: Off. J. Int. Soc. Cereb. Blood Flow Metabolism*.
- Dougherty, T.J., Gomer, C.J., Henderson, B.W., *et al.*, 1998. Photodynamic therapy. *J. Natl. Cancer Inst.* 90 (12), 889–905.
- Douris, P., Southard, V., Ferrigi, R., *et al.*, 2006. Effect of phototherapy on delayed onset muscle soreness. *Photomed. Laser Surg.* 24 (3), 377–382.
- Dragieva, G., Hafner, J., Dummer, R., *et al.*, 2004. Topical photodynamic therapy in the treatment of actinic keratoses and Bowen's disease in transplant recipients. *Transplantation* 77 (1), 115–121.
- Eells, J.T., Gopalakrishnan, S., Valter, K., 2016. Near-infrared photobiomodulation in retinal injury and disease. *Adv. Exp. Med. Biol.* 854, 437–441.
- Eells, J.T., Henry, M.M., Summerfelt, P., *et al.*, 2003. Therapeutic photobiomodulation for methanol-induced retinal toxicity. *Proc. Natl. Acad. Sci. USA* 100 (6), 3439–3444.
- Egorin, M.J., Zuhowski, E.G., Sentz, D.L., *et al.*, 1999. Plasma pharmacokinetics and tissue distribution in CD2F1 mice of Pc4 (NSC 676418), a silicone phthalocyanine photodynamic sensitizing agent. *Cancer Chemother. Pharmacol.* 44 (4), 283–294.
- Eichler, M., Lavi, R., Shainberg, A., Lubart, R., 2005. Flavins are source of visible-light-induced free radical formation in cells. *Lasers Surg. Med.* 37 (4), 314–319.
- El Massri, N., Moro, C., Torres, N., *et al.*, 2016. Near-infrared light treatment reduces astrogliosis in MPTP-treated monkeys. *Exp. Brain Res.*
- F  h, H., 1968. Long-range coherence and energy storage in biological systems. *Int. J. Quantum Chem.* 2 (5), 641–649.

- FAH, A.-W., XY, Z., 1996. Comparison of the effects of laser therapy on wound healing using different laser wavelengths. *Laser Ther* 1996 (8), 127–135.
- Fauci, A.S., Morens, D.M., 2012. The perpetual challenge of infectious diseases. *New England J. Med.* 366 (5), 454–461.
- Ferraresi, C., de Brito Oliveira, T., de Oliveira Zafalon, L., *et al.*, 2011. Effects of low level laser therapy (808 nm) on physical strength training in humans. *Lasers Med. Sci.* 26 (3), 349–358.
- Ferraresi, C., Dos Santos, R.V., Marques, G., *et al.*, 2015. Light-emitting diode therapy (LEDT) before matches prevents increase in creatine kinase with a light dose response in volleyball players. *Lasers Med. Sci.* 30 (4), 1281–1287.
- Ferreira, D.D., Reynoso, E., Cordero, P., *et al.*, 2016. Synthesis and properties of 5,10,15,20-tetrakis[4-(3-N,N-dimethylaminopropoxy)phenyl] chlorin as potential broad-spectrum antimicrobial photosensitizers. *J. Photochem. Photobiol. B* 158, 243–251.
- Fingar, V.H., Kik, P.K., Haydon, P.S., *et al.*, 1999. Analysis of acute vascular damage after photodynamic therapy using benzoporphyrin derivative (BPD). *Br. J. Cancer* 79 (11–12), 1702–1708.
- Fingar, V.H., Wieman, T.J., Wiehle, S.A., Cerrito, P.B., 1992. The role of microvascular damage in photodynamic therapy: the effect of treatment on vessel constriction, permeability, and leukocyte adhesion. *Cancer Res.* 52 (18), 4914–4921.
- Freitas, I., 1990. Lipid accumulation: the common feature to photosensitizer-retaining normal and malignant tissues [news]. *J. Photochem. Photobiol. B* 7 (2–4), 359–361.
- Friedberg, J.S., Mick, R., Stevenson, J.P., *et al.*, 2004. Phase II trial of pleural photodynamic therapy and surgery for patients with non-small-cell lung cancer with pleural spread. *J. Clin. Oncol.* 22 (11), 2192–2201.
- Frisoli, J.K., Tudor, E.G., Flotte, T.J., *et al.*, 1993. Pharmacokinetics of a fluorescent drug using laser-induced fluorescence. *Cancer Res.* 53 (24), 5954–5961.
- Frohlich, H., 1970. Long range coherence and the action of enzymes. *Nature* 228 (5276), 1093.
- Frohlich, H., 1975. The extraordinary dielectric properties of biological materials and the action of enzymes. *Proc. Natl Acad. Sci. USA* 72 (11), 4211–4215.
- Gad, F., Zahra, T., Francis, K.P., Hasan, T., Hamblin, M.R., 2004. Targeted photodynamic therapy of established soft-tissue infections in mice. *Photochem. Photobiol. Sci.* 3 (5), 451–458.
- Garg, A.D., Galluzzi, L., Apetoh, L., *et al.*, 2015. Molecular and translational classifications of DAMPs in immunogenic cell death. *Front. Immunol.* 6, 588.
- Garg, A.D., Krysko, D.V., Vandenabeele, P., Agostinis, P., 2011. DAMPs and PDT-mediated photo-oxidative stress: exploring the unknown. *Photochem. Photobiol. Sci.* 10 (5), 670–680.
- Geiger, P.G., Korytowski, W., Girotti, A.W., 1995. Photodynamically generated 3-beta-hydroxy-5 alpha-cholest-6-ene-5- hydroperoxide: toxic reactivity in membranes and susceptibility to enzymatic detoxification. *Photochem. Photobiol.* 62 (3), 580–587.
- Geneva, I.I., 2016. Photobiomodulation for the treatment of retinal diseases: a review. *Int. J. Ophthalmol.* 9 (1), 145–152.
- Geronemus, R.G., Weiss, R.A., Weiss, M.A., *et al.*, 2003. Non-ablative LED photomodulation light activated fibroblast stimulation clinical trial. *Lasers Surg. Med.* 25, 22.
- Giuliani, F., Martinelli, M., Cocchi, A., *et al.*, 2010. In vitro resistance selection studies of RLP068/Cl, a new Zn(II) phthalocyanine suitable for antimicrobial photodynamic therapy. *Antimicrob Agents Chemother.* 54 (2), 637–642.
- Gladwin, M.T., Shiva, S., 2009. The ligand binding battle at cytochrome c oxidase: how no regulates oxygen gradients in tissue. *Circ. Res.* 104 (10), 1136–1138.
- Goldman, L., Blaney, D.J., Kindel Jr., D.J., Franke, E.K., 1963. Effect of the laser beam on the skin. [Preliminary report] *J Invest Dermatol* 40, 121–122.
- Gollnick, S.O., Evans, S.S., Baumann, H., *et al.*, 2003. Role of cytokines in photodynamic therapy-induced local and systemic inflammation. *Br J Cancer* 88 (11), 1772–1779.
- Gollnick, S.O., Kabingu, E., Kousis, P.C., Henderson, B.W., 2004. Stimulation of the host immune response by photodynamic therapy (PDT). *Proc SPIE* 5319, 60–70.
- Gollnick, S.O., Owczarczak, B., Maier, P., 2006. Photodynamic therapy and anti-tumor immunity. *Lasers Surg Med* 38 (5), 509–515.
- Gollnick, S.O., Vaughan, L., Henderson, B.W., 2002. Generation of effective antitumor vaccines using photodynamic therapy. *Cancer Res.* 62 (6), 1604–1608.
- Gomer, C.J., Ferrario, A., Murphree, A.L., 1987. The effect of localized porphyrin photodynamic therapy on the induction of tumour metastasis. *Br J Cancer* 56 (1), 27–32.
- Gomer, C.J., Rucker, N., Murphree, A.L., 1988. Differential cell photosensitivity following porphyrin photodynamic therapy. *Cancer Res.* 48 (16), 4539–4542.
- Gomer, C.J., Ryter, S.W., Ferrario, A., *et al.*, 1996. Photodynamic therapy-mediated oxidative stress can induce expression of heat shock proteins. *Cancer Res.* 56 (10), 2355–2360.
- Gonzalez-Lima, F., Barrett, D.W., 2014. Augmentation of cognitive brain functions with transcranial lasers. *Front. Syst. Neurosci.* 8, 36.
- Goodson, W.H., Hunt, T.K., 1979. Wound healing and the diabetic patient. *Surg. Gynecol. Obstet.* 149, 600–608.
- Gorgey, A.S., Wade, A.N., Sobhi, N.N., 2008. The effect of low-level laser therapy on electrically induced muscle fatigue: A pilot study. *Photomed. Laser Surg* 26 (5), 501–506.
- Gougeon, M.L., Kroemer, G., 2003. Charming to death: caspase-dependent or -independent? *Cell Death Differ* 10 (3), 390–392.
- Gudgin Dickson, E.F., Holmes, H., Jori, G., *et al.*, 1995. On the source of the oscillations observed during in vivo zinc phthalocyanine fluorescence pharmacokinetic measurements in mice. *Photochem. Photobiol.* 61 (5), 506–509.
- Guyer, B., Freedman, M.A., Strobino, D.M., Sondik, E.J., 2000. Annual summary of vital statistics: Trends in the health of Americans during the 20th century. *Pediatrics* 106 (6), 1307–1317.
- Halliwell, B., Chirico, S., 1993. Lipid peroxidation: its mechanism, measurement, and significance. *Am. J. Clin. Nutr.* 57 (5 Suppl), 715S–724S. (discussion 724S–725S).
- Halliwell, B., Gutteridge, J.M.C., 2015. *Free Radicals in Biology and Medicine*, 5 edn, 1. Oxford: Oxford Press.
- Hamblin, M.R., Dai, T., 2010. Can surgical site infections be treated by photodynamic therapy? *Photodiagnosis Photodyn. Ther.* 7 (2), 134–136.
- Hamblin, M.R., Hasan, T., 2004. Photodynamic therapy: A new antimicrobial approach to infectious disease? *Photochem. Photobiol. Sci.* 3 (5), 436–450.
- Hamblin, M.R., Mroz, P., 2008. *Advances in Photodynamic Therapy: Basic, Translational and Clinical*. Norwood, MA: Artech House.
- Hamblin, M.R., Newman, E.L., 1994. On the mechanism of the tumour-localising effect in photodynamic therapy. *J Photochem. Photobiol. B* 23 (1), 3–8.
- Hashmi, J.T., Huang, Y.Y., Sharma, S.K., *et al.*, 2010. Effect of pulsing in low-level light therapy. *Lasers Surg. Med.* 42 (6), 450–466.
- Hawkins, D., Abrahamse, H., 2005. Biological effects of helium-neon laser irradiation on normal and wounded human skin fibroblasts. *Photomed. Laser Surg.* 23, 251–259.
- Hawkins, D., Houreld, N., Abrahamse, H., 2005. Low level laser therapy (LLLT) as an effective therapeutic modality for delayed wound healing. *Ann. NY Acad. Sci.* 1056, 486–493.
- He, C., Agharkar, P., Chen, B., 2008. Intravital microscopic analysis of vascular perfusion and macromolecule extravasation after photodynamic vascular targeting therapy. *Pharmaceutical Res.* 25 (8), 1873–1880.
- Henderson, B.W., Fingar, V.H., 1989. Oxygen limitation of direct tumor cell kill during photodynamic treatment of a murine tumor model. *Photochem. Photobiol.* 49 (3), 299–304.
- Henderson, B.W., Gollnick, S.O., 2003. Mechanistic Principles of Photodynamic Therapy. In: Vo-Dinh, T. (Ed.), *Biomedical Photonics Handbook*. Boca Raton, FL: CRC Press, pp. 36.1–36.27.
- Henderson, B.W., Gollnick, S.O., Snyder, J.W., *et al.*, 2004. Choice of oxygen-conserving treatment regimen determines the inflammatory response and outcome of photodynamic therapy of tumors. *Cancer Res.* 64 (6), 2120–2126.
- Henderson, B.W., Miller, A.C., 1986. Effects of scavengers of reactive oxygen and radical species on cell survival following photodynamic treatment in vitro: comparison to ionizing radiation. *Radiat. Res.* 108 (2), 196–205.
- Henderson, B.W., Waldow, S.M., Mang, T.S., *et al.*, 1985. Tumor destruction and kinetics of tumor cell death in two experimental mouse tumors following photodynamic therapy. *Cancer Res.* 45 (2), 572–576.
- Henderson, T.A., Morris, L.D., 2015. SPECT perfusion imaging demonstrates improvement of traumatic brain injury with transcranial near-infrared laser phototherapy. *Adv. Mind Body Med.* 29 (4), 27–33.
- Hou, Y.C., Janczuk, A., Wang, P.G., 1999. Current trends in the development of nitric oxide donors. *Curr. Pharm. Des.* 5 (6), 417–441.

- Huang, Y.Y., Chen, A.C., Carroll, J.D., Hamblin, M.R., 2009. Biphasic dose response in low level light therapy. *Dose Response* 7 (4), 358–383.
- Huisa, B.N., Stemer, A.B., Walker, M.G., *et al.*, 2013. Transcranial laser therapy for acute ischemic stroke: a pooled analysis of NEST-1 and NEST-2. *Int. J. Stroke* 8 (5), 315–320.
- Hunt, D.W., Levy, J.G., 1998. Immunomodulatory aspects of photodynamic therapy. *Expert Opin. Investig. Drugs* 7 (1), 57–64.
- Y.D. Ignatov, A.I.V., Vlasov, T.D., *et al.*, 2005. Effects of helium-neon laser irradiation and local anesthetics on potassium channels in pond snail neurons. *Neurosci. Behav. Physiol.* 35, 871–875.
- Igney, F.H., Krammer, P.H., 2002a. Death and anti-death: Tumour resistance to apoptosis. *Nat. Rev. Cancer* 2 (4), 277–288.
- Igney, F.H., Krammer, P.H., 2002b. Immune escape of tumors: apoptosis resistance and tumor counterattack. *J. Leukoc. Biol.* 71 (6), 907–920.
- Ivancic, B.T., Ivancic, T., 2008. Low-level laser therapy improves vision in patients with age-related macular degeneration. *Photomed. Laser Surg.* 26 (3), 241–245.
- Ivancic, B.T., Ivancic, T., 2012. Low-level laser therapy improves visual acuity in adolescent and adult patients with amblyopia. *Photomed. Laser Surg.* 30 (3), 167–171.
- Ivancic, B.T., Ivancic, T., 2014. Low-level laser therapy improves vision in a patient with retinitis pigmentosa. *Photomed. Laser Surg.* 32 (3), 181–184.
- Jalili, A., Makowski, M., Switaj, T., *et al.*, 2004. Effective photoimmunotherapy of murine colon carcinoma induced by the combination of photodynamic therapy and dendritic cells. *Clin. Cancer Res.* 10 (13), 4498–4508.
- J.I. O., 2000. *Energy Medicine: The Scientific Basis*, 1st edn. Edinburgh: Churchill Livingstone.
- Johnstone, D.M., el Massri, N., Moro, C., *et al.*, 2014. Indirect application of near infrared light induces neuroprotection in a mouse model of parkinsonism – an abscopal neuroprotective effect. *Neuroscience* 274, 93–101.
- Johnstone, D.M., Moro, C., Stone, J., Benabid, A.L., Mitrofanis, J., 2015. Turning on lights to stop neurodegeneration: the potential of near infrared light therapy in Alzheimer's and Parkinson's disease. *Front. Neurosci.* 9, 500.
- Jori, G., 1989. In vivo transport and pharmacokinetic behavior of tumour photosensitizers. *Ciba Found. Symp.* 146, 78–86.
- Jori, G., Reddi, E., 1993. The role of lipoproteins in the delivery of tumour-targeting photosensitizers. *Int. J. Biochem.* 25 (10), 1369–1375.
- Kabingu, E., Oseroff, A.R., Wilding, G.E., Gollnick, S.O., 2009. Enhanced systemic immune reactivity to a Basal cell carcinoma associated antigen following photodynamic therapy. *Clin. Cancer Res.* 15 (13), 4460–4466.
- Kabingu, E., Vaughan, L., Owczarczak, B., Ramsey, K.D., Gollnick, S.O., 2007. CD8+ T cell-mediated control of distant tumours following local photodynamic therapy is independent of CD4+ T cells and dependent on natural killer cells. *Br. J. Cancer* 96 (12), 1839–1848.
- Kana, J.S., Hutschenreiter, G., Haina, D., Waidelich, W., 1981. Effect of low-power density laser radiation on healing of open skin wounds in rats. *Arch. Surg.* 116, 293–296.
- Karu, T.I., 2008. Mitochondrial signaling in mammalian cells activated by red and near-IR radiation. *Photochem. Photobiol.* 84 (5), 1091–1099.
- Karu, T.I., Pyatibrat, L.V., Afanasyeva, N.I., 2004. A novel mitochondrial signaling pathway activated by visible-to-near infrared radiation. *Photochem. Photobiol.* 80 (2), 366–372.
- Karu, T.I., Pyatibrat, L.V., Afanasyeva, N.I., 2005. Cellular effects of low power laser therapy can be mediated by nitric oxide. *Lasers Surg. Med.* 36 (4), 307–314.
- Kearns, D.R., 1971. Physical and chemical properties of singlet molecular oxygen. *Chem. Rev.* 71 (4), 32.
- Kessel, D., 2006. Death pathways associated with photodynamic therapy. *Med Laser Appl.* 21 (4), 219–224.
- Kessel, D., Arroyo, A.S., 2007. Apoptotic and autophagic responses to Bcl-2 inhibition and photodamage. *Photochem. Photobiol. Sci.* 6 (12), 1290–1295.
- Kessel, D., Oleinick, N.L., 2009. Initiation of autophagy by photodynamic therapy. *Methods Enzymol.* 453, 1–16.
- Kessel, D., Poretz, R.D., 2000. Sites of photodamage induced by photodynamic therapy with a chlorin e6 triacetoxymethyl ester (CAME). *Photochem. Photobiol.* 71 (1), 94–96.
- Kessel, D., Reiners Jr., J.J., 2007. Apoptosis and autophagy after mitochondrial or endoplasmic reticulum photodamage. *Photochem Photobiol* 83 (5), 1024–1028.
- Kessel, D., Vicente, M.G., Reiners Jr., J.J., 2006. Initiation of apoptosis and autophagy by photodynamic therapy. *Lasers Surg. Med.* 38 (5), 482–488.
- Khuman, J., Zhang, J., Park, J., *et al.*, 2012. Low-level laser light therapy improves cognitive deficits and inhibits microglial activation after controlled cortical impact in mice. *J. Neurotrauma* 29 (2), 408–417.
- Khurana, M., Moriyama, E.H., Mariampillai, A., Wilson, B.C., 2008. Intravital high-resolution optical imaging of individual vessel response to photodynamic treatment. *J. Biomed. Opt.* 13 (4), 040502.
- Kikuchi, T., Mogi, M., Okabe, I., *et al.*, 2015. Adjunctive application of antimicrobial photodynamic therapy in nonsurgical periodontal treatment: A review of literature. *Int. J. Mol. Sci.* 16 (10), 24111–24126.
- Kim, S.S., Park, M.W., Lee, C.J., 2007. Phototherapy of androgenetic alopecia with low level narrow band 655-nm red light and 780-nm infrared light. *J. Am. Acad. Dermatol.* 56, AB112. *American Academy of Dermatology Proceedings of the 65th Annual Meeting.*
- Kim, W.S., Calderhead, R.G., 2011. Is light-emitting diode phototherapy (LED-LLLT) really effective? *Laser Ther.* 20 (3), 205–215.
- Kitsis, R.N., Molkenin, J.D., 2010. Apoptotic cell death "Nixed" by an ER-mitochondrial necrotic pathway. *Proc. Natl Acad. Sci. USA* 107 (20), 9031–9032.
- Klaper, M., Fudickar, W., Linker, T., 2016. Role of distance in singlet oxygen applications: A model system. *J. Am. Chem. Soc.* 138 (22), 7024–7029.
- Kolari, P.J., Airaksinen, O., 1993. Poor penetration of infra-red and helium neon low power laser light into the dermal tissue. *Acupunct. Electrother. Res.* 18 (1), 17–21.
- Kongshaug, M., Moan, J., Brown, S.B., 1989. The distribution of porphyrins with different tumour localising ability among human plasma proteins. *Br. J. Cancer* 59 (2), 184–188.
- Korbelik, M., 1992. Low density lipoprotein receptor pathway in the delivery of Photofrin: How much is it relevant for selective accumulation of the photosensitizer in tumors? *J. Photochem. Photobiol. B* 12 (1), 107–109.
- Korbelik, M., 2006. PDT-associated host response and its role in the therapy outcome. *Lasers Surg. Med.* 38 (5), 500–508.
- Korbelik, M., Cecic, I., 1999. Contribution of myeloid and lymphoid host cells to the curative outcome of mouse sarcoma treatment by photodynamic therapy. *Cancer Lett.* 137 (1), 91–98.
- Korbelik, M., Cecic, I., 2003. Mechanism of tumor destruction by photodynamic therapy. In: Nalwa, H.S. (Ed.), *Handbook of Photochemistry and Photobiology*. Stevenson Ranch, CA, USA: American Scientific Publishers, pp. 39–77.
- Korbelik, M., Dougherty, G.J., 1999. Photodynamic therapy-mediated immune response against subcutaneous mouse tumors. *Cancer Res.* 59 (8), 1941–1946.
- Korbelik, M., Krosi, G., 1995. Photofrin accumulation in malignant and host cell populations of a murine fibrosarcoma. *Photochem. Photobiol.* 62 (1), 162–168.
- Korbelik, M., Krosi, G., Krosi, J., Dougherty, G.J., 1996. The role of host lymphoid populations in the response of mouse EMT6 tumor to photodynamic therapy. *Cancer Res.* 56 (24), 5647–5652.
- Korbelik, M., Merchant, S., Huang, N., 2009. Exploitation of immune response-eliciting properties of hypocrellin photosensitizer SL052-based photodynamic therapy for eradication of malignant tumors. *Photochem. Photobiol.* 85 (6), 1418–1424.
- Korbelik, M., Stott, B., Sun, J., 2007. Photodynamic therapy-generated vaccines: Relevance of tumour cell death expression. *Br J Cancer* 97 (10), 1381–1387.
- Korbelik, M., Sun, J., 2006. Photodynamic therapy-generated vaccine for cancer therapy. *Cancer Immunol. Immunother.* 55 (8), 900–909.
- Korbelik, M., Sun, J., Cecic, I., 2005. Photodynamic therapy-induced cell surface expression and release of heat shock proteins: Relevance for tumor response. *Cancer Res.* 65 (3), 1018–1026.
- Kousis, P.C., Henderson, B.W., Maier, P.G., Gollnick, S.O., 2007. Photodynamic therapy6 enhancement of antitumor immunity is regulated by neutrophils. *Cancer Res.* 67 (21), 10501–10510.
- Kraus, C.N., 2008. Low hanging fruit in infectious disease drug development. *Curr. Opin. Microbiol.* 11 (5), 434–438.
- Kroemer, G., Jaattela, M., 2005. Lysosomes and autophagy in cell death control. *Nat Rev Cancer* 5 (11), 886–897.
- Krosi, G., Korbelik, M., Dougherty, G.J., 1995. Induction of immune cell infiltration into murine SCCVII tumour by photofrin-based photodynamic therapy. *Br. J. Cancer* 71 (3), 549–555.

- Kucuk, B.B., Oral, K., Selcuk, N.A., Toklu, T., Civi, O.G., 2010. The anti-inflammatory effect of low-level laser therapy on experimentally induced inflammation of rabbit temporomandibular joint retrodiscal tissues. *J. Orofac. Pain* 24 (3), 293–297.
- Lampl, Y., Zivin, J.A., Fisher, M., *et al.*, 2007. Infrared laser therapy for ischemic stroke: A new treatment strategy: Results of the NeuroThera Effectiveness and Safety Trial-1 (NEST-1). *Stroke* 38 (6), 1843–1849.
- Lapchak, P.A., 2013a. Fast neuroprotection (fast-NPRX) for acute ischemic stroke victims: The time for treatment is now. *Transl. Stroke Res.* 4 (6), 704–709.
- Lapchak, P.A., 2013b. Recommendations and practices to optimize stroke therapy: Developing effective translational research programs. *Stroke* 44 (3), 841–843.
- Lapchak, P.A., Boitano, P.D., 2016. Transcranial near-infrared laser therapy for stroke: How to recover from futility in the NEST-3 clinical trial. *Acta Neurochir. Suppl.* 121, 7–12.
- Lapchak, P.A., Zhang, J.H., Noble-Haesslein, L.J., 2013. RIGOR guidelines: Escalating STAIR and STEPS for effective translational research. *Transl. Stroke Res.* 4 (3), 279–285.
- Larroque, C., Pelegrin, A., Van Lier, J.E., 1996. Serum albumin as a vehicle for zinc phthalocyanine: Photodynamic activities in solid tumour models. *Br J Cancer* 74 (12), 1886–1890.
- Lavie, G., Kaplinsky, C., Toren, A., *et al.*, 1999. A photodynamic pathway to apoptosis and necrosis induced by dimethyl tetrahydroxyhelianthone and hypericin in leukaemic cells: Possible relevance to photodynamic therapy. *Br. J. Cancer* 79 (3–4), 423–432.
- Lavi, R., Shainberg, A., Shneyvays, V., *et al.*, 2010. Detailed analysis of reactive oxygen species induced by visible light in various cell types. *Lasers Surg. Med.* 42 (6), 473–480.
- Leal Junior, E.C., Lopes-Martins, R.A., Dalan, F., *et al.*, 2008. Effect of 655-nm low-level laser therapy on exercise-induced skeletal muscle fatigue in humans. *Photomed. Laser Surg.* 26 (5), 419–424.
- Leal Junior, E.C., Lopes-Martins, R.A., de Almeida, P., *et al.*, 2010a. Effect of low-level laser therapy (GaAs 904 nm) in skeletal muscle fatigue and biochemical markers of muscle damage in rats. *Eur. J. Appl. Physiol.* 108 (6), 1083–1088.
- Leal Junior, E.C., Lopes-Martins, R.A., Frigo, L., *et al.*, 2010b. Effects of low-level laser therapy (LLLT) in the development of exercise-induced skeletal muscle fatigue and changes in biochemical markers related to postexercise recovery. *J. Orthop. Sports Phys. Ther.* 40 (8), 524–532.
- Leavitt, M., Charles, G., Heyman, E., Michaels, D., 2009. HairMax Laser comb laser phototherapy device in the treatment of male androgenetic alopecia: A randomized, double-blind, sham device-controlled, multicentre trial. *Clin. Drug Investig.* 29 (5), 283–292.
- Lee, J.H., Chang, S.Y., Moy, W.J., *et al.*, 2016. Simultaneous bilateral laser therapy accelerates recovery after noise-induced hearing loss in a rat model. *Peer J.* 4, e2252.
- Lee, S.Y., Park, K.H., Choi, J.W., *et al.*, 2007. A prospective, randomized, placebo-controlled, double-blinded, and split-face clinical study on LED phototherapy for skin rejuvenation: Clinical, profilometric, histologic, ultrastructural, and biochemical evaluations and comparison of three different treatment settings. *J. Photochem. Photobiol. B* 88 (1), 51–67.
- Little, F.M., Gomer, C.J., Hyman, S., Apuzzo, M.L., 1984. Observations in studies of quantitative kinetics of tritium labelled hematoporphyrin derivatives (HpDI and HpDII) in the normal and neoplastic rat brain model. *J. Neurooncol.* 2 (4), 361–370.
- Liu, X.G., Zhou, Y.J., Liu, T.C., Yuan, J.Q., 2009. Effects of low-level laser irradiation on rat skeletal muscle injury after eccentric exercise. *Photomed. Laser Surg.* 27 (6), 863–869.
- Loevschall, H., Arenholt-Bindslev, D., 1994. Effect of low level diode laser irradiation of human oral mucosa fibroblasts in vitro. *Lasers Surg. Med.* 14, 347–354.
- Lopes-Martins, R.A., Marcos, R.L., Leonardo, P.S., *et al.*, 2006. Effect of low-level laser (Ga-Al-As 655 nm) on skeletal muscle fatigue induced by electrical stimulation in rats. *J. Appl. Physiol.* (1985) 101 (1), 283–288.
- Lubart, R., Friedmann, H., 2011. LLLT and PDT. *Laser Ther.* 20 (3), 233.
- Lundeberg, T., Malm, M., 1991. Low-power HeNe laser treatment of venous leg ulcers. *Ann. Plast. Surg.* 27, 537–539.
- Madar-Balakirski, N., Tempel-Brami, C., Kalchenko, V., *et al.*, 2010. Permanent occlusion of feeding arteries and draining veins in solid mouse tumors by vascular targeted photodynamic therapy (VTP) with Tookad. *PLOS ONE* 5 (4), e10282.
- Maeda, H., Tsukigawa, K., Fang, J., 2016. A retrospective 30 years after discovery of the enhanced permeability and retention effect of solid tumors: Next-generation chemotherapeutics and photodynamic therapy – problems, solutions, and prospects. *Microcirculation* 23 (3), 173–182.
- Maeurer, M.J., Gollin, S.M., Storkus, W.J., *et al.*, 1996. Tumor escape from immune recognition: Loss of HLA-A2 melanoma cell surface expression is associated with a complex rearrangement of the short arm of chromosome 6. *Clin. Cancer Res.* 2 (4), 641–652.
- Magna, M., Pisetsky, D.S., 2016. The alarmin properties of DNA and DNA-associated nuclear proteins. *Clin. Ther.* 38 (5), 1029–1041.
- Maiman, T.H., 1960. Stimulated optical radiation in ruby. *Nature* 187 (4736), 493–494.
- Maisch, T., 2007. Anti-microbial photodynamic therapy: useful in the future? *Lasers Med. Sci.* 22 (2), 83–91.
- Maisch, T., 2015. Resistance in antimicrobial photodynamic inactivation of bacteria. *Photochem. Photobiol. Sci.* 14 (8), 1518–1526.
- Maisch, T., Hackbarth, S., Regensburger, J., *et al.*, 2011. Photodynamic inactivation of multi-resistant bacteria (PIB) – a new approach to treat superficial infections in the 21st century. *J. Deutschen Dermatologischen Gesellschaft = J. German Soc. Dermatol.: JDDG* 9 (5), 360–366.
- Maksimovich, I.V., 2015. Dementia and cognitive impairment reduction after laser transcatheter treatment of Alzheimer's disease. *World J. Neurosci.* 5, 189–203.
- Malik, Z., Ladan, H., Nitzan, Y., 1992. Photodynamic inactivation of Gram-negative bacteria: problems and possible solutions. *J. Photochem. Photobiol. B* 14 (3), 262–266.
- Maloney, R., Shanks, S., Maloney, J., 2010. The application of low-level laser therapy For the symptomatic care of late stage Parkinson's disease: A non-controlled, non-randomized study (abstract). *Lasers Surg. Med.* 185.
- Mathew, R., Karantz-Wadsworth, V., White, E., 2007. Role of autophagy in cancer. *Nat. Rev. Cancer.* 7 (12), 961–967.
- Maugain, E., Sasnouski, S., Zorin, V., *et al.*, 2004. Foscan-based photodynamic treatment in vivo: correlation between efficacy and Foscan accumulation in tumor, plasma and leukocytes. *Oncol. Rep* 12 (3), 639–645.
- Maziere, J.C., Santus, R., Morliere, P., *et al.*, 1990. Cellular uptake and photosensitizing properties of anticancer porphyrins in cell membranes and low and high density lipoproteins. *J. Photochem. Photobiol. B* 6 (1–2), 61–68.
- McClure, J., 2011. The role of causal attributions in public misconceptions about brain injury. *Rehab. Psychol.* 56 (2), 85–93.
- McDaniel, D.H., Newman, J., Geronemus, R., *et al.*, 2003. Non-ablative non-thermal LED photomodulation, $\sqrt{N}/\sqrt{Ea}$ multicenter clinical photoaging trial. *Lasers Surg. Med.* 15, 22.
- McGuff, P.E., Deterling Jr., R.A., Gottlieb, L.S., 1965. Tumoricidal effect of laser energy on experimental and human malignant tumors. *New England J. Med.* 273 (9), 490–492.
- Medrado, A.R., Pugliese, L.S., Reis, S.R., Andrade, Z.A., 2003. Influence of low level laser therapy on wound healing and its biological action upon myofibroblasts. *Lasers Surg. Med.* 32, 239–244.
- Meneguzzo, D.T., Lopes, L.A., Pallota, R., *et al.*, 2012. Prevention and treatment of mice paw edema by near-infrared low-level laser therapy on lymph nodes. *Lasers Med. Sci.*
- Merchat, M., Spikes, J.D., Bertoloni, G., Jori, G., 1996. Studies on the mechanism of bacteria photosensitization by meso-substituted cationic porphyrins. *J. Photochem. Photobiol. B* 35 (3), 149–157.
- Mester, E., Ludany, G., Sellyei, M., *et al.*, 1968a. Studies on the inhibiting and activating effects of laser beams. *Langenbecks Arch. Chir.* 322, 1022–1027.
- Mester, E., Ludany, G., Sellyei, M., Szende, B., Tota, J., 1968b. The stimulating effect of low power laser rays on biological systems. *Laser Rev.* 1, 3.
- Mester, E., Nagylucskay, S., Doklen, A., Tisza, S., 1976. Laser stimulation of wound healing. *Acta Chir. Acad. Sci. Hung.* 17, 49–55.
- Mester, E., Spiry, T., Szende, B., Tota, J.G., 1971. Effect of laser rays on wound healing. *Am. J. Surg.* 122, 532–535.
- Mester, E., Szende, B., Gartner, P., 1968c. The effect of laser beams on the growth of hair in mice. *Radiobiol. Radiother. (Berl)* 9 (5), 621–626.
- Mester, E., Szende, B., Spiry, T., Scher, A., 1972. Stimulation of wound healing by laser rays. *Acta Chir. Acad. Sci. Hung.* 13 (3), 315–324.

- Meyers, A.D., 1990. Lasers and wound healing. *Arch. Otolaryngol. Head Neck Surg.* 116, 1128.
- Michalikova, S., Ennaceur, A., van Rensburg, R., Chazot, P.L., 2008. Emotional responses and memory performance of middle-aged CD1 mice in a 3D maze: Effects of low infrared light. *Neurobiol Learn Mem* 89 (4), 480–488.
- Minnock, A., Vernon, D.I., Schofield, J., *et al.*, 1996. Photoinactivation of bacteria. Use of a cationic water-soluble zinc phthalocyanine to photoinactivate both gram-negative and gram-positive bacteria. *J. Photochem. Photobiol. B* 32 (3), 159–164.
- Miura, Y.Y.M., Tsuboi, R., Ogawa, H., 1999. Promotion of rat hair growth by irradiation using Super LizerTM. *Jpn. J. Dermatol.* 109 (13), 2149–2152.
- Moan, J., Berg, K., 1991. The photodegradation of porphyrins in cells can be used to estimate the lifetime of singlet oxygen. *Photochem. Photobiol.* 53 (4), 549–553.
- Moges, H., Vasconcelos, O.M., Campbell, W.W., *et al.*, 2009. Light therapy and supplementary Riboflavin in the SOD1 transgenic mouse model of familial amyotrophic lateral sclerosis (FALS). *Lasers Surg. Med.* 41, 52–59.
- Moreno-Arias, G.A., Castelo-Branco, C., Ferrando, J., 2002b. Side-effects after IPL photodepilation. *Dermatol. Surg.* 28 (12), 1131–1134.
- Moreno-Arias, G., Castelo-Branco, C., Ferrando, J., 2002a. Paradoxical effect after IPL photoepilation. *Dermatol. Surg.* 28 (11), 1013–1016. (discussion 1016).
- Moro, C., Massri, N.E., Darlot, F., *et al.*, 2016. Effects of a higher dose of near-infrared light on clinical signs and neuroprotection in a monkey model of Parkinson's disease. *Brain Res.*
- Moro, C., Massri, N.E., Torres, N., *et al.*, 2014. Photobiomodulation inside the brain: a novel method of applying near-infrared light intracranially and its impact on dopaminergic cell survival in MPTP-treated mice. *J. Neurosurg.* 120 (3), 670–683.
- Mroz, P., Szokalska, A., Wu, M.X., Hamblin, M.R., 2010. Photodynamic therapy of tumors can lead to development of systemic antigen-specific immune response. *PLoS One* 5 (12), e15194.
- Mroz, P., Vatansever, F., Muchowicz, A., Hamblin, M.R., 2013. Photodynamic therapy of murine mastocytoma induces specific immune responses against the cancer/testis antigen P1A. *Cancer Res.*
- Munoz-Price, L.S., Poirel, L., Bonomo, R.A., *et al.*, 2013. Clinical epidemiology of the global expansion of *Klebsiella pneumoniae* carbapenemases. *Lancet Infectious Dis.* 13 (9), 785–796.
- Naeser, M.A., Hamblin, M.R., 2015. Traumatic brain injury: a major medical problem that could be treated using transcranial, red/near-infrared LED photobiomodulation. *Photomed. Laser Surg.*
- Naeser, M.A., Martin, P.I., Lundgren, K., *et al.*, 2010. Improved language in a chronic nonfluent aphasia patient after treatment with CPAP and TMS. *Cogn. Behav. Neurol.* 23 (1), 29–38.
- Naeser, M.A., Saltmarche, A., Kregel, M.H., Hamblin, M.R., Knight, J.A., 2011. Improved cognitive function after transcranial, light-emitting diode treatments in chronic, traumatic brain injury: Two case reports. *Photomed. Laser Surg.* 29 (5), 351–358.
- Naeser, M.A., Zafonte, R., Kregel, M.H., *et al.*, 2014. Significant improvements in cognitive performance post-transcranial, red/near-infrared light-emitting diode treatments in chronic, mild traumatic brain injury: Open-protocol study. *J. Neurotrauma* 31 (11), 1008–1017.
- Naeser, M., Ho, M., Martin, P.E., *et al.*, 2012. Improved language after scalp application of red/near-infrared light-emitting diodes: pilot study supporting a new, noninvasive treatment for chronic aphasia. *Procedia – Social Behav. Sci.* 61, 138–139.
- Natoli, R., Valter, K., Barbosa, M., *et al.*, 2013. 670nm photobiomodulation as a novel protection against retinopathy of prematurity: evidence from oxygen induced retinopathy models. *PLOS ONE* 8 (8), e72135.
- Neira, R., *et al.* 2000. Low level assisted lipoplasty: A new technique. In: *Proceedings of the World Congress on Liposuction*, Dearborn, MI.
- Neira, R., Arroyave, J., Ramirez, H., *et al.*, 2002. Fat liquefaction: effect of low-level laser energy on adipose tissue. *Plast Reconstr Surg* 110 (3), 912–922. (discussion 923–5).
- Nitzan, Y., Balzani-Sudakevitz, A., Ashkenazi, H., 1998. Eradication of *Acinetobacter baumannii* by photosensitized agents in vitro. *J. Photochem. Photobiol. B* 42 (3), 211–218.
- Nitzan, Y., Guterman, M., Malik, Z., Ehrenberg, B., 1992. Inactivation of gram-negative bacteria by photosensitized porphyrins. *Photochem. Photobiol.* 55 (1), 89–96.
- Noble, P.B., Shields, E.D., Blecher, P.D., Bentley, K.C., 1992. Locomotory characteristics of fibroblasts within a three-dimensional collagen lattice: modulation by a Helium/Neon soft laser. *Lasers Surg. Med.* 12, 669–674.
- Okada, H., Mak, T.W., 2004. Pathways of apoptotic and non-apoptotic death in tumour cells. *Nat. Rev. Cancer* 4 (8), 592–603.
- Oleinick, N.L., Evans, H.H., 1998. The photobiology of photodynamic therapy: Cellular targets and mechanisms. *Radiat. Res.* 150 (5 Suppl), S146–S156.
- Oleinick, N.L., Morris, R.L., Belichenko, I., 2002. The role of apoptosis in response to photodynamic therapy: What, where, why, and how. *Photochem Photobiol Sci* 1 (1), 1–21.
- O'Neill, J., 2015. Tackling a global health crisis: Initial steps The Review on Antimicrobial Resistance Chaired by Jim O'Neill 2015.
- Oron, A., Oron, U., Chen, J., *et al.*, 2006. Low-level laser therapy applied transcranially to rats after induction of stroke significantly reduces long-term neurological deficits. *Stroke* 37 (10), 2620–2624.
- Oron, A., Oron, U., Streeter, J., *et al.*, 2007. low-level laser therapy applied transcranially to mice following traumatic brain injury significantly reduces long-term neurological deficits. *J. Neurotrauma* 24 (4), 651–656.
- Peoples, C., Spana, S., Ashkan, K., *et al.*, 2012. Photobiomodulation enhances nigral dopaminergic cell survival in a chronic MPTP mouse model of Parkinson's disease. *Parkinsonism Relat. Disord.* 18 (5), 469–476.
- Peplow, P.V., 2015. Neuroimmunomodulatory effects of transcranial laser therapy combined with intravenous tPA administration for acute cerebral ischemic injury. *Neural Regen. Res.* 10 (8), 1186–1190.
- Pereira, A.N., Eduardo Cde, P., Matson, E., Marques, M.M., 2002. Effect of low-power laser irradiation on cell growth and procollagen synthesis of cultured fibroblasts. *Lasers Surg. Med.* 31, 263–267.
- Perfettini, J.L., Kroemer, G., 2003. Caspase activation is not death. *Nat. Immunol.* 4 (4), 308–310.
- Pettigrew, C.A., Cotter, T.G., 2009. Deregulation of cell death (apoptosis): implications for tumor development. *Discov. Med.* 8 (41), 61–63.
- Pottier, R., Kennedy, J.C., 1990. The possible role of ionic species in selective biodistribution of photochemotherapeutic agents toward neoplastic tissue. *J. Photochem. Photobiol. B* 8 (1), 1–16.
- PR, K., 1989. Low level laser therapy: a review. *Laser Med. Sci.* 4 (3), 141–150.
- Proctor, P.H., 1989. Endothelium-derived relaxing factor and minoxidil: Active mechanisms in hair growth. *Arch. Dermatol.* 125 (8), 1146.
- Proskuryakov, S.Y., Gabai, V.L., 2010. Mechanisms of tumor cell necrosis. *Curr. Pharm. Des.* 16 (1), 56–68.
- Purushothuman, S., Nandasena, C., Johnstone, D.M., Stone, J., Mitrofanis, J., 2013. The impact of near-infrared light on dopaminergic cell survival in a transgenic mouse model of parkinsonism. *Brain Res.* 1535, 61–70.
- Qu, C., Cao, W., Fan, Y., Lin, Y., 2010. Near-infrared light protect the photoreceptor from light-induced damage in rats. *Adv. Exp. Med. Biol.* 664, 365–374.
- Quirk, B.J., Torbey, M., Buchmann, E., Verma, S., Whelan, H.T., 2012. Near-infrared photobiomodulation in an animal model of traumatic brain injury: improvements at the behavioral and biochemical levels. *Photomed. Laser Surg.* 30 (9), 523–529.
- Raskin, P., Marks, J.F., Burns, H., Plumer, M.D., Siperstein, M.D.L., 1975. Capillary basement membrane within diabetic children. *Am. J. Med.* 58, 365–375.
- Rathmell, J.C., Thompson, C.B., 1999. The central effectors of cell death in the immune system. *Annu. Rev. Immunol.* 17, 781–828.
- Reidy, G.K., Stehno-Bittel, L., Enwemeka, C.S., 2001. Laser photostimulation accelerates wound healing in diabetic rats. *Wound Repair Regen.* 9, 248–255.
- Reiners, R., J.J., Agostinis, P., Berg, K., Oleinick, N.L., Kessel, D., 2010. Assessing autophagy in the context of photodynamic therapy. *Autophagy* 6 (1), 7–18.
- Reis e Sousa, C., 2004. Activation of dendritic cells: Translating innate into adaptive immunity. *Curr. Opin. Immunol.* 16 (1), 21–25.
- Glen Calderhead, R., Kubota, J., Trelles, M.A., Ohshiro, T., 2008. One mechanism behind LED phototherapy for wound healing and skin rejuvenation: Key role of the mast cell. *Laser Ther.* 17, 141–148.

- Rhee, C.K., He, P., Jung, J.Y., *et al.*, 2013. Effect of low-level laser treatment on cochlea hair-cell recovery after ototoxic hearing loss. *J. Biomed. Opt.* 18 (12), 128003.
- Richter, A.M., Cerruti-Sola, S., Sternberg, E.D., Dolphin, D., Levy, J.G., 1990. Biodistribution of tritiated benzoporphyrin derivative (3H-BPD-MA), a new potent photosensitizer, in normal and tumor-bearing mice. *J. Photochem. Photobiol. B* 5 (2), 231–244.
- Rojas, J.C., Bruchey, A.K., Gonzalez-Lima, F., 2012. Low-level light therapy improves cortical metabolic capacity and memory retention. *J. Alzheimers Dis.* 32 (3), 741–752.
- Rossi, A., Cantisani, C., Melis, L., *et al.*, 2012. Minoxidil use in dermatology, side effects and recent patents. *Recent Pat. Inflamm. Allergy Drug Discov.* 6 (2), 130–136.
- Russell, B.A., Kellett, N., Reilly, L.R., 2005. A study to determine the efficacy of combination LED light therapy (633 nm and 830 nm) in facial skin rejuvenation. *J. Cosmet. Laser Ther.* 7 (3–4), 196–200.
- Rustin, P., 2002. Mitochondria, from cell death to proliferation. *Nat. Genet.* 30 (4), 352–353.
- Ryter, S.W., Gomer, C.J., 1993. Nuclear factor kappa B binding activity in mouse L1210 cells following photofrin II-mediated photosensitisation. *Photochem. Photobiol.* 58, 753–756.
- Salehpour, F., Rasta, S.H., Mohaddes, G., Sadigh-Eteghad, S., Salarirad, S., 2016. Therapeutic effects of 10-Hz pulsed wave lasers in rat depression model: A comparison between near-infrared and red wavelengths. *Lasers Surg. Med.*
- Salgado, A.S., Zangaro, R.A., Parreira, R.B., Kerppers, I.I., 2015. The effects of transcranial LED therapy (TCLT) on cerebral blood flow in the elderly women. *Lasers Med. Sci.* 30 (1), 339–346.
- Saltmarche, A.E., Naeser, M.A., Ho, K.F., Hamblin, M.R., Lim, L., 2016. Significant improvement in cognition after transcranial and intranasal photobiomodulation: A controlled, single-blind pilot study in participants with dementia (Abstract). In: Alzheimer's Association International Conference, Toronto, Canada.
- Sanchez-Chavez, G., Pena-Rangel, M.T., Riesgo-Escovar, J.R., Martinez-Martinez, A., Salceda, R., 2012. Insulin stimulated-glucose transporter Glut 4 is expressed in the retina. *PLOS ONE* 7 (12), e52959.
- Sandford, M.A., Walsh, L.J., 1994. Thermal effects during desensitisation of teeth with gallium-aluminium-arsenide lasers. *Periodontology* 15, 25–30.
- Santos, L.A., Marcos, R.L., Tomazoni, S.S., *et al.*, 2014. Effects of pre-irradiation of low-level laser therapy with different doses and wavelengths in skeletal muscle performance, fatigue, and skeletal muscle damage induced by tetanic contractions in rats. *Lasers Med. Sci.* 29 (5), 1617–1626.
- Sasnauskienė, A., Kadziauskas, J., Vezelyte, N., Jonusiene, V., Kirveliene, V., 2009. Apoptosis, autophagy and cell cycle arrest following photodamage to mitochondrial interior. *Apoptosis* 14 (3), 276–286.
- Satino, J.L.M.M., 2003. Hair regrowth and increased hair tensile strength using the hair max laser comb for low-level laser therapy. *Int. J. Cos. Surg. Aest. Dermatol.* 5, 113–117.
- Savill, J., Fadok, V., 2000. Corpse clearance defines the meaning of cell death. *Nature* 407, 784–788.
- Scherz-Shouval, R., Elazar, Z., 2007. ROS, mitochondria and the regulation of autophagy. *Trends Cell. Biol.* 17 (9), 422–427.
- Scherz-Shouval, R., Shvets, E., Fass, E., *et al.*, 2007. Reactive oxygen species are essential for autophagy and specifically regulate the activity of Atg4. *EMBO J.* 26 (7), 1749–1760.
- Schiffer, F., Johnston, A.L., Ravichandran, C., *et al.*, 2009. Psychological benefits 2 and 4 weeks after a single treatment with near infrared light to the forehead: A pilot study of 10 patients with major depression and anxiety. *Behav. Brain Funct.* 5, 46.
- Schindl, A., Heinze, G., Schindl, M., Pernerstorfer-Schon, H., Schindl, L., 2002. Systemic effects of low-intensity laser irradiation on skin microcirculation in patients with diabetic microangiopathy. *Microvasc. Res.* 64, 240–246.
- Schindl, A., Schindl, M., Pernerstorfer-Schon, H., 2000. Low intensity laser irradiation in the treatment of recalcitrant radiation ulcers in patients with breast cancer – long-term results of 3 cases. *Photodermatol. Photoimmunol. Photomed.* 16, 34–37.
- Schindl, A., Schindl, M., Schon, H., *et al.*, 1998. Low-intensity laser irradiation improves skin circulation in patients with diabetic microangiopathy. *Diabetes Care* 21, 580–584.
- Schuitmaker, J.J., Feitsma, R.I., Journee-De Korver, J.G., Dubbelman, T.M., Pauwels, E.K., 1993. Tissue distribution of bacteriochlorin a labelled with ^{99m}Tc-pertechnetate in hamster Greene melanoma. *Int. J. Radiat. Biol.* 64 (4), 451–458.
- Shaw, V.E., Spana, S., Ashkan, K., *et al.*, 2010. Neuroprotection of midbrain dopaminergic cells in MPTP-treated mice after near-infrared light treatment. *J. Comp. Neurol.* 518 (1), 25–40.
- Sheng, C., Pogue, B.W., Wang, E., Hutchins, J.E., Hoopes, P.J., 2004. Assessment of photosensitizer dosimetry and tissue damage assay for photodynamic therapy in advanced-stage tumors. *Photochem. Photobiol.* 79 (6), 520–525.
- Skinner, S.M., Gage, J.P., Wilce, P.A., Shaw, R.M., 1996. A preliminary study of the effects of laser radiation on collagen metabolism in cell culture. *Aust. Dent. J.* 41, 188–192.
- Smith, R., Coast, J., 2013. The true cost of antimicrobial resistance. *BMJ* 346, f1493.
- Spanheimer, R.G., Umpierrez, G.E., Stumpf, V., 1988. Decreased collagen production in diabetic rats. *Diabetes* 37, 371–376.
- Sroka, R., Fuchs, C., Schaffer, M., *et al.*, 1997. Biomodulation effects on cell mitosis after laser irradiation using different wavelengths. *Laser Surg. Med. Supplement* (9), 6.
- Star, W.M., Marijnissen, H.P., van den Berg-Blok, A.E., *et al.*, 1986. Destruction of rat mammary tumor and normal tissue microcirculation by hematoporphyrin derivative photoradiation observed in vivo in sandwich observation chambers. *Cancer Res.* 46 (5), 2532–2540.
- Stott, B., Korbek, M., 2007. Activation of complement C3, C5, and C9 genes in tumors treated by photodynamic therapy. *Cancer Immunol. Immunother.* 56 (5), 649–658.
- Sun, J., Cecic, I., Parkins, C.S., Korbek, M., 2002. Neutrophils as inflammatory and immune effectors in photodynamic therapy-treated mouse SCCVII tumours. *Photochem. Photobiol. Sci.* 1 (9), 690–695.
- Sur, B.W., Nguyen, P., Sun, C.H., Tromberg, B.J., Nelson, E.L., 2008. Immunophototherapy using PDT combined with rapid intratumoral dendritic cell injection. *Photochem. Photobiol.* 84 (5), 1257–1264.
- Sussai, D.A., Carvalho Pde, T., Dourado, D.M., *et al.*, 2010. Low-level laser therapy attenuates creatine kinase levels and apoptosis during forced swimming in rats. *Lasers Med. Sci.* 25 (1), 115–120.
- Tafur, J., Mills, P.J., 2008. Low-intensity light therapy: Exploring the role of redox mechanisms. *Photomed. Laser Surg.* 26 (4), 323–328.
- Tamura, A., Matsunobu, T., Mizutani, K., *et al.*, 2015. Low-level laser therapy for prevention of noise-induced hearing loss in rats. *Neurosci. Lett.* 595, 81–86.
- Tanaka, M., Mroz, P., Dai, T., *et al.*, 2012. Photodynamic therapy can induce a protective innate immune response against murine bacterial arthritis via neutrophil accumulation. *PLOS ONE* 7 (6), e39823.
- Tang, J., Du, Y., Lee, C.A., *et al.*, 2013. Low-intensity far-red light inhibits early lesions that contribute to diabetic retinopathy: In vivo and in vitro. *Invest. Ophthalmol. Vis. Sci.* 54 (5), 3681–3690.
- Tang, J., Herda, A.A., Kern, T.S., 2014. Photobiomodulation in the treatment of patients with non-center-involving diabetic macular oedema. *Br. J. Ophthalmol.* 98 (8), 1013–1015.
- Tauber, S., Schorn, K., Beyer, W., Baumgartner, R., 2003. Transmeatal cochlear laser (TCL) treatment of cochlear dysfunction: a feasibility study for chronic tinnitus. *Lasers Med. Sci.* 18 (3), 154–161.
- Teggi, R., Bellini, C., Fabiano, B., Bussi, M., 2008. Efficacy of low-level laser therapy in Meniere's disease: a pilot study of 10 patients. *Photomed. Laser Surg.* 26 (4), 349–353.
- Thomas, D.W., O'Neill, I.D., Harding, K.G., Shepherd, J.P., 1995. Cutaneous wound healing: A current perspective. *J. Oral Maxillofac. Surg.* 53, 442–447.
- Thong, P.S., Olivo, M., Kho, K.W., *et al.*, 2008. Immune response against angiosarcoma following lower fluence rate clinical photodynamic therapy. *J. Environ. Pathol. Toxicol. Oncol.* 27 (1), 35–42.
- Thong, P.S., Ong, K.W., Goh, N.S., *et al.*, 2007. Photodynamic-therapy-activated immune response against distant untreated tumours in recurrent angiosarcoma. *Lancet Oncol* 8 (10), 950–952.

- Trueb, R.M., 2009. Chemotherapy-induced alopecia. *Semin. Cutan Med. Surg.* 28 (1), 11–14.
- Tseng, M.T., Reed, M.W., Ackermann, D.M., *et al.*, 1988. Photodynamic therapy induced ultrastructural alterations in microvasculature of the rat cremaster muscle. *Photochem. Photobiol.* 48 (5), 675–681.
- Tuby, H., Maltz, L., Oron, U., 2006. Modulations of VEGF and iNOS in the rat heart by low level laser therapy are associated with cardioprotection and enhanced angiogenesis. *Lasers Surg. Med.* 38 (7), 682–688.
- Vabulas, R.M., Wagner, H., Schild, H., 2002. Heat shock proteins as ligands of toll-like receptors. *Curr. Top. Microbiol. Immunol.* 270, 169–184.
- Valduga, G., Bertoloni, G., Reddi, E., Jori, G., 1993. Effect of extracellularly generated singlet oxygen on gram-positive and gram-negative bacteria. *J. Photochem. Photobiol. B* 21 (1), 81–86.
- van Breugel, H.H., Bar, P.R., 1992. Power density and exposure time of He-Ne laser irradiation are more important than total energy dose in photo-biomodulation of human fibroblasts in vitro. *Lasers Surg. Med.* 12 (5), 528–537.
- van Duijnhoven, F.H., Aalbers, R.I., Rovers, J.P., Terpstra, O.T., Kuppen, P.J., 2003. Immunological aspects of photodynamic therapy of liver tumors in a rat model for colorectal cancer. *Photochem. Photobiol.* 78 (3), 235–240.
- Vera, D.M., Haynes, M.H., Ball, A.R., *et al.*, 2012. Strategies to potentiate antimicrobial photoinactivation by overcoming resistant phenotypes. *Photochem. Photobiol.* 88 (3), 499–511.
- Vieira, W.H., Ferraresi, C., Perez, S.E., Baldissera, V., Parizotto, N.A., 2012. Effects of low-level laser therapy (808 nm) on isokinetic muscle performance of young women submitted to endurance training: A randomized controlled clinical trial. *Lasers Med. Sci.* 27 (2), 497–504.
- Vinck, E., Cagnie, B., Coorevits, P., Vanderstraeten, G., Cambier, D., 2006. Pain reduction by infrared light-emitting diode irradiation: A pilot study on experimentally induced delayed-onset muscle soreness in humans. *Lasers Med. Sci.* 21 (1), 11–18.
- Vlachos, S.P., Kontoes, P.P., 2002. Development of terminal hair following skin lesion treatments with an intense pulsed light source. *Aesthetic Plast. Surg.* 26 (4), 303–307.
- Vohra, F., Al-Rifaiy, M.Q., Lillywhite, G., Abu Hassan, M.I., Javed, F., 2014. Efficacy of mechanical debridement with adjunct antimicrobial photodynamic therapy for the management of peri-implant diseases: a systematic review. *Photochem. Photobiol. Sci.* 13 (8), 1160–1168.
- Wachowska, M., Gabrysiak, M., Muchowicz, A., *et al.*, 2014. 5-Aza-2'-deoxycytidine potentiates antitumour immune response induced by photodynamic therapy. *Eur. J. Cancer.*
- Wahl, G., Bastanier, S., 1991. Soft laser in postoperative care in dentoalveolar treatment. *ZWR* 100, 512–515.
- Wainwright, M., 1998. Photodynamic antimicrobial chemotherapy (PACT). *J. Antimicrob. Chemother.* 42 (1), 13–28.
- Wajima, Z., Shitara, T., Inoue, T., Ogawa, R., 1996. Linear polarized light irradiation around the stellate ganglion area increases skin temperature and blood flow. *Masui* 45 (4), 433–438.
- Weishaupt, K.R., Gomer, C.J., Dougherty, T.J., 1976. Identification of singlet oxygen as the cytotoxic agent in photoinactivation of a murine tumor. *Cancer Res.* 36 (7 PT 1), 2326–2329.
- Weiss, R.A., McDaniel, D.H., Geronemus, R., *et al.*, 2004a. Non-ablative, non-thermal light emitting diode (LED) phototherapy of photoaged skin. *Laser Surg. Med.* 16, 31.
- Weiss, R.A., McDaniel, D.H., Geronemus, R.G., Weiss, M.A., 2005a. Clinical trial of a novel non-thermal LED array for reversal of photoaging: Clinical, histologic, and surface profilometric results. *Lasers Surg. Med.* 36 (2), 85–91.
- Weiss, R.A., McDaniel, D.H., Geronemus, R.G., *et al.*, 2005b. Clinical experience with light-emitting diode (LED) photomodulation. *Dermatol. Surg.* 31 (9 Pt 2), 1199–1205.
- Weiss, R.A., Weiss, M.A., Geronemus, R.G., McDaniel, D.H., 2004b. A novel non-thermal non-ablative full panel LED photomodulation device for reversal of photoaging: Digital microscopic and clinical results in various skin types. *J. Drugs Dermatol.* 3 (6), 605–610.
- West, C.M., West, D.C., Kumar, S., Moore, J.V., 1990. A comparison of the sensitivity to photodynamic treatment of endothelial and tumour cells in different proliferative states. *Int. J. Radiat. Biol.* 58 (1), 145–156.
- Wikramanayake, T.C., Rodriguez, R., Choudhary, S., *et al.*, 2012a. Effects of the Lexington LaserComb on hair regrowth in the C3H/HeJ mouse model of alopecia areata. *Lasers Med. Sci.* 27 (2), 431–436.
- Wikramanayake, T.C., Villasante, A.C., Mauro, L.M., *et al.*, 2012b. Low-level laser treatment accelerated hair regrowth in a rat model of chemotherapy-induced alopecia (CIA). *Lasers Med. Sci.* 28 (3), 701–706.
- Wilks, A.F., 1989. Two putative protein-tyrosine kinases identified by application of the polymerase chain reaction. *Proc. Natl Acad. Sci. USA* 86 (5), 1603–1607.
- Willner, P., 1997. Validity, reliability and utility of the chronic mild stress model of depression: A 10-year review and evaluation. *Psychopharmacology (Berl)* 134 (4), 319–329.
- Wilson, M., 2004. Lethal photosensitisation of oral bacteria and its potential application in the photodynamic therapy of oral infections. *Photochem. Photobiol. Sci.* 3 (5), 412–418.
- Woodburn, K.W., Stylli, S., Hill, J.S., *et al.*, 1992. Evaluation of tumour and tissue distribution of porphyrins for use in photodynamic therapy. *Br. J. Cancer* 65 (3), 321–328.
- Wu, Q., Xuan, W., Ando, T., *et al.*, 2012a. Low-level laser therapy for closed-head traumatic brain injury in mice: effect of different wavelengths. *Lasers Surg. Med.* 44 (3), 218–226.
- Wu, S., Zhou, F., Zhang, Z., Xing, D., 2011. Bax is essential for Drp1-mediated mitochondrial fission but not for mitochondrial outer membrane permeabilization caused by photodynamic therapy. *J. Cell Physiol.* 226 (2), 530–541.
- Wu, X., Alberico, S.L., Moges, H., *et al.*, 2012b. Pulsed light irradiation improves behavioral outcome in a rat model of chronic mild stress. *Lasers Surg. Med.* 44 (3), 227–232.
- Xuan, W., Agrawal, T., Huang, L., Gupta, G.K., Hamblin, M.R., 2015. Low-level laser therapy for traumatic brain injury in mice increases brain derived neurotrophic factor (BDNF) and synaptogenesis. *J. Biophotonics* 8 (6), 502–511.
- Xuan, W., Vatansever, F., Huang, L., Hamblin, M.R., 2014. Transcranial low-level laser therapy enhances learning, memory, and neuroprogenitor cells after traumatic brain injury in mice. *J. Biomed. Opt.* 19 (10), 108003.
- Xuan, W., Vatansever, F., Huang, L., *et al.*, 2013. Transcranial low-level laser therapy improves neurological performance in traumatic brain injury in mice: Effect of treatment repetition regimen. *PLOS ONE* 8 (1), e53454.
- Xue, L.Y., Chiu, S.M., Oleinick, N.L., 2001. Photodynamic therapy-induced death of MCF-7 human breast cancer cells: A role for caspase-3 in the late steps of apoptosis but not for the critical lethal event. *Exp. Cell Res.* 263 (1), 145–155.
- Yamazaki, M., Miura, Y., Tsuboi, R., Ogawa, H., 2003. Linear polarized infrared irradiation using Super Lizer is an effective treatment for multiple-type alopecia areata. *Int. J. Dermatol.* 42 (9), 738–740.
- Yano, K.B.L.F., Detmar, M., 2001. Control of hair growth and follicle size by VEGF-mediated angiogenesis. *J. Clin. Invest.* 107, 409–417.
- Yoneyama, H., Katsumata, R., 2006. Antibiotic resistance in bacteria and its future for novel antibiotic development. *Biosci. Biotechnol. Biochem.* 70 (5), 1060–1075.
- Youngblood, W.J., Gryko, D.T., Lammi, R.K., *et al.*, 2002. Glaser-mediated synthesis and photophysical characterization of diphenylbutadiyne-linked porphyrin dyads. *J. Organic Chem.* 67 (7), 2111–2117.
- Yuan, F., Leunig, M., Berk, D.A., Jain, R.K., 1993. Microvascular permeability of albumin, vascular surface area, and vascular volume measured in human adenocarcinoma LS174T using dorsal chamber in SCID mice. *Microvasc. Res.* 45 (3), 269–289.
- Yusuf, N., Katiyar, S.K., Elmets, C.A., 2008. The immunosuppressive effects of phthalocyanine photodynamic therapy in mice are mediated by CD4+ and CD8+ T cells and can be adoptively transferred to naive recipients. *Photochem. Photobiol.* 84 (2), 366–370.
- Yu, W., Naim, J.O., Lanzafame, J., 1997. Effects of photostimulation on wound healing in diabetic mice. *Lasers Surg. Med.* 20, 56–63.
- Yu, W., Naim, J.O., Lanzafame, R.J., 1994. The effect of laser irradiation on the release of bFGF from 3T3 fibroblasts. *Photochem. Photobiol.* 59 (2), 167–170.
- Zhang, L., Xing, D., Zhu, D., Chen, Q., 2008. Low-power laser irradiation inhibiting Abeta25-35-induced PC12 cell apoptosis via PKC activation. *Cell Physiol. Biochem.* 22, 215–222.

- Zhang, Q., Zhou, C., Hamblin, M.R., Wu, M.X., 2014. Low-level laser therapy effectively prevents secondary brain injury induced by immediate early responsive gene X-1 deficiency. *J. Cereb. Blood Flow Metabolism: Off. J. Int. Soc. Cereb. Blood Flow Metabolism* 34 (8), 1391–1401.
- Zhao, Y., Vanhoutte, P.M., Leung, S.W., 2015. Vascular nitric oxide: Beyond eNOS. *J. Pharmacol. Sci.* 129 (2), 83–94.
- Zivin, J.A., Albers, G.W., Bornstein, N., *et al.*, 2009. Effectiveness and safety of transcranial laser therapy for acute ischemic stroke. *Stroke* 40 (4), 1359–1364.
- Zivin, J.A., Sehra, R., Shoshoo, A., *et al.*, 2014. Neuro thera(R) efficacy and safety trial-3 (NEST-3): a double-blind, randomized, sham-controlled, parallel group, multicenter, pivotal study to assess the safety and efficacy of transcranial laser therapy with the neuro thera(R) laser system for the treatment of acute ischemic stroke within 24h of stroke onset. *Int. J. Stroke* 9 (7), 950–955.

Relevant Website

<https://www.sciencebasedmedicine.org/science-by-press-release-a-helmet-to-fight-alzheimers-disease/>
Science-Based Medicine.

Resolution and Multiple Scattering in Imaging

Edwin A Marengo, Northeastern University, Boston, MA, United States and Technological University of Panama, Panama City, Panama

Zambrano-Nunez Maytee, Technological University of Panama, Panama City, Panama

Edson S Galagarza, Northeastern University, Boston, MA, United States and Technological University of Panama, Panama City, Panama

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Light carries information about the sources from which it comes and the objects with which it interacts. Imaging can be defined as the physical or computational processing of this information in order to create a real or virtual replica (an “image”) of the source or object or its associated optical wavefield. This function can be performed by a physical imaging system, such as a classical lens-based system, as shown in Fig. 1. Alternatively, it can be implemented computationally, by a two-layer system composed of a physical interface, which senses the field produced by the object and a computational block that implements inversion algorithms to estimate the radiating object’s properties, which can, in turn, render image representations of that object and its surrounding field (see Fig. 2). The most critical property of an imaging system, be it physical or computational, is the degree of accuracy with which objects under surveillance can be successfully imaged or reproduced by that system. This characterization of the imaging system, usually termed “imaging resolution,” is affected by both fundamental physical limits (the inherent finite resolving power of the system, also termed the “diffraction limit”) as well as environmental effects (propagation model deviations) and sensor imperfections (aberrations). This contribution outlines the key principles to define and compute the resolution of typical imaging systems including those in which the phenomenon of multiple scattering can play a major role. Particular attention is given to the

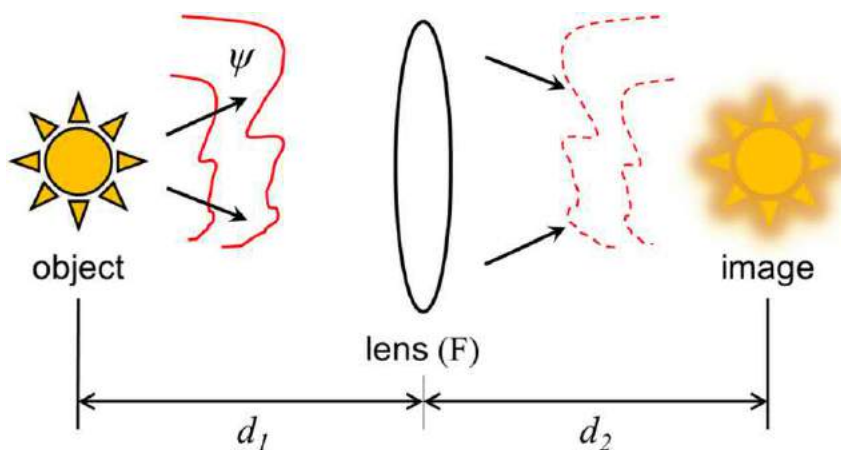


Fig. 1 Physical imaging. In this lens-based system, imaging occurs if $1/d_1 + 1/d_2 = 1$ where F is the lens focal length.

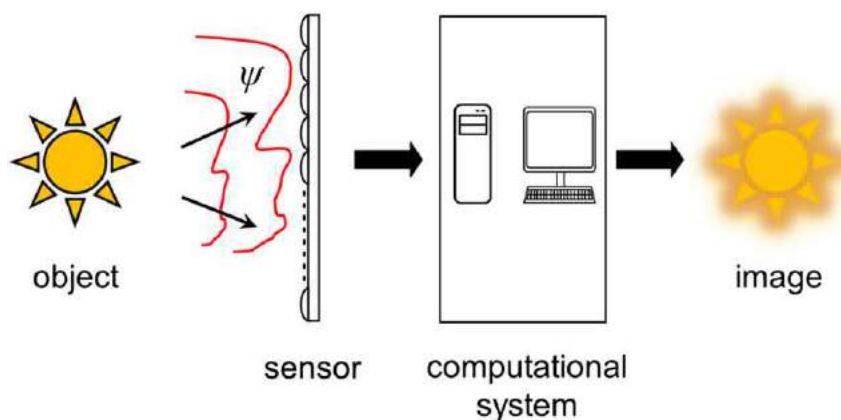


Fig. 2 Computational Imaging.

intrinsic, fundamental resolution limit of the systems considered, which can be thought of as a fundamental bound to the resolution. The latter is indeed attainable in state-of-the-art systems (termed “diffraction-limited”).

Inverse Source and Scattering Problems

Two different forms of imaging are of interest in this contribution. The most basic one, to be termed here “source-field imaging,” involves the imaging of sources and their fields, and can be addressed within the theoretical framework of the inverse source problem (reconstruction of an unknown source from knowledge of its field in a region located outside the source support, such as in the far zone). This includes as a special case the inverse diffraction problem of reconstructing the field in a boundary or region from knowledge of the field in another region. The sources in question can be self-emitting (primary) or induced (secondary), as in a scatterer that is illuminated by light from another source (see Fig. 3). The second kind of imaging problem under consideration, to be termed “scatterer imaging,” involves the determination of a scatterer’s constitutive properties (e.g., spatial distribution of its index of refraction) from scattering data corresponding to a number of transmit-and-receive experiments.

In source-field imaging the associated mapping from the sought-after source or field (the object) to the field data is linear. In the classical theoretical formulations of this problem one can talk about a well-defined resolution limit arising from bandlimitation considerations in Fourier space. In contrast, in the scatterer imaging or inverse scattering problem, the associated forward map from scatterer properties to the data is, in general, nonlinear, so that alternative methods, such as the statistical Cramer–Rao bound (CRB) (Kay, 1993, chapter 3) need to be adopted so as to quantify the resolution (Sentenac et al., 2007). Scattering has, however, a weak scatterer regime, called “the Born approximation,” in which the map in question is linear. This specialized regime is of interest in many practical applications and within it there is an associated classical resolution limit which has characteristics similar to those of the related source-field imaging problem.

Background and Literature

Resolution is traditionally quantified as the minimum distance that two point sources or scatterers can have while being distinguishable in the corresponding image. The image of a point source or scatterer is usually called the point spread function (PSF) of the imaging system. According to the well-known Rayleigh criterion (Goodman, 2005, p. 158), the resolution corresponds to the separation for which the peak of the image of one source occurs at the first null or zero of the image of the other. This is, however, only an approximate measure of resolution which applies, in fact, to a quite specific type of object. As pointed out by Toraldo di Francia in his classical studies on resolving power and information (Toraldo di Francia, 1952, 1955), imaging resolution is more appropriately characterized by the entire PSF and its associated information content, for example, in the spectral domain, from which one can in turn determine the effective dimensionality or number of degrees of freedom (NDF) of the entire imaging system. In many of the imaging problems that arise in optics, this detailed resolution characterization can be done quite conveniently in spatial Fourier domain, as we explain next for both inverse source and inverse scattering scenarios.

Among the major contributions to our understanding of imaging resolution and the associated field data “information content” and degrees of freedom of classical imaging systems, it is worthwhile mentioning the pioneering work of Abbe who established the resolution limit of the microscope around 1873 and Lord Rayleigh who further expanded the relevant theory in 1879 (Lord Rayleigh, 1879). Other contributions include those of Sparrow (1916), Schuster (1924), Houston (1927), and Buxton (1937), who proposed alternatives to the Rayleigh resolution criterion. Important advances at the crossroads of antenna theory, information theory, and optics were done in the 1940s and 1950s by several authors including Schelkunoff, Bouwkamp and Bruijn (1946), Woodward and Lawson (1948), and Toraldo di Francia (1955, 1952), which addressed the theoretical and practical

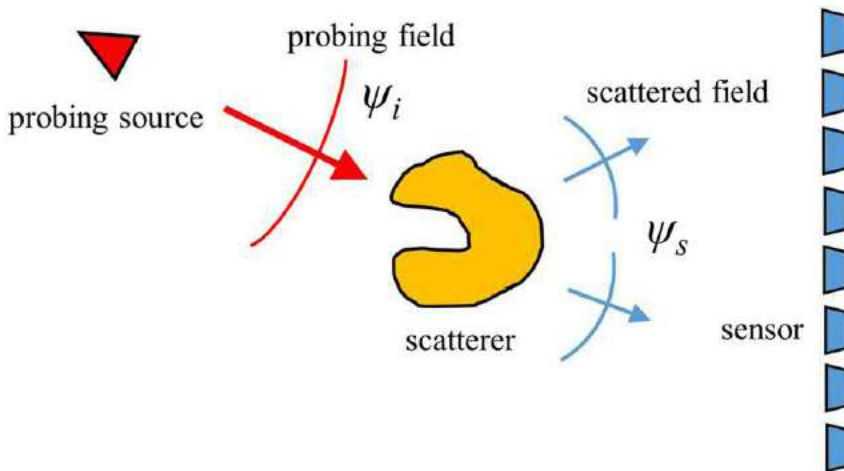


Fig. 3 Scattering experiment. The scatterer generates scattered field $\psi_s(\mathbf{r})$ in response to excitation or probing with incident field $\psi_i(\mathbf{r})$.

limits associated to the launching of and imaging with far fields. Among other results, these and other subsequent investigations (Ronchi, 1961; Rushforth and Harris, 1968) conclusively rendered the important insight that the resolution limits of far-field data are, for the most part, only practical (as opposed to theoretical), insofar as they are the result of noise and perturbations in realistic data and signal models. For example, if data were perfect and thus lacked noise, then it would be possible for many imaging systems (see Sections Propagating and Evanescent Spectra of the Radiated Field and Transverse Resolution of this article), to analytically extend the data corresponding to a region in Fourier space (called the visible region) into another one (called invisible) so as to increase the signal bandwidth and improve the resolution (Goodman, 2005; Section 6.6). This is one of the computational approaches to “superresolution” but it is useful only for perfect field data. In realistic systems, data are not perfect, and thus this approach is of limited practical applicability. But this illustrates that the conventional “fundamental” resolution limits are mostly practical, due to the almost inevitable presence of noise and perturbations, such as systematic errors. For instance, it is not impossible to physically process the far field, via a suitable device or instrument, so as to achieve superresolution. But (as in the computational case) such an analog or “physical processing” approach is viable only insofar as the physical field arriving at the instrument carries itself little noise or interference from spurious sources. Several promising approaches along these lines are mentioned in Section Superresolution, which are based on metamaterials. There we also mention other, more robust and already proven approaches to superresolution. Recent overviews on resolution and the related topic of superresolution can be found in den Dekker and van den Bos (1997), Heintzmann and Ficz (2006), Zheludev (2008), Testorf and Fiddy (2010).

Resolution in Source-Field Imaging

Scalar Wave Model

In the scalar approximation, the light emitted in free space by a monochromatic source oscillating at temporal frequency f (in Hertz) can be represented by a scalar wavefield whose spatial dependence is defined by a wavefunction $\psi(\mathbf{r})$ which obeys the inhomogeneous Helmholtz equation

$$(\nabla^2 + k_0^2)\psi(\mathbf{r}) = \rho(\mathbf{r}) \quad (1)$$

where $\mathbf{r} \in \mathbb{R}^3$, ρ represents the spatial dependence of the source (at said temporal frequency), ∇^2 is the Laplacian operator, and the wavenumber

$$k_0 = \omega/c = 2\pi/\lambda \quad (2)$$

where $\omega = 2\pi f$ is the angular oscillation frequency (in radians per second), c is the speed of light in free space, and λ is the wavelength. In this model of wave propagation, the wavefield $\psi(\mathbf{r})$ radiated by the source $\rho(\mathbf{r})$ is given by

$$\psi(\mathbf{r}) = \int_{\tau} d^3r' \rho(\mathbf{r}') G_0(\mathbf{r} - \mathbf{r}') \quad (3)$$

where d^3r' denotes differential volume, τ is the spatial support of the source, and G_0 is the free space Green function defined by

$$G_0(\mathbf{r} - \mathbf{r}') = -\frac{e^{ik_0|\mathbf{R}|}}{4\pi|\mathbf{R}|} \quad (4)$$

where $i = \sqrt{-1}$ and

$$\mathbf{R} = \mathbf{r} - \mathbf{r}' \quad (5)$$

Green's function G_0 admits a number of alternative representations, such as the Fourier-integral representation,

$$G_0(\mathbf{R}) = -\frac{1}{8\pi^3} \int_{-\infty}^{\infty} d^3K \frac{\exp(i\mathbf{K} \cdot \mathbf{R})}{k_0^2 - K^2 + i\varepsilon} \quad (6)$$

where $\mathbf{K} \in \mathbb{R}^3$, d^3K is differential volume in Fourier space, $K = |\mathbf{K}|$, and ε is an infinitesimal value added to remove the singularity at $K = k_0$.

Three-Dimensional Resolution of Far Fields

In many applications the fields are sensed in the far zone of the source region (τ), where $|\mathbf{r} - \mathbf{r}'| \simeq r - \hat{\mathbf{r}} \cdot \mathbf{r}'$, where $r = |\mathbf{r}|$ and $\hat{\mathbf{r}} = \mathbf{r}/r$, which when substituted in Eq. (4) gives the far zone behavior of G_0 ,

$$G_0(r\hat{\mathbf{r}} - \mathbf{r}') = -\frac{e^{ik_0r}}{4\pi r} e^{-ik_0\hat{\mathbf{r}} \cdot \mathbf{r}'} \quad (7)$$

Substituting this in Eq. (3) gives the far zone behavior of the field ψ ,

$$\psi(r\hat{\mathbf{r}}) \sim f(\hat{\mathbf{r}}) \frac{e^{ik_0r}}{r} \quad (8)$$

where the quantity $f(\hat{\mathbf{r}})$ is the far-field radiation pattern defined by

$$f(\hat{\mathbf{r}}) = -\frac{1}{4\pi} \int_{\tau} d^3r \rho(\mathbf{r}) e^{-ik_0 \hat{\mathbf{r}} \cdot \mathbf{r}} \quad (9)$$

The latter is, apart from a multiplicative factor, equal to the spatial Fourier transform of the source

$$\tilde{\rho}(\mathbf{K}) = \int_{\tau} d^3r \rho(\mathbf{r}) e^{-i\mathbf{K} \cdot \mathbf{r}} \quad (10)$$

evaluated at $\mathbf{K} = k_0 \hat{\mathbf{r}}$, i.e.,

$$f(\hat{\mathbf{r}}) = -\frac{1}{4\pi} \tilde{\rho}(\mathbf{K})|_{\mathbf{K}=k_0 \hat{\mathbf{r}}} \quad (11)$$

This means that the information about the source that is contained in the far field is the spatial Fourier transform of the source evaluated over the surface of a sphere, called the Ewald sphere (shown in Fig. 4), that has radius $K = k_0$ and is centered at the origin in Fourier space. The $K = k_0$ and $K \neq k_0$ components are sometimes referred to as the “on-the-shell” and “off-the-shell” components of the Fourier spectrum.

The fact that the far field carries only on-the-shell information about the spatial Fourier transform of ρ means that there is a loss of information about the source in the wave propagation process (from source to far field). This, in turn, implies that there is a fundamental limit to our capacity to reconstruct an unknown source from knowledge of the far field alone. It is, in fact, impossible to uniquely deduce the unknown source from the far field alone, since there are infinite nontrivial sources (called nonradiating sources) for which the spatial Fourier transform vanishes identically over the Ewald sphere (Devaney and Wolf, 1973; Bleistein and Cohen, 1977; Marengo and Ziolkowski, 2000; Gbur, 2003). One can, however, evaluate the unique, wavelike source which best fits the Fourier data over that sphere and has zero off-the-shell Fourier components. This is the classical approach to source imaging, and is the one that will lead us next to the fundamental resolution limit. Mathematically, this corresponds to the solution of minimum functional energy or 2 norm, or “minimum energy solution” (Devaney, 2012, chapter 5) to the inverse problem, without prior information about the source support. Minimum energy solutions consistent with a known source support can also be considered in certain applications (Marengo et al., 1999, 2000). In this case, the prior support information reliably enables superresolution (Schmidt-Weinmar, 1978).

The inverse problem of reconstructing an unknown source ρ , whose support τ is unknown, and that generates a given far-field radiation pattern $f(\hat{\mathbf{r}})$, with full-view data corresponding to 4π steradians, can thus be cast as computing the inverse Fourier transform (IFT) of the “on-the-shell” Fourier transform of the source. This is equivalent to Fourier inverting the quantity $\tilde{\rho}(\mathbf{K})H(\mathbf{K})$, where $H(\mathbf{K}) = \delta(K - k_0)/K^2$, which implies in view of the convolution theorem that the reconstructed source $\hat{\rho}$ is related to the source ρ by

$$\hat{\rho}(\mathbf{r}) = \int_{\tau} d^3r' \rho(\mathbf{r}') h(\mathbf{r} - \mathbf{r}') \quad (12)$$

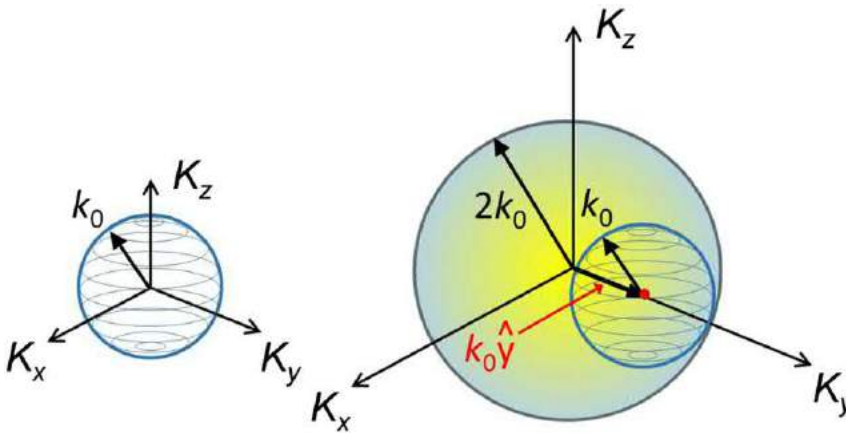


Fig. 4 Ewald sphere (left), and limiting Ewald sphere and ball (right) in 3D Fourier space. The Ewald sphere ($|\mathbf{K}| = k_0$) defines the Fourier domain region populated by far-field data of inverse source problems. The limiting Ewald ball ($|\mathbf{K}| \leq 2k_0$) defines the corresponding data region for full-view far-field inverse scattering within the weak scattering or first Born approximation. The figure also shows a shifted Ewald sphere ($|\mathbf{K} - k_0 \hat{\mathbf{y}}| = k_0$). It is contained inside the limiting Ewald ball, and corresponds to far-zone scattered field data under probing with a plane wave traveling in the direction of the unit vector $-\hat{\mathbf{y}}$. By probing the scatterer with plane waves traveling in all directions, the resulting scattered field data populates the entire limiting Ewald ball.

where the PSF h of this imaging system, equal to the IFT of H , can be shown to be given by

$$h(\mathbf{r} - \mathbf{r}') = \frac{1}{2\pi^2} j_0(k_0 |\mathbf{r} - \mathbf{r}'|) = -\frac{2}{\pi k_0} \text{Im} G_0(\mathbf{r} - \mathbf{r}') \quad (13)$$

where j_0 is the spherical Bessel function of the first kind and order 0 and Im denotes the imaginary part. A plot of this PSF is shown in Fig. 5. The first zero of $j_0(k_0 \Delta_{3D}^R)$ corresponding to $k_0 \Delta_{3D}^R = \pi$, gives a separation

$$\Delta_{3D}^R = \lambda/2 \quad (14)$$

which can be regarded, in view of the Rayleigh criterion, as the sought-after resolution that is attainable with this 3D, source-inversion-based computational imaging system. In particular, without the involvement of prior information about the source (such as support, parametric form, point-like nature, sparsity, or other), then two point sources are well resolved with far-field data only if they are separated by a distance equal to or greater than about half a wavelength.

Propagating and Evanescent Spectra of the Radiated Field

In many applications, field sensing is done over a plane, say $z = z_1$, which is itself contained in a half-space $z > z_0$ that is exterior to the source support τ , as shown in Fig. 6. In these applications, the field can be conveniently analyzed using a quite insightful classical representation called the angular-spectrum expansion, whose main features are outlined next.

Consider the Cartesian representation of $\mathbf{r}$ and $\mathbf{r}'$, in particular, $\mathbf{r} = (x, y, z)$ and $\mathbf{r}' = (x', y', z')$. Similarly, let $\mathbf{K} = (K_x, K_y, K_z)$. Let $\mathbf{R}_t = (x - x', y - y')$ and $\mathbf{K}_t = (K_x, K_y)$ denote the transverse components of $\mathbf{R}$ and $\mathbf{K}$, respectively, and introduce $Z = z - z'$. Then

$$\begin{aligned} \mathbf{R} &= \mathbf{R}_t + Z\hat{z} \\ \mathbf{K} &= \mathbf{K}_t + K_z\hat{z} \end{aligned} \quad (15)$$

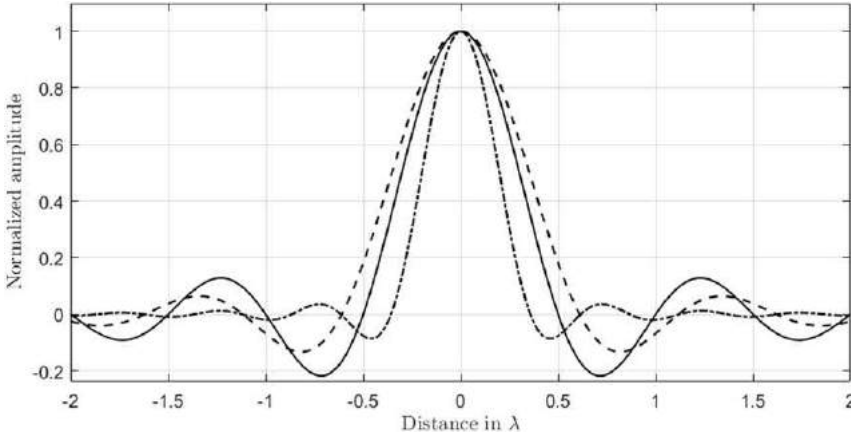


Fig. 5 Plots of the normalized point spread function (PSF) $h(d)/h(0)$ vs. d (in λ) relevant to three-dimensional (3D) source-field imaging (solid line), two-dimensional (2D) transverse field imaging (dashed line), and 3D Born-approximable inverse scattering (dashed-dotted line).

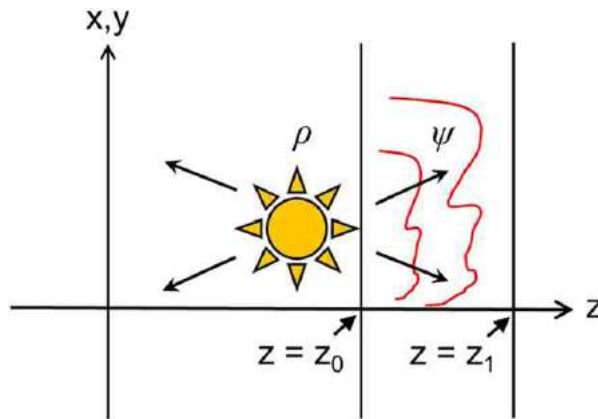


Fig. 6 Geometry used in the discussion of the plane wave expansion of the radiated field.

Integrating over K_z in Eq. (6) one arrives, via classical contour integration in the complex plane (see Devaney, 2012, Section 4.1), to the Weyl expansion of Green's function G_0 , in particular,

$$G_0(\mathbf{R}) = -\frac{i}{8\pi^2} \int_{-\infty}^{\infty} d^2 K_t \frac{e^{i(\mathbf{K}_t \cdot \mathbf{R}_t \pm \gamma Z)}}{\gamma(\mathbf{K}_t)} \quad (16)$$

where "+" is used if $Z > 0$ and "-" if $Z < 0$ and

$$\gamma(\mathbf{K}_t) = \begin{cases} \sqrt{k_0^2 - K_t^2} & \text{if } K_t \leq k_0 \\ i\sqrt{K_t^2 - k_0^2} & \text{if } K_t > k_0 \end{cases} \quad (17)$$

where $K_t = |\mathbf{K}_t|$.

In view of Eq. (17), the integral in Eq. (16) can be conveniently decomposed into two parts, one corresponding to integration over the ball in two-dimensional $\mathbf{K}_t$ space defined by $K_t \leq k_0$, and another corresponding to integration over the complementary region (the exterior of this ball). These two regions are illustrated in Fig. 7. The waves associated to the $K_t \leq k_0$ region, sometimes called "the visible region," are propagating, and go to the far zone, while the waves corresponding to the $K_t > k_0$ region, sometimes referred to as "the invisible region," decay exponentially with $|Z|$ and are called evanescent. The latter carry the high-frequency features of G_0 but cannot be sensed effectively for $|Z|$ greater than a few wavelengths. Moreover, since this field representation can be applied to any $\hat{\mathbf{z}} \in S^2$ over the unit sphere S^2 , this further implies that these high-frequency, evanescent features of G_0 cannot be sensed a few wavelengths away from the source point $\mathbf{r}'$. For $|Z|$ greater than a few wavelengths, $G_0(\mathbf{R})$ can thus be approximated as

$$G_0(\mathbf{R}) \simeq -\frac{i}{8\pi^2} \int_{K_t \leq k_0} d^2 K_t \frac{e^{i(\mathbf{K}_t \cdot \mathbf{R}_t \pm \gamma Z)}}{\gamma(\mathbf{K}_t)} \quad (18)$$

This is a free field, obeying the homogeneous Helmholtz equation in all space.

Physically, the Weyl expansion (Eq. (16)) thus characterizes Green's function G_0 as superpositions (an spectrum) of two kinds of plane waves, namely, propagating and evanescent waves. The expansion coefficient $1/\gamma(\mathbf{K}_t)$ appearing in Eq. (16) can be thought of as the plane wave spectrum of G_0 . Its value in the visible and invisible regions defines the respective propagating and evanescent spectra of G_0 . Interestingly, it can be shown that (Devaney, 2012, p. 122)

$$\text{Im} G_0(\mathbf{R}) = -\frac{i}{8\pi^2} \int_{K_t \leq k_0} \frac{d^2 K_t}{\gamma(\mathbf{K}_t)} e^{i\mathbf{K}_t \cdot \mathbf{R}_t} \cos[\gamma(\mathbf{K}_t)Z] \quad (19)$$

hence the imaginary part of G_0 is determined only by the propagating spectrum. This is consistent with the fact, demonstrated earlier, that the imaginary part of G_0 governs, under conventional far-field sensing, the 3D imaging resolution, giving rise to the fundamental Rayleigh-criterion-based $\lambda/2$ resolution limit.

Now, using Eqs. (3) and (16) one finds that the field $\psi(\mathbf{r})$ can be represented for $z > z_0$ as

$$\psi(\mathbf{r}) = -\frac{i}{8\pi^2} \int_{-\infty}^{\infty} \frac{d^2 K_t}{\gamma(\mathbf{K}_t)} \hat{\rho} \left[\mathbf{K}_t + \gamma(\mathbf{K}_t) \hat{\mathbf{z}} \right] e^{i[\mathbf{K}_t \cdot \mathbf{r}_t + \gamma(\mathbf{K}_t)Z]} \quad (20)$$

where $\mathbf{r}_t = (x, y)$. Thus the corresponding plane wave spectrum of the field is governed by the spatial Fourier transform $\hat{\rho}$ of the source ρ , and its analytic continuation for complex argument, as required in connection with the evanescent portion of the field spectrum. For z near the boundary $z = z_0$ this field has significant evanescent components, which carry high-frequency information about the source and can thus lead to high imaging resolution in the associated inversion. On the other hand, $z - z_0$ greater than a

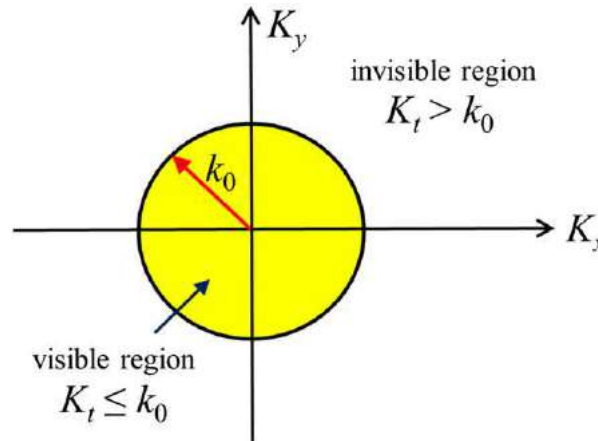


Fig. 7 Regions in Fourier domain corresponding to the visible (propagating) and invisible (evanescent) field components. The visible region is populated by far-field data.

few wavelengths the field is mostly propagating and can be approximated as

$$\psi(\mathbf{r}) = -\frac{i}{8\pi^2} \int_{K_t \leq k_0} \frac{d^2 K_t}{\gamma(K_t)} \tilde{\rho}[\mathbf{K}_t + \gamma(K_t)\hat{\mathbf{z}}] e^{i[\mathbf{K}_t \cdot \mathbf{r}_t + \gamma(K_t)z]} \quad (21)$$

so that it carries only the low-frequency information about the source, which is in fact identical to the far-field one that lies in the Ewald sphere in the Fourier domain, as was explained earlier.

Transverse Resolution

In many optical imaging systems, such as in conventional microscopy, the object to be imaged lies over a planar surface. In this context it is convenient to characterize the resolution limits associated to the imaging of the field over a planar surface. Furthermore, this characterization is also relevant to the imaging of more arbitrary objects. In particular, it sheds insight on the resolution with which the near field in the vicinity of the object can be imaged, which is in turn linked to the resolution with which the object itself can be resolved. Thus, referring to the geometry in Fig. 6, we consider next the classical resolution associated to the imaging of the field over the boundary $z=z_0$ that bounds the support τ of radiating source ρ . It is assumed that the field data are gathered over the boundary $z=z_1$ where $z_1 - z_0$ is greater than several wavelengths so that the evanescent components cannot be reliably sensed there.

This can be addressed within the plane wave spectrum representation of the field by considering an equivalent Huygens' source linked to the field, for example, the double-layer singular surface source

$$\rho_s(\mathbf{r}) = -2\psi(\mathbf{r}_t, z_0)\delta'(z - z_0) \quad (22)$$

where $\delta'(z) = \frac{d}{dz}\delta(z)$, which leads to the first Rayleigh–Sommerfeld formula (Devaney, 2012, p. 75; Goodman, 2005, p. 48)

$$\psi(\mathbf{r}) = -2 \int_{-\infty}^{\infty} d^2 r'_t \psi(\mathbf{r}'_t, z_0) \frac{\partial}{\partial z'} G_0(\mathbf{r} - \mathbf{r}'), \quad z > z_0 \quad (23)$$

Using this equivalent source in Eq. (21) one obtains

$$\psi(\mathbf{r}_t, z_1) = \frac{i}{4\pi^2} \int_{K_t \leq k_0} d^2 K_t \tilde{\psi}(\mathbf{K}_t, z_0) e^{i[\mathbf{K}_t \cdot \mathbf{r}_t + \gamma(K_t)z_1]} \quad (24)$$

Therefore, the information about the near field at $z=z_0$ that is contained in the measured field at $z=z_1$ is its 2D spatial Fourier transform $\tilde{\psi}(\mathbf{K}_t, z_0)$ evaluated over the ball of radius k_0 centered at the origin in 2D Fourier space (the visible region shown in Fig. 7). This means that the largest spatial frequency magnitude contained in the data is $k_0 = 2\pi/\lambda$, so that the minimum spatial period that can be resolved is λ . This is a fundamental measure of the achievable resolution, in particular,

$$\Delta_{2D} = \lambda \quad (25)$$

based on standard bandlimitation considerations. In addition, one can also consider the Rayleigh resolution criterion. For this purpose, we note that the transfer function of the propagation map is $H(\mathbf{K}_t) = U(k_0 - K_t)$ where U is the unit-step function. The respective IFT corresponds to the PSF h of the imaging system which can be shown to be equal to

$$h(\mathbf{r}_t) = 2\pi k_0 J_1(k_0 \rho) / \rho \quad (26)$$

where $\rho = |\mathbf{r}_t|$. A plot of this PSF is shown in Fig. 5. The first zero of this PSF occurs at $k_0 \rho \approx 1.22\pi$, corresponding to a (Rayleigh-criterion-based) 2D or transverse field resolution of

$$\Delta_{2D}^R = 0.61\lambda \quad (27)$$

This corresponds to the resolution limit if the field is sensed over the entire plane $z=z_1$. In addition to the intrinsic loss of information (the evanescent wave part) due to wave propagation, there is also an additional loss due to the finite size of the sensing apparatus. In this realistic case if the aperture covers a polar angle 2θ then the detected transverse frequencies are such that $K_t \leq k_0 \sin \theta$ and therefore the respective PSF is

$$h(\mathbf{r}_t) = 2\pi k_0 \sin \theta J_1(k_0 \sin \theta \rho) / \rho \quad (28)$$

which gives the Rayleigh criterion transverse resolution

$$\Delta_{2D}^R = \frac{0.61\lambda}{\text{NA}} \quad (29)$$

where $\text{NA} = \sin \theta$ is the numerical aperture of the apparatus.

Resolution in Scatterer Imaging

In many application areas (such as radar, lidar, biomedical imaging, geophysical imaging, nondestructive testing, to mention a few) a medium is probed using known incident waves, and the medium's response, in the form of the resulting perturbation or scattered waves, is measured with sensors located in the far zone (see Fig. 3). By processing the measured scattered field data, it is possible to obtain images of the scattering medium or target, which give information about its localization, shape, and material

properties (e.g., index of refraction). If the surrounding background medium is free space, then the scattered field, to be denoted as $\psi_s(\mathbf{r})$, obeys, within the scalar formulation, a differential equation of the form

$$(\nabla^2 + k_0^2)\psi_s(\mathbf{r}) = V(\mathbf{r})\psi(\mathbf{r}) \quad (30)$$

where k_0 is the wavenumber in free space, $V(\mathbf{r})$ is the space-dependent scattering potential, which is related to the scatterer's space-dependent index of refraction $n(\mathbf{r})$ via

$$V(\mathbf{r}) = k_0^2[1 - n^2(\mathbf{r})] \quad (31)$$

and $\psi(\mathbf{r})$ is the total field, which is given by

$$\psi(\mathbf{r}) = \psi_i(\mathbf{r}) + \psi_s(\mathbf{r}) \quad (32)$$

where $\psi_i(\mathbf{r})$ is the probing or incident field and the scattered field $\psi_s(\mathbf{r})$ obeys the radiation condition at infinity, as required by causality considerations.

It follows from Eqs. (1), (3), and (30) (via the substitution $\rho \rightarrow V\psi$) that

$$\psi_s(\mathbf{r}) = \int_{\tau} d^3r' G_0(\mathbf{r} - \mathbf{r}') V(\mathbf{r}') \psi(\mathbf{r}') \quad (33)$$

where τ denotes the scatterer's support. Furthermore, using Eq. (7) in Eq. (33) we obtain (in analogy to Eqs. (8) and (9)) the far zone behavior of the scattered field:

$$\psi(r\mathbf{s}) \sim f_s(\mathbf{s}; \psi_i) \frac{e^{ik_0 r}}{r} \quad (34)$$

where the quantity $f_s(\mathbf{s}; \psi_i)$ is the far-zone scattering amplitude, which depends on the observation direction $\mathbf{s} \in S^2$ and the incident field ψ_i , and is given by

$$f(\hat{\mathbf{r}}) = -\frac{1}{4\pi} \int_{\tau} d^3r V(\mathbf{r}) \psi(\mathbf{r}) e^{-ik_0 \mathbf{s} \cdot \mathbf{r}} \quad (35)$$

For the typical situation in which the incident field is a uniform plane wave traveling in the direction of the unit vector $\mathbf{s}_0$, i.e.,

$$\psi_i(\mathbf{r}) = e^{ik_0 \mathbf{s}_0 \cdot \mathbf{r}} \quad (36)$$

then it is convenient to denote $f_s(\mathbf{s}; \psi_i)$ as $f_s(\mathbf{s}; \mathbf{s}_0)$. For this case we obtain from Eqs. (32), (35), and (36)

$$f_s(\mathbf{s}; \mathbf{s}_0) = f_s^B(\mathbf{s}; \mathbf{s}_0) + f_s^{MS}(\mathbf{s}; \mathbf{s}_0) \quad (37)$$

where f_s^B is the scattering amplitude under the weak scattering or Born approximation, which is given by

$$f_s^B(\mathbf{s}; \mathbf{s}_0) = -\frac{1}{4\pi} \int_{\tau} d^3r V(\mathbf{r}) e^{-ik_0(\mathbf{s} - \mathbf{s}_0) \cdot \mathbf{r}} = -\frac{1}{4\pi} \tilde{V}(\mathbf{K})|_{\mathbf{K} = k_0(\mathbf{s} - \mathbf{s}_0)} \quad (38)$$

where $\tilde{V}$ denotes the spatial Fourier transform of the scattering potential function V , and f_s^{MS} is the multiple scattering contribution to the total scattering amplitude, which is given by

$$f_s^{MS}(\mathbf{s}; \mathbf{s}_0) = -\frac{1}{4\pi} \int_{\tau} d^3r V(\mathbf{r}) \psi_s(\mathbf{r}) e^{-ik_0 \mathbf{s} \cdot \mathbf{r}} \quad (39)$$

If multiple scattering is weak, so that $|\psi_s| \ll |\psi_i|$ and consequently $|f_s^{MS}| \ll |f_s^B|$ then from Eq. (37) approximately

$$f_s(\mathbf{s}; \mathbf{s}_0) \simeq f_s^B(\mathbf{s}; \mathbf{s}_0) \quad (40)$$

and according to Eq. (38) the scattering amplitude is given by the spatial Fourier transform $\tilde{V}$ of the scattering potential V evaluated at $\mathbf{K} = k_0(\mathbf{s} - \mathbf{s}_0)$. If the scattering amplitude is measured for a fixed incidence direction $\mathbf{s}_0$ and for all the observation directions $\mathbf{s}$, the resulting scattering data populates the shifted Ewald sphere where $|\mathbf{K} - k_0 \mathbf{s}_0| = k_0$ (see Fig. 4). If probing is done using all the incidence directions $\mathbf{s}_0$ and sensing is done in all observation directions $\mathbf{s}$ then the resulting scattering data populates in Fourier space a ball of radius $2k_0$ that is centered at the origin. It is bounded by the so-called limiting Ewald sphere, and is illustrated and referred to as the limiting Ewald ball in Fig. 4. We focus next on this situation of full-view probing and sensing. In this weak scattering regime, the map from the scattering potential V to the scattering amplitude is linear, so that one can compute a well-defined PSF for the relevant computational imaging. The PSF in question is given by (Simonetti, 2006, Eq. (17))

$$h(\mathbf{r}) = \frac{8j_1(2k_0 r)}{\lambda^2 r} \quad (41)$$

A plot of this PSF is shown in Fig. 5. Based on the first zero of the spherical Bessel function j_1 (whose approximate value is 4.493), an approximate value of the Born approximation scattering resolution based on Rayleigh's criterion is

$$\Delta_{3D}^R \simeq 0.358\lambda \quad (42)$$

But this is only a rough resolution estimate. Moreover, it is pertinent only for the special case of two point-like scatterers. It is more general to simply state that the largest spatial frequency magnitude contained in the data is $2k_0 = 4\pi/\lambda$ which implies, in turn, that the minimum spatial period that can be resolved is $\lambda/2$, and this is the familiar resolution limit discussed in Wolf's classical treatment of the inverse scattering problem in the framework of the Born approximation (Wolf, 1969, Eq. (22)). In particular,

based on this minimum spatial period criterion which is relevant to arbitrary scatterers,

$$\Delta_{3D,s} = \lambda/2 \quad (43)$$

On the other hand, if multiple scattering is not negligible, then the scattering amplitude is given by Eq. (37) where the term f_s^{MS} is given by Eq. (39) and cannot be neglected. Substituting from Eq. (33) in Eq. (39) we find that

$$\begin{aligned} f_s^{\text{MS}}(\mathbf{s}; \mathbf{s}_0) &= -\frac{1}{4\pi} \int_{\tau} d^3r V(\mathbf{r}) \psi_s(\mathbf{r}) e^{-ik_0 \mathbf{s} \cdot \mathbf{r}} \\ &= -\frac{1}{4\pi} \int_{\tau} d^3r V(\mathbf{r}) (\mathbf{r}) e^{-ik_0 \mathbf{s} \cdot \mathbf{r}} \int_{\tau} d^3r' G_0(\mathbf{r} - \mathbf{r}') V(\mathbf{r}') \psi(\mathbf{r}') \end{aligned} \quad (44)$$

which clearly showcases the nonlinear nature of the map from V to f_s^{MS} (and, in turn, to f_s). This prohibits the use of bandlimitation considerations in Fourier domain to quantify the corresponding imaging resolution. Alternative criteria that apply also to nonlinear systems, such as a criterion based on the fundamental CRB have been proposed (Sentenac *et al.*, 2007), and applied to quantify the effects of multiple scattering in imaging (Sentenac *et al.*, 2007; Simonetti *et al.*, 2008; Marengo and Berestesky, 2013; Marengo *et al.*, 2012).

Simonetti (2006) has shown that multiple scattering can enhance imaging resolution, relative to the Born-approximation-based resolution limits, thanks to the conversion of evanescent waves in the scatterer into propagating waves that are accessible in the far zone. This mechanism implies the encoding of high-frequency information about the scatterer, beyond that associated (in the weak scattering approximation) to the limiting Ewald ball. Thus superresolution is possible thanks to multiple scattering. The key argument can be outlined with the help of Eqs. (6) and (44). Using Eq. (6) in Eq. (44) one obtains

$$f_s^{\text{MS}}(\mathbf{s}; \mathbf{s}_0) = -\frac{1}{4\pi} \int d^3K \frac{\tilde{V}(k_0 \mathbf{s}_0 - \mathbf{K}) T(\mathbf{K}, k_0 \mathbf{s}_0)}{k_0^2 - K^2 + i\epsilon} \quad (45)$$

where

$$T(\mathbf{K}, k_0 \mathbf{s}_0) = \int_{\tau} d^3r \rho(\mathbf{r}) e^{-i\mathbf{K} \cdot \mathbf{r}} V(\mathbf{r}) \psi(\mathbf{r}) \quad (46)$$

The integral in Eq. (45) is over the entire $\mathbf{K}$ space, so that it depends on the value of the spatial Fourier transform $\tilde{V}$ of V for all spatial frequencies, including those outside the limiting Ewald ball. As a result, it is possible, thanks to multiple scattering, to encode information in the far-field scattering amplitude about the high spatial frequencies of the object corresponding to the exterior of the limiting Ewald ball, leading to the possibility of superresolution.

There is a lot of evidence that multiple scattering can, indeed, lead to superresolution. This includes evidence based on experimental data (Chen and Chew, 1998), numerical inversions incorporating multiple scattering (Cui *et al.*, 2004; Belkebir *et al.*, 2006), and algorithm-independent, estimation-theoretic studies based on the fundamental CRB corresponding to a given noise level (Sentenac *et al.*, 2007; Simonetti *et al.*, 2008; Marengo and Berestesky, 2013). On the other hand, it has been found that while multiple scattering is typically beneficial and enables superresolution beyond what is possible for weak scatterers obeying the Born approximation, for certain sensing conditions its impact is minor and can even be detrimental (Sentenac *et al.*, 2007; Marengo and Berestesky, 2013). The quantification of the role of multiple scattering in superresolution is so far inconclusive and is an important area of ongoing research.

Superresolution

The resolution limit discussed so far applies to conventional imaging systems employing far-field probing (involving propagating waves only) and far-field sensing. If the interrogation of the object involves probing or sensing in the near zone, then imaging resolution is greatly improved (leading to “superresolution”) thanks to the sensing of the high-frequency information about the object that is contained in the corresponding near-field evanescent waves. This is the idea behind near-field optics techniques such as near-field scanning microscopy (Betzig and Trautman, 1992) and total internal reflection tomography (Carney and Schotland, 2001; Chaumet *et al.*, 2004).

Another approach to superresolution is based on the analyticity of the radiated fields, and consists of analytically continuing the field data associated to the propagating (or visible) region of the plane wave spectrum into the complementary evanescent (or invisible) region of the spectrum (Goodman, 2005, Section 6.6). However, the required analytic continuation is highly sensitive to noise in the data, and consequently this method is hard to implement in practice. A more reliable methodology which is well-known (Schmidt-Weinmar, 1978) to render superresolution in many applications is to incorporate into the inversion and imaging algorithms prior information about the object, such as support information (see Marengo *et al.*, 1999, 2000; Schmidt-Weinmar, 1978; Marengo, 2014; Goodman, 2005, Section 6.6.4). In the latter approach, the superresolution is due to the exploitation of prior information about the sought-after object. Here it is important to emphasize that the focus of the present contribution has been on the classical assessment of resolution, in which no prior information about the object is assumed. In modern imaging, several techniques have emerged which successfully exploit prior information, and can even achieve superresolution thanks to the relevant prior. In this connection, the incorporation of sparsity priors by means of the methods of compressive sensing has been particularly successful (Candes, 2006).

Recently, there has also been a lot of interest in achieving superresolution via direct physical imaging using innovative devices and sensors. One of these approaches is the so-called superlens proposed by Pendry (2000), in which left-handed metamaterials play a key role. In these approaches, the idea is to build a device that can amplify or sense in the far zone the evanescent waves emanating from a radiating object. The experimental realization has been pursued by many groups (Grbic and Eleftheriades, 2004), and several alternative technologies such as the hyperlens (Jacob *et al.*, 2006) have also been proposed.

References

- Belkebir, K., Chaumet, P.C., Sentenac, A., 2006. Influence of multiple scattering on three-dimensional imaging with optical diffraction tomography. *Journal of the Optical Society of America A* 23, 586–595.
- Betzig, E., Trautman, J.K., 1992. Near-field optics: Microscopy, spectroscopy, and surface modification beyond the diffraction limit. *Science* 257, 189–195.
- Bleistein, N., Cohen, J.K., 1977. Nonuniqueness in the inverse source problem in acoustics and electromagnetics. *Journal of Mathematical Physics* 18, 194–201.
- Bouwkamp, C.J., de Bruijn, N.G., 1946. The problem of optimum antenna current distribution. *Philips Research Reports* 1, 135–158.
- Buxton, A., 1937. Note on optical resolution. *Philosophical Magazine* 23, 440–442.
- Candes, E.J., 2006. Compressive sampling. In: *Proceedings of the International Congress of Mathematicians, Madrid*, pp. 1–20.
- Carney, P.S., Schotland, J.C., 2001. Three-dimensional total internal reflection tomography. *Optics Letters* 26, 1072–1074.
- Chaumet, P.C., Belkebir, K., Sentenac, A., 2004. Superresolution of three-dimensional optical imaging by use of evanescent waves. *Optics Letters* 29, 2740–2742.
- Chen, F.-C., Chew, W.C., 1998. Experimental verification of super resolution in nonlinear inverse scattering. *Applied Physics Letters* 72, 3080–3082.
- Cui, T.J., Chew, W.C., Yin, X.X., Hong, W., 2004. Study of resolution and super resolution in electromagnetic imaging for half-space problems. *IEEE Transactions on Antennas and Propagation* 52, 1398–1411.
- den Dekker, A.J., van den Bos, A., 1997. Resolution: A survey. *Journal of the Optical Society of America A* 14, 547–557.
- Devaney, A.J., 2012. *Mathematical Foundations of Imaging, Tomography and Wavefield Inversion*. Cambridge: Cambridge University Press.
- Devaney, A.J., Wolf, E., 1973. Radiating and nonradiating classical current distributions and the fields they generate. *Physical Review* 8, 1044–1047.
- Gbur, G., 2003. Nonradiating sources and other invisible objects. In: Wolf, E. (Ed.), *Progress in Optics*, vol. 45. Amsterdam: Elsevier.
- Grbic, A., Eleftheriades, G.V., 2004. Overcoming the diffraction limit with a planar left-handed transmission-line lens. *Physical Review Letters* 92, 117403.
- Goodman, J.W., 2005. *Introduction to Fourier Optics*, third ed. Englewood, CO: Roberts & Company.
- Heintzmann, R., Fic, G., 2006. Breaking the resolution limit in light microscopy. *Briefings in Functional Genomics and Proteomics* 5, 289–301.
- Houston, W.V., 1927. A compound interferometer for fine structure work. *Physical Review* 29, 478–484.
- Jacob, Z., Alekseyev, L.V., Narimanov, E., 2006. Optical hyperlens: Far-field imaging beyond the diffraction limit. *Optics Express* 14, 8247–8256.
- Kay, S.M., 1993. *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. New Jersey, NJ: Prentice Hall.
- Lord Rayleigh, F.R.S., 1879. Investigations in optics, with special reference to the spectroscope. London, Edinburgh, and Dublin *Philosophical Magazine and Journal of Science*, Series 5 8, 261–274.
- Marengo, E.A., 2014. Inverse diffraction theory and computation of minimum source regions of far fields. *Mathematical Problems in Engineering* 2014, 1–18. Article ID: 513953.
- Marengo, E.A., Berestetsky, P., 2013. Cramer–Rao bound study of multiple scattering effects in target separation estimation. *International Journal of Antennas and Propagation* 2013, 1–10. Article ID: 572923.
- Marengo, E.A., Devaney, A.J., Ziolkowski, R.W., 1999. New aspects of the inverse source problem with far field data. *Journal of the Optical Society of America A* 16, 1612–1622.
- Marengo, E.A., Devaney, A.J., Ziolkowski, R.W., 2000. Inverse source problem and minimum energy sources. *Journal of the Optical Society of America A* 17, 34–45.
- Marengo, E.A., Zambrano-Nunez, M., Berestetsky, P., 2012. Cramer–Rao study of multiple scattering effects in target localization. *International Journal of Antennas and Propagation* 2012, 1–13. Article ID: 390312.
- Marengo, E.A., Ziolkowski, R.W., 2000. Nonradiating and minimum energy sources and their fields: Generalized source inversion theory and applications. *IEEE Transactions on Antennas and Propagation* 48, 1553–1562.
- Pendry, J.B., 2000. Negative refraction makes a perfect lens. *Physical Review Letters* 85, 3966–3969.
- Ronchi, V., 1961. Resolving power of calculated and detected images. *Journal of the Optical Society of America* 51, 458–460.
- Rushforth, C.K., Harris, R.W., 1968. Restoration, resolution, and noise. *Journal of the Optical Society of America* 58, 539–545.
- Schuster, A., 1924. *An Introduction to the Theory of Optics*, third ed. London: Arnold.
- Schmidt-Weinmar, H.G., 1978. Spatial resolution of subwavelength sources from optical far-zone data. In: Baltes, H.P. (Ed.), *Inverse Source Problems in Optics*. Berlin: Springer-Verlag, pp. 83–118.
- Sentenac, A., Guerin, C.-A., Chaumet, P.C., *et al.*, 2007. Influence of multiple scattering on the resolution of an imaging system: A Cramer–Rao analysis. *Optics Express* 15, 1340–1347.
- Simonetti, F., 2006. Multiple scattering: The key to unravel the subwavelength world from the far-field pattern of a scattered wave. *Physical Review E* 73, 036619.
- Simonetti, F., Fleming, M., Marengo, E.A., 2008. Illustration of the role of multiple scattering in subwavelength imaging from far-field measurements. *Journal of the Optical Society of America A* 25, 292–303.
- Sparrow, C.M., 1916. On spectroscopic resolving power. *Astrophysics Journal* 44, 76–86.
- Testorf, M.E., Fiddy, M.A., 2010. Superresolution imaging: Revisited. In: Hawkes, P.W. (Ed.), *Advances in Imaging and Electron Physics*, vol. 163. San Diego, CA: Academic Press.
- Toraldo di Francia, G., 1952. Super-gain antennas and optical resolving power. *Il Nuovo Cimento* 9, 426–438.
- Toraldo di Francia, G., 1955. Resolving power and information. *Journal of the Optical Society of America* 45, 497–501.
- Wolf, E., 1969. Three-dimensional structure determination of semi-transparent objects from holographic data. *Optics Communications* 1, 153–156.
- Woodward, P.M., Lawson, J.D., 1948. The theoretical precision with which an arbitrary radiation-pattern may be obtained from a source of a given size. *Journal of Electrical Engineering* 95, 363–370.
- Zheludev, N.I., 2008. What diffraction limit? *Nature Materials* 7, 420–422.

Coherent Diffractive Imaging

Lei Tian, Boston University, Boston, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction

The infamous “phase problem” stems from that optical detection devices (e.g., charge-coupled devices, photodetectors, and human eyes) cannot directly record the phase of the electromagnetic field (at 10^{15} Hz and higher), but only the average power, which is proportional to the magnitude square of the field. As a result, obtaining phase involves additional complexity, both experimentally and computationally.

In general, the first step of solving the phase problem is to convert the originally “invisible” phase information to the observable intensity variations, i.e., phase contrast. In fact, any *complex* transfer function will produce some phase contrast. For example, the Zernike phase contrast is achieved by introducing a $\pi/2$ phase shift (and thus a pure imaginary transfer function component) to the direct illumination beam. The second step is to carry out a phase retrieval procedure to computationally recover the quantitative phase values.

Many techniques have been developed to recover phase information. Depending on the physical mechanism of realizing phase contrast, they can be broadly categorized into interferometric and “non-interferometric” techniques – the former relies on interferometry and the latter does not. Here, we focus on the non-interferometric techniques. A distinctive advantage of these techniques is that phase contrast can be produced simply by some diffraction process, which means that the experimental setup can be made simple (e.g., lensless). The non-interferometric techniques are often referred to as “coherent diffractive imaging (CDI),” emphasizing the role of diffraction in forming phase contrast. As compared to interferometry, CDI often requires more sophisticated computational algorithms to reconstruct the phase, since the diffraction measurements are often related to the phase by nonlinear equations.

From a theoretical point of view, phase retrieval may not have a unique solution; furthermore, even if a unique solution exists, it is not guaranteed that the algorithm will converge to the global solution because the underlying problem is nonlinear and nonconvex. The success of the phase retrieval algorithms depends on not only the numerical procedure itself, but also the physical measurement process that defines the nonlinear problem to solve. Generally speaking, successful phase retrieval requires both redundancy and diversity in the data. The optimal design involves trade-offs between experimental complexity, cost, accuracy, and computational requirement. The optimal implementation is highly application dependent, which requires joint designs of physical optical instrumentation and computation.

Fourier Measurement-Based Techniques

The Fourier measurement is the most widely used method in CDI. A typical experimental setup based on far-field diffraction is illustrated in Fig. 1. Under coherent illumination, the measured far-field intensity is simply the Fourier intensity of the object,

$$I(\mathbf{u}) = \left| \int g(\mathbf{r}) \exp\{-i2\pi\mathbf{r} \cdot \mathbf{u}\} d^2\mathbf{r} \right|^2 = |\mathcal{F}\{g\}|^2 \quad (1)$$

where $\mathcal{F}\{\cdot\}$ denotes the two-dimensional (2D) Fourier transform (FT), $g(\mathbf{r}) = |g(\mathbf{r})| \exp\{i\phi(\mathbf{r})\}$ is the object's complex transmittance function having amplitude $|g(\mathbf{r})|$ and phase $\phi(\mathbf{r})$, and $\mathbf{r}$ and $\mathbf{u}$ denote the spatial and Fourier coordinates, respectively. The knowledge of the Fourier intensity alone does not allow unique recovery of the object, since there are infinitely many possible phase functions that satisfy the equation. The general strategy is to introduce additional prior knowledge and/or constraints to

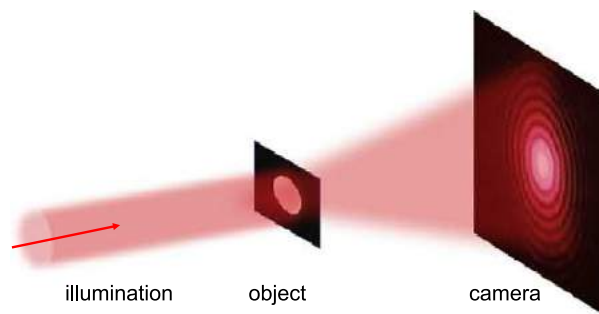


Fig. 1 Schematics of the Fourier measurement-based coherent diffractive imaging (CDI) technique.

regularize the problem. In the following, we will describe different algorithms that have been designed for solving problems that incorporate different priors.

Alternating Projection Algorithms

Alternating projection is the most popular class of phase retrieval algorithms. The general idea is to treat the constraint at the Fourier domain as one set, and the constraint at the real-space (at the object plane) as the other. The algorithm then iteratively updates the solution by projecting the estimated field between the two sets in an alternating order.

The pioneering work of Gerchberg and Saxton (GS) in 1972 assumes both the Fourier magnitude $|G(\mathbf{u})| = \sqrt{I(\mathbf{u})}$ and the real-space magnitude $|g(\mathbf{r})|$ are known. The algorithm starts with some phase initial guess at the real space, for example, all-zero phase or uniformly random-distributed values within certain range. It then iteratively imposes the real-space and Fourier magnitude constraints in the following four steps, as illustrated in Fig. 2. In the s th iteration, the algorithm first takes the FT of the estimated object $g_s(\mathbf{r})$ to get the estimated Fourier spectrum $G_s(\mathbf{u})$,

$$G_s(\mathbf{r}) = \mathcal{F}\{g_s(\mathbf{u})\} \quad (2)$$

Second, the computed Fourier magnitude is then replaced by the measurement, while keeping the estimated phase unchanged.

$$G'_s(\mathbf{u}) = G_s(\mathbf{u}) \cdot \frac{\sqrt{I(\mathbf{u})}}{|G_s(\mathbf{u})|} \quad (3)$$

Third, it takes the inverse FT, $\mathcal{F}^{-1}\{\cdot\}$, of the updated Fourier spectrum $G'_s(\mathbf{u})$ to update the object field $g'_s(\mathbf{r})$,

$$g'_s(\mathbf{r}) = \mathcal{F}^{-1}\{G'_s(\mathbf{u})\} \quad (4)$$

Fourth, the real-space magnitude constraint is imposed by replacing the estimate with the known values, while keeping the estimated phase unchanged.

$$g_{s+1}^{\text{GS}}(\mathbf{r}) = |g(\mathbf{r})| \cdot \frac{g'_s(\mathbf{r})}{|g'_s(\mathbf{r})|} \quad (5)$$

The iteration continues until the reconstruction error between the estimated and the measured Fourier magnitude, $\sum_{\mathbf{u}} ||G_s(\mathbf{u})| - \sqrt{I(\mathbf{u})}|$, below certain value.

The error-reduction (ER) algorithm proposed by Fienup in 1978 extends the GS algorithm to incorporate other types of priors, such as the object support and nonnegativity constraints. The ER algorithm follows the similar four steps in each iteration.

The only difference is in the fourth step, in which other real-space constraints are imposed by

$$g_{s+1}^{\text{ER}}(\mathbf{r}) = \begin{cases} g'_s(\mathbf{r}), & m \notin \gamma \\ 0, & m \in \gamma \end{cases} \quad (6)$$

where γ defines the set of indices for which $g'_s(\mathbf{r})$ violates the real-space constraints.

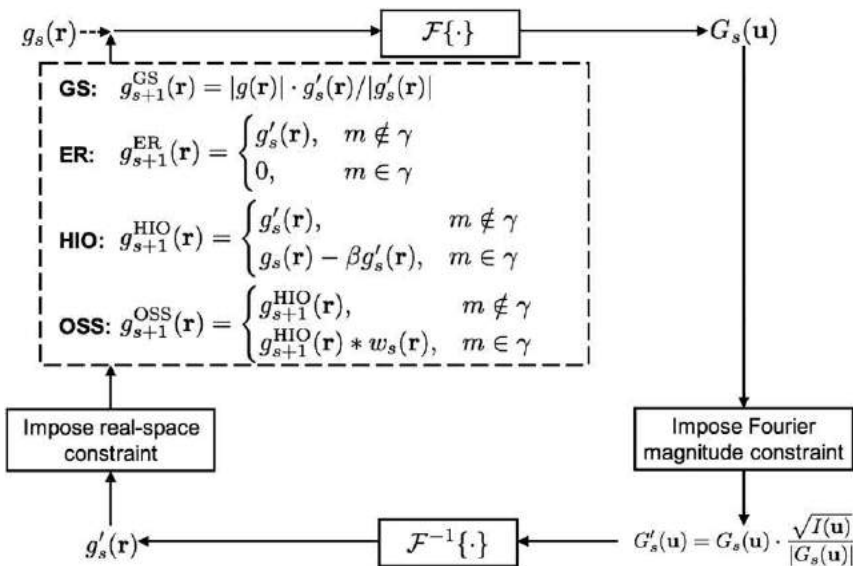


Fig. 2 Block diagram of the Gerchberg–Saxton (GS), error reduction (ER), hybrid input–output (HIO), and oversampling smoothness (OSS) algorithms.

The reconstruction errors of both GS and ER algorithms are shown to be monotonically decreasing with iterations. Despite this convergence property, it is not guaranteed that the true solution can be found, since the algorithm may converge to a local minimum. The underlying problem, which can be written in the form of a multivariate minimization of a cost function with respect to the unknown phase function, is both nonlinear and nonconvex in a high-dimensional space. Here, the (non)convexity characterizes the hyper-surface formed by the cost function. A nonconvex surface consists of multiple peaks and valleys, which implies the existence of multiple local minimal solutions.

The hybrid input-output (HIO) algorithm is another commonly used alternating projection algorithm. It follows the same iterative framework as GS and ER with a modification to the fourth step, as given by

$$g_{s+1}^{\text{HIO}}(\mathbf{r}) = \begin{cases} g_s'(\mathbf{r}), & m \notin \gamma \\ g_s(\mathbf{r}) - \beta g_s'(\mathbf{r}), & m \in \gamma \end{cases} \quad (7)$$

where β is a feedback constant, with a typical value between 0.5 and 1. Empirically, it is shown that HIO is able to avoid local minima and converges to the global solution in the noise-free case. In practice, the solution tends to oscillate as the iteration increases. The algorithm may stagnate and fails to converge to the global minimum.

Many algorithms have been developed to overcome these limitations, including the combination of HIO and ER, solvent flipping (SF), different map (DM), hybrid projection-reflection (HPR), average successive reflections (ASR), relaxed averaged alternating reflectors (RAAR), oversampling smoothness (OSS), and continuous HIO (CHOI). As an example, both CHIO and OSS attempt to smooth the updates by removing sharp transitions in the estimates. The OSS update is related to HIO by

$$g_{s+1}^{\text{OSS}}(\mathbf{r}) = \begin{cases} g_{s+1}^{\text{HIO}}(\mathbf{r}), & m \notin \gamma \\ g_{s+1}^{\text{HIO}}(\mathbf{r}) * w_s(\mathbf{r}), & m \in \gamma \end{cases} \quad (8)$$

where $*$ denotes convolution. For the set of indices that violates the real-space constraints, the HIO solution is low-pass filtered by a Gaussian function $w_s(\mathbf{r})$, whose variance increases with iterations.

The convergence property of HIO and its variants is greatly affected by the support constraint. In many applications, the support is not exactly known a priori. Estimation techniques have been developed based on the support of the object's auto-correlation function (i.e., FT of the measured Fourier intensity). The support can be further refined during iterations using the "shrink-wrap" algorithm.

The performance of the alternating projection phase retrieval algorithm is also dependent on the initial guess and it may converge to different solutions due to the nonconvexity of the problem. Initializations with different random phase distributions have been proposed. The final solution is then constructed by some numerical averaging scheme. For example, the phase retrieval transform function (PRTF) is proposed to quantify the computationally recovered spatial resolution, as defined by

$$\text{PRTF}(\mathbf{u}) = \langle |\mathcal{F}\{\hat{g}_t(\mathbf{r})\}| \rangle_t / \sqrt{I(\mathbf{u})} \quad (9)$$

where $\hat{g}_t(\mathbf{r})$ denotes the estimated object from the t th reconstruction, and $\langle \cdot \rangle$ takes the average over all the reconstructions using different initial guesses. In cases where a low-resolution object is known or can be first estimated and then used as the initial guess, the algorithm tends to converge to the global solution more reliably and reconstruct high resolution. Examples include the "expanding Fourier modulus" method and ptychographic algorithms.

Partial Coherence Algorithms

The FT model in Eq. (1) assumes that the illumination beam is spatially and temporally fully coherent, which does not hold for many applications. Directly applying the coherent phase retrieval algorithm suffers from significant reconstruction artifacts in partially coherent CDI problem.

The diffraction intensity $I^{\text{PC}}(\mathbf{u})$ under partially coherent illumination is described through the mutual intensity function $J(\mathbf{r}_1, \mathbf{r}_2)$, which is a measure of the statistical correlations between a pair of points at $\mathbf{r}_1$ and $\mathbf{r}_2$ in the partially coherent field,

$$I^{\text{PC}}(\mathbf{u}) = \int \int J(\mathbf{r}_1, \mathbf{r}_2) g(\mathbf{r}_1) g^*(\mathbf{r}_2) \exp\{-i2\pi\mathbf{u} \cdot (\mathbf{r}_1 - \mathbf{r}_2)\} d^2\mathbf{r}_1 d^2\mathbf{r}_2 \quad (10)$$

The first approach to simplify Eq. (10) is to use the coherent mode decomposition model, which states that any partially coherent field can be decomposed into a set of coherent fields that are mutually incoherent,

$$J(\mathbf{r}_1, \mathbf{r}_2) = \sum_{n=1}^N \mu_n \psi_n(\mathbf{r}_1) \psi_n^*(\mathbf{r}_2) \quad (11)$$

The coherent modes are described by a set of orthonormal functions $\psi_n(\mathbf{r})$. The real, nonnegative number μ_n characterizes the amount of energy contained in each mode. Eq. (11) assumes that the majority of the energy in the illumination is contained in the first N modes. The diffraction intensity can be rewritten as

$$I^{\text{PC}}(\mathbf{u}) = \sum_{n=1}^N \mu_n \left| \int \psi_n(\mathbf{r}) g(\mathbf{r}) \exp\{-i2\pi\mathbf{u} \cdot \mathbf{r}\} d^2\mathbf{r} \right|^2 \quad (12)$$

which is the incoherent sum of the coherent diffraction intensities from the object illuminated with each coherent mode.

The second approach further assumes the illumination is statistically stationary and the mutual intensity function depends only on the spatial separations $\Delta\mathbf{r}=\mathbf{r}_1-\mathbf{r}_2$. The diffraction intensity takes the following convolution form,

$$I^{\text{PC}}(\mathbf{u}) = \left| \int g(\mathbf{r}) \exp \{-i2\pi\mathbf{u} \cdot \mathbf{r}\} d^2\mathbf{r} \right|^2 * S(\mathbf{u}) = I^{\text{C}}(\mathbf{u}) * S(\mathbf{u}) \quad (13)$$

where $I^{\text{C}}(\mathbf{u})$ is the intensity under coherent illumination; $S(\mathbf{u})$ is the FT of $J(\Delta\mathbf{r})$, which represents the intensity distribution of the primary source according to the Van Cittert-Zernike theorem.

When the coherence function of the illumination is known, the partially coherent phase retrieval can be achieved using a modified alternating projection algorithm that follows the similar four-step process. The only modification is to the second Fourier spectrum estimation step,

$$G'_s(\mathbf{u}) = G_s(\mathbf{u}) \cdot \sqrt{\frac{I(\mathbf{u})}{I_s^{\text{PC}}(\mathbf{u})}} \quad (14)$$

where $I_s^{\text{PC}}(\mathbf{u})$ denotes the estimated partially coherent diffraction intensity using either Eq. (12) or (13) in the s th iteration.

It has been shown heuristically that simultaneous recovery of the primary source (Eq. (13)) and the object function is possible from a single Fourier measurement and the object support constraint. The solution is found by solving the minimization problem,

$$\min_{\mathbf{g}, \mathbf{S}} \|\mathbf{I} - \mathbf{I}^{\text{C}} * \mathbf{S}\|_2^2 \quad (15)$$

where $\|\cdot\|_2^2$ represents the ℓ_2 norm of a vector, $\mathbf{g}$, $\mathbf{I}$, and $\mathbf{S}$ are vectors presenting the discretized object, intensity measurement, and source function, respectively. An alternating minimization strategy can be used that alternates between the two subproblems of coherent phase retrieval and source recovery in each iteration. The coherent phase retrieval can be implemented using any previous algorithm. The source is found by linear deconvolution assuming the coherent intensity is known. Since the full problem is nonlinear and nonconvex, the solution may still converge to a local minimum.

Three-Dimensional Recovery Problems

The 2D Fourier model in Eq. (1) relies on the assumption that the object is thin; the object function is determined by the projection approximation. When the object is thick and *weakly scattering*, the three-dimensional (3D) scattering process can be modeled using the first Born approximation, which accounts for only single scattering. The 3D CDI technique attempts to reconstruct the volumetric distribution of the object, which is characterized by the complex refractive index distribution $n(\mathbf{r}, z)$ as

$$g(\mathbf{r}, z) = \frac{\pi}{\lambda^2} (1 - n^2(\mathbf{r}, z)) \quad (16)$$

Denote the 3D FT of the object as $G(\mathbf{u}, w) = \mathcal{F}_{3D}\{g(\mathbf{r}, z)\}$, where w denotes the spatial frequency along the axis of illumination (z -axis). A single diffraction intensity $I(\mathbf{u})$ corresponds to a spherical cap on the Ewald sphere,

$$I(\mathbf{u}) = \left| G\left(\mathbf{u}, \sqrt{\frac{1}{\lambda^2} - |\mathbf{u}|^2} - \frac{1}{\lambda}\right) \right|^2 \quad (17)$$

To better cover the object's 3D Fourier space, diffraction data need to be collected at various orientations by either rotating the object or the illumination. In the reconstruction, alternating projection algorithms in Section Alternating Projection Algorithms have been extended to solve the 3D phase retrieval problem.

Defocus-Based Techniques

Defocus is another popular CDI method that is based on diffraction in free-space. The experimental procedure involves simply moving the camera (or the object) axially, while capturing multiple images through-focus, as illustrated in Fig. 3. The in-focus

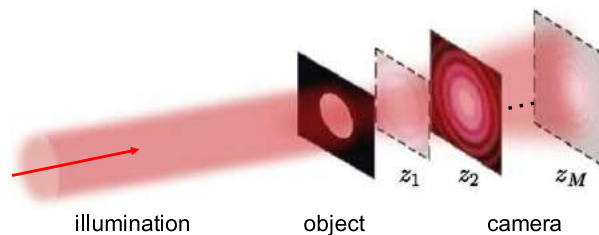


Fig. 3 Schematics of defocus-based coherent diffractive imaging (CDI) techniques.

intensity image contains no phase information. The defocused intensity is characterized by the depth (z) dependent free-space point-spread-function (PSF) $h(\mathbf{r}, z)$. The measured intensity at z under coherent illumination is

$$I(\mathbf{r}; z) = \left| \int g(\mathbf{r}') h(\mathbf{r} - \mathbf{r}', z) d^2 \mathbf{r}' \right|^2 = |\mathcal{P}_z\{g\}|^2 \quad (18)$$

where $\mathcal{P}_z\{\cdot\}$ denotes the propagator at defocus distance z .

Many phase retrieval algorithms have been demonstrated for defocus measurements. The validity of each method depends on many experimental parameters, such as the defocus range, the object property, and the illumination coherence.

Iterative Algorithms

To solve the nonlinear Eq. (18), one can employ iterative algorithms. When only a single defocused intensity is available, the technique is also known as Fresnel CDI, which is analogous to the Fourier technique. All the iterative algorithms discussed in Section Fourier Measurement-Based Techniques can be applied by simply replacing $\mathcal{F}\{\cdot\}$ and $\mathcal{F}^{-1}\{\cdot\}$ by the (Fresnel) forward-propagator $\mathcal{P}_z\{\cdot\}$ and back-propagator $\mathcal{P}_z^{-1}\{\cdot\} = \mathcal{P}_{-z}\{\cdot\}$, respectively. All the limitations in the Fourier techniques still apply – the algorithm may stagnant and not converge due to the nonlinearity and nonconvexity.

In many practical applications, it is possible to take a series of defocused intensity measurement, $I(\mathbf{r}; z_m)$, $m = 1, 2, \dots, M$. The most widely applied phase retrieval algorithm is the modified GS algorithm. In each iteration, one propagates the estimated field to each defocused planes, where the estimated amplitude is replaced by the measurement. To impose the amplitude constraints across multiple planes, two updating schemes have been implemented: sequential and global. The sequential GS algorithm imposes the amplitude constraint one plane at a time as the field propagates between the defocused and the object planes. The global GS algorithm instead updates the object field by averaging over estimates by imposing amplitude constraints from all defocused planes.

An alternative approach is based on solving the minimization of a cost function using the nonlinear optimization framework. For example, consider the amplitude-based least-square problem,

$$\min_{\mathbf{g}} E(\mathbf{g}) = \frac{1}{2} \sum_{m=1}^M \left\| \sqrt{\mathbf{I}_m} - |\mathbf{P}_{z_m} \mathbf{g}| \right\|_2^2 \quad (19)$$

where $\mathbf{g}$ and $\mathbf{I}_m$ are vectors presenting the discretized object function and intensity measurement at z_m , respectively. The discrete propagator $\mathbf{P}_{z_m}$ can be efficiently implemented using the fast Fourier transform (FFT), $\mathbf{P}_{z_m} = \mathbf{K}^H \mathbf{H}_{z_m} \mathbf{K}$, where $\mathbf{K}$ is the 2D discrete Fourier transform (DFT) matrix, $\mathbf{H}_{z_m} = \text{diag}(\exp[-i\pi\lambda z_m |\mathbf{u}|^2])$ is a diagonal matrix describing the Fresnel transfer function at the distance z_m , $\mathbf{u}$ is the discrete spatial frequency coordinates along the lateral dimensions, λ is the wavelength, $\text{diag}(\mathbf{a})$ denotes the diagonal matrix with its corresponding diagonal entries equal to the elements in the vector $\mathbf{a}$, and $\cdot^H$ denotes the matrix Hermitian. The back-propagation matrix $\mathbf{P}_{z_m}^{-1}$ satisfies $\mathbf{P}_{z_m}^{-1} = \mathbf{P}_{z_m}^H$, since $\mathbf{P}_{z_m}$ is unitary.

The reconstruction algorithm is then implemented following standard gradient-based optimization procedure. It starts with some initial guess of the object $\mathbf{g}_0$. At the s th iteration, the general equation for updating the object is

$$\mathbf{g}_{s+1} = \mathbf{g}_s + \alpha_s \mathbf{d}_s \quad (20)$$

where α_s is a step size, which can be a constant or found by a line-search algorithm. The search direction $\mathbf{d}_s$ is crucial for the algorithm convergence, which is determined by the gradient and/or second-order derivative (Hessian) of the cost function.

Both the gradient and Hessian can be conveniently calculated using $\mathbb{C}\mathbb{R}$ -calculus, in which the complex variable $\mathbf{g}$ is augmented as $[\mathbf{g}, \mathbf{g}^*]^T$ and $\bullet^*$ denotes the complex conjugate. The gradient of the cost function in Eq. (19) is

$$\nabla E(\mathbf{g}, \mathbf{g}^*) = \sum_{m=1}^M \begin{bmatrix} -\mathbf{P}_{z_m}^H \cdot \text{diag}(\mathbf{P}_{z_m} \mathbf{g} / |\mathbf{P}_{z_m} \mathbf{g}|) \\ -\mathbf{P}_{z_m}^T \cdot \text{diag}(\mathbf{P}_{z_m}^* \mathbf{g}^* / |\mathbf{P}_{z_m} \mathbf{g}|) \end{bmatrix} [\sqrt{\mathbf{I}_m} - |\mathbf{P}_{z_m} \mathbf{g}|] \quad (21)$$

where $\cdot^T$ denotes the matrix transpose.

In the first-order gradient descent (GD) methods, the search direction is simply the gradient of the cost function. In the second-order methods, the Hessian of the cost function $\nabla^2 E(\mathbf{g}, \mathbf{g}^*)$ is used to improve convergence speed. The search direction is then obtained by solving the linear equation

$$\nabla^2 E(\mathbf{g}, \mathbf{g}^*) \begin{bmatrix} \mathbf{d} \\ \mathbf{d}^* \end{bmatrix} = -\nabla E(\mathbf{g}, \mathbf{g}^*) \quad (22)$$

The Hessian matrix in practical phase retrieval problem is huge (size of N^2 , where N is the size of the object), so directly computing the inverse of Hessian is computationally prohibitive. Quasi-Newton's methods, such as the limited-memory Broyden–Fletcher–Goldfarb–Shanno method (L-BFGS) circumvents these problems by approximating the inverse of Hessian from the gradients of a few previous iterations. In modified Gauss–Newton methods, the inverse of Hessian is computed iteratively (e.g., using the conjugate gradient method) by solving the linear Eq. (22). Both the L-BFGS and modified Gauss–Newton methods have been applied to defocus phase retrieval problems and achieve better performance than the GS and GD algorithms.

Recursive Algorithms

The recursive algorithms, such as Kalman filter, attempts to progressively improve the phase estimate as each new intensity is included in the reconstruction. The intensity evolution upon propagation is treated with a complex augmented state space model. Denote the field at the m th plane and its FT as $\mathbf{a}_m$ as $\mathbf{b}_m$, respectively. Propagation of the intensity between consecutive defocus planes can be written in the Fourier domain as

$$\mathbf{b}_{m-1} = \mathbf{K}\mathbf{a}_{m-1} \quad (23)$$

$$\mathbf{b}_m = \mathbf{H}_m \mathbf{b}_{m-1} \quad (24)$$

where $\mathbf{H}_m$ is the matrix describing propagation by a distance $z_m - z_{m-1}$. The observation model assumes noisy intensity measurements,

$$\mathbf{I}_m = |\mathbf{a}_m|^2 + \mathbf{v}_m \quad (25)$$

where $\mathbf{v}_m$ assumes pixel-dependent gaussian noise with zero mean and a covariance matrix $\mathbf{R}_m \equiv \langle \mathbf{v}_m \mathbf{v}_m^T \rangle = \text{diag}(|\mathbf{a}_m|^2)$.

In the reconstruction, the Kalman filter phase retrieval (KF-PR) updates the estimates of both the field as well as its covariance matrix. The method is shown to be robust in the presence of significant noise corruption. The main computational requirement of KF-PR is the inversion of the covariance matrix. Sparsity priors in the covariance matrix have been exploited to significantly reduce the computational cost. Partial coherence KF-PR algorithm has also been demonstrated using the convolution model in Section Partial Coherence Algorithms.

Direct-Inversion Methods

When the object is weakly scattering in the sense of the first Born approximation being valid, and/or the defocus distance is small as compared to the Fresnel distance, the nonlinear measurement can be linearized, which allows implementations of direct-inversion algorithms. The transport of intensity equation (TIE) relates phase and amplitude to variations of intensity along the optical axis z ,

$$\frac{\partial I(\mathbf{r})}{\partial z} = -\frac{\lambda}{2\pi} \nabla_{\perp} \cdot [I(\mathbf{r}) \nabla_{\perp} \phi(\mathbf{r})] \quad (26)$$

where $I(\mathbf{r})$ is the intensity at focus, $\phi(\mathbf{r})$ is the phase, and $\nabla_{\perp}$ denotes the gradient operator in lateral directions $\mathbf{r}$ only. The problem is linear for small defocus distances.

Inversion algorithms for the TIE can be understood by considering the case of a pure-phase object, where $I(\mathbf{r})$ is constant. In this situation, the TIE simplifies to a Poisson equation, whose solution can be efficiently found by an FFT-based algorithm,

$$\phi(\mathbf{r}) = \mathcal{F}^{-1} \left\{ \mathcal{F} \left\{ \frac{2\pi}{\lambda I} \frac{\partial I(\mathbf{r})}{\partial z} \right\} / (4\pi^2 |\mathbf{u}|^2 + \delta) \right\} \quad (27)$$

The small regularization constant δ prevents instabilities around zero frequencies. When $I(\mathbf{r})$ is not constant, the inversion of the TIE requires solving two Poisson equations. The phase reconstruction algorithm does not require additional phase unwrapping step and is computationally efficient. In practice, one cannot measure intensity derivative directly and so it is usually estimated by finite difference methods. However, intensity is not linear upon propagation, which corrupts the derivative estimate. Higher-order TIE method has been demonstrated to mitigate this nonlinearity error by performing pixel-wise spatial domain polynomial fitting.

Contrast transfer function (CTF) is a linear model derived from the weak-object approximation,

$$g(\mathbf{r}) = \exp[-\alpha(\mathbf{r}) + i\phi(\mathbf{r})] \approx 1 - \alpha(\mathbf{r}) + i\phi(\mathbf{r}) \quad (28)$$

where $\alpha(\mathbf{r})$ characterizes the object's absorption. The FT of intensity at a distance z , $\mathcal{I}(\mathbf{u}; z)$ is linearly related to the FT of the phase $\Phi(\mathbf{u})$ and absorption $A(\mathbf{u})$ as

$$\mathcal{I}(\mathbf{u}; z) = \delta(\mathbf{u}) - 2A(\mathbf{u}) \cos(\pi\lambda|\mathbf{u}|^2 z) + 2\Phi(\mathbf{u}) \sin(\pi\lambda|\mathbf{u}|^2 z) \quad (29)$$

where $\delta(\mathbf{u})$ denotes the Dirac delta function.

The CTF suggests that the low-frequency phase information is only captured at large defocus distances. The intensity's FT follows a sinusoidal oscillation along the propagation direction with a frequency $\pi\lambda|\mathbf{u}|^2$, which increases as the spatial frequency becomes larger. The optimal design of defocus distances has shown to be exponentially spaced, which significantly reduces the number of images required to ensure high-quality phase recovery.

Phase reconstruction can be achieved by solving a linear deconvolution problem,

$$\min_{A, \Phi} \sum_m \sum_{\mathbf{u}} \left| \mathcal{I}(\mathbf{u}; z_m) - [\delta(\mathbf{u}) - 2A(\mathbf{u}) \cos(\pi\lambda|\mathbf{u}|^2 z_m) + 2\Phi(\mathbf{u}) \sin(\pi\lambda|\mathbf{u}|^2 z_m)] \right|^2 + \sum_{\mathbf{u}} [\alpha \|A(\mathbf{u})\|_2^2 + \beta \|\Phi(\mathbf{u})\|_2^2] \quad (30)$$

An alternative algorithm is based on pixel-wise spatial frequency domain curve-fitting to sinusoids. A Gaussian process regression-based method takes the CTF as the guideline to apply a thresholding operation in the curve-fitting process for suppressing the unwanted high frequency components in the fitted functions. The algorithm has been empirically shown to be robust even beyond the weak-object approximation, where the actual data deviates from the CTF model.

Transverse Translation-Based Techniques

The main limitation of the Fourier and defocus-based CDI techniques is the requirement of an isolated object. In the Fourier CDI technique, the size of the object cannot exceed the reciprocal of the detector's pixel size, due to the sampling requirement and Fourier transform relation between the object and image space. In the defocus based CDI technique, the object should be smaller than the detector. The transverse translation-based technique, commonly known as "ptychography," circumvents this problem by taking a series of Fourier diffraction images, while translating the illumination or the object, as illustrated in Fig. 4. A phase retrieval algorithm is then devised to *coherently* "stitch" all the intensity images from each sub field-of-view (FOV).

At the j th transverse position $\mathbf{r}_j$, the intensity $I_j(\mathbf{u})$ is simply the far-field diffraction of the exit wave from each sub-FOV, which can be modeled as the product of the object $g(\mathbf{r})$ with the shifted "probe" (illumination) function $p(\mathbf{r} - \mathbf{r}_j)$, assuming the object is thin,

$$I_j(\mathbf{u}) = \left| \int g(\mathbf{r}) p(\mathbf{r} - \mathbf{r}_j) \exp \{-i2\pi \mathbf{r} \cdot \mathbf{u}\} d^2 \mathbf{r} \right|^2 = |\mathcal{F}\{g(\mathbf{r}) p(\mathbf{r} - \mathbf{r}_j)\}|^2, \quad j = 1, 2, \dots, J \quad (31)$$

The phase retrieval algorithm reconstructs $g(\mathbf{r})$ from multiple intensity measurements at different probe positions. The success of this technique is highly dependent on the experimental parameters, such as the amount of overlap in the probe regions, the object property, and the illumination coherence. For example, a minimum of 60% of probe overlap has been reported both numerically and experimentally. It should be noted that even with the multiple intensity measurements, the underlying problem is still nonlinear and nonconvex, so the solution can get stuck in local minima. It has been shown empirically that transverse translation-based techniques significantly improve the convergence property, as compared to Fourier and defocused based techniques, because of the redundancy and diversity in the dataset.

Alternating Projection Algorithms

The most widely used ptychographic phase retrieval algorithm is based on the alternating projection framework. Similar to the modified GS algorithm in Section Iterative Algorithms, the multiple translational measurements can be processed using either the sequential or global scheme. The iterative algorithm allows simultaneous recovery of both the object and the probe functions.

The sequential algorithm is initialized with $g_0(\mathbf{r})$ and $p_0(\mathbf{r})$. In each iteration, it repeatedly applies two projection operations to each measurement one-by-one. In the j th subroutine within the s th iteration, it starts by calculating each exit field $\psi_{j,s}(\mathbf{r})$ from the j th sub-FOV,

$$\psi_{j,s}(\mathbf{r}) = g_{j,s}(\mathbf{r}) p_{j,s}(\mathbf{r} - \mathbf{r}_j) \quad (32)$$

The Fourier projection Π_F^S is applied to each exit field by imposing the amplitude constraint in two steps. It first replaces the Fourier magnitude by the measurement, while keeping the phase unchanged. An updated exit field, $\psi'_{j,s}(\mathbf{r})$ is then calculated by taking the inverse FT:

$$\Pi_F^S[\psi_{j,s}(\mathbf{r})] : \quad \Psi'_{j,s}(\mathbf{u}) = \mathcal{F}\{\psi_{j,s}(\mathbf{r})\} \cdot \frac{\sqrt{I_j(\mathbf{u})}}{|\mathcal{F}\{\psi_{j,s}(\mathbf{r})\}|} \quad (33)$$

$$\psi'_{j,s}(\mathbf{r}) = \mathcal{F}^{-1}\{\Psi'_{j,s}(\mathbf{u})\} \quad (34)$$

Next, the algorithm implements the overlap projection Π_O^S , which updates the object and probe functions:

$$\Pi_O^S[\psi_{j,s}(\mathbf{r})] : \quad g_{j+1,s}(\mathbf{r}) = g_{j,s}(\mathbf{r}) + \alpha \frac{p_{j,s}^*(\mathbf{r} - \mathbf{r}_j)}{|p_{j,s}(\mathbf{r})|_{\max}^2} [\psi'_{j,s}(\mathbf{r}) - \psi_{j,s}(\mathbf{r})] \quad (35)$$

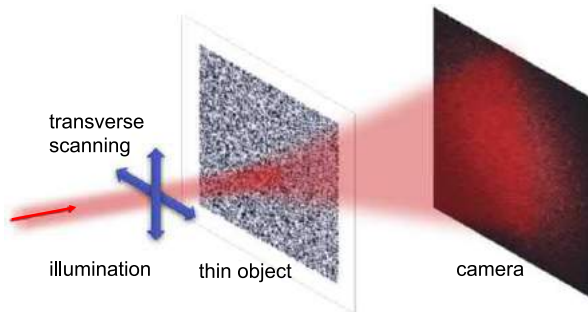


Fig. 4 Schematics of transverse translation-based coherent diffractive imaging (CDI) techniques.

$$p_{j+1,s}(\mathbf{r}) = p_{j,s}(\mathbf{r}) + \beta \frac{g_{j,s}^*(\mathbf{r} + \mathbf{r}_j)}{|g_{j,s}(\mathbf{r})|_{\max}^2} \left[\psi'_{j,s}(\mathbf{r} + \mathbf{r}_j) - \psi_{j,s}(\mathbf{r} + \mathbf{r}_j) \right] \quad (36)$$

where α and β are step sizes and are often set to 1. The results are then used as the input to the $(j+1)$ th subroutine until all the diffraction intensities have been used.

In the global alternating projection scheme, the Fourier projection Π_F^G is accomplished by imposing all the amplitude constraints simultaneously to all the exit wave $\psi_s(\mathbf{r}) = (\psi_{1,s}(\mathbf{r}), \psi_{2,s}(\mathbf{r}), \dots, \psi_{j,s}(\mathbf{r}))$:

$$\Pi_F^G[\psi_s(\mathbf{r})] = \left(\Pi_F^S[\psi_{1,s}(\mathbf{r})], \Pi_F^S[\psi_{2,s}(\mathbf{r})], \dots, \Pi_F^S[\psi_{j,s}(\mathbf{r})] \right) \quad (37)$$

The overlap projection, Π_O^G , updates the object and probe functions by taking weighted averages:

$$\Pi_O^G[\psi_s(\mathbf{r})] : g_{s+1}(\mathbf{r}) = \frac{\sum_{j=1}^J p_s^*(\mathbf{r} - \mathbf{r}_j) \psi'_{j,s}(\mathbf{r})}{\sum_{j=1}^J |p_s(\mathbf{r} - \mathbf{r}_j)|^2} \quad (38)$$

$$p_{s+1}(\mathbf{r}) = \frac{\sum_{j=1}^J g_s^*(\mathbf{r} + \mathbf{r}_j) \psi'_{j,s}(\mathbf{r} + \mathbf{r}_j)}{\sum_{j=1}^J |g_s(\mathbf{r} + \mathbf{r}_j)|^2} \quad (39)$$

Other alternating projection algorithms, as introduced in Section Alternating Projection Algorithms, can also be applied. For example, the DM algorithm has been widely used in ptychographic phase retrieval and takes the following procedure to update the exit field,

$$\psi_{s+1}(\mathbf{r}) = \psi_s(\mathbf{r}) + \Pi_F^G[2\Pi_O^G[\psi_s(\mathbf{r})] - \psi_s(\mathbf{r})] - \Pi_O^G[\psi_s(\mathbf{r})] \quad (40)$$

Mode Multiplexing Algorithms

The redundancy and diversity in the ptychographic dataset makes it possible to recover additional information besides the object and probe functions. Consider applying the mode decomposition model (Eq. (11)) to both the probe and the object, the diffraction pattern at the j th transverse position is,

$$I_j(\mathbf{u}) = \sum_{n=1}^N \sum_{m=1}^M \left| \int g_{(m)}(\mathbf{r}) p_{(n)}(\mathbf{r} - \mathbf{r}_j) \exp \{-i2\pi \mathbf{r} \cdot \mathbf{u}\} d^2 \mathbf{r} \right|^2 = \sum_{n=1}^N \sum_{m=1}^M |\mathcal{F}\{g_{(m)}(\mathbf{r}) p_{(n)}(\mathbf{r} - \mathbf{r}_j)\}|^2, \quad j = 1, 2, \dots, J \quad (41)$$

where $g_{(m)}(\mathbf{r})$ is the m th mode of the object and $p_{(n)}(\mathbf{r})$ the n th mode of the probe.

The alternating projection algorithms can be extended to the mode decomposition problem to simultaneously reconstruct the M mutually incoherent object modes and the N mutually incoherent probe modes. Each iteration starts by calculating the $M \times N$ coherent exit field from all possible pairs of the object and probe modes,

$$\psi_{(m,n)}(\mathbf{r}) = g_{(m)}(\mathbf{r}) p_{(n)}(\mathbf{r} - \mathbf{r}_j) \quad (42)$$

The Fourier and overlap projections are then applied similar to the coherent case. In the sequential scheme, the Fourier update Eq. (33) is replaced by

$$\Psi'_{(m,n)j,s}(\mathbf{u}) = \mathcal{F}\{\psi_{(m,n)j,s}(\mathbf{r})\} \cdot \frac{\sqrt{I_j(\mathbf{u})}}{\sqrt{\sum_{n=1}^N \sum_{m=1}^M |\mathcal{F}\{\psi_{(m,n)j,s}(\mathbf{r})\}|^2}} \quad (43)$$

which is applied to all the exit field. The rest of the steps follow the same operations as described in Eqs. (34–36), which are applied simultaneously to each exit wave.

The global scheme makes the same modifications to the Fourier projection as the sequential scheme. The overlap projection updates all the object and probe modes by taking the weighted average over all the modes and measurements,

$$g_{(m)s+1}(\mathbf{r}) = \frac{\sum_{j=1}^J \sum_{n=1}^N p_{(n)s}^*(\mathbf{r} - \mathbf{r}_j) \psi'_{(m,n)j,s}(\mathbf{r})}{\sum_{j=1}^J \sum_{n=1}^N |p_{(n)s}(\mathbf{r} - \mathbf{r}_j)|^2} \quad (44)$$

and

$$p_{(n)s+1}(\mathbf{r}) = \frac{\sum_{j=1}^J \sum_{m=1}^M g_{(m)s}^*(\mathbf{r} + \mathbf{r}_j) \psi'_{(m,n)j,s}(\mathbf{r} + \mathbf{r}_j)}{\sum_{j=1}^J \sum_{m=1}^M |g_{(m)s}(\mathbf{r} + \mathbf{r}_j)|^2} \quad (45)$$

The general mode decomposition model above has been applied to various situations. The probe mode decomposition model is useful to account for the effects of partially coherent illumination. A major limitation of ptychography is the long acquisition time required to collect all the images. “Fly-scan” ptychography attempts to improve the data collection speed by continuously scanning the probe during the exposure time. It has been shown that the motion blurs in fly-scan measurements can be accounted for using the probe mode decomposition model. Multi-spectral and polarization multiplexed ptychographic measurements have been attempted. Spectrally and polarization resolved reconstructions have been demonstrated by adapting the mode multiplexing algorithms to the problem of decomposing both the object and probe by their respective spectral and polarization channels.

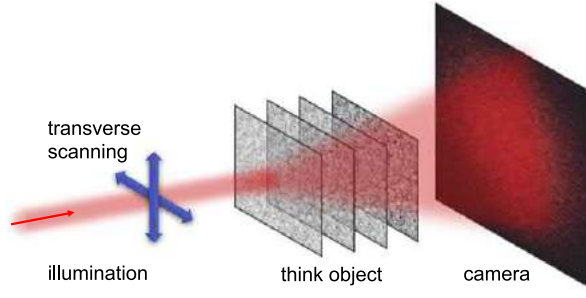


Fig. 5 Schematics of the multi-slice ptychography.

Multi-Slice Algorithms

The multiplicative model used in 2D ptychography breaks down when the object is thick and multiple scattering effects become nonnegligible. The multi-slice model is proposed to extend the depth of field and partially accounts for the forward multiple scattering effects.

A thick object is modeled as M thin slices, each described by a transmittance function $g^{(m)}(\mathbf{r})$ ($m = 1, 2, \dots, M$) (Fig. 5). The layer in-between is assumed uniform (e.g., air) of thickness Δz_m . The exit field from the object is modeled by a series of multiply-and-propagate operations,

$$\psi_j(\mathbf{r}) = \mathcal{P}_{\Delta z_{M-1}} \left\{ \dots \mathcal{P}_{\Delta z_2} \left\{ \mathcal{P}_{\Delta z_1} \left\{ p_j(\mathbf{r} - \mathbf{r}_j) g^{(1)}(\mathbf{r}) \right\} g^{(2)}(\mathbf{r}) \right\} \dots \right\} g^{(M)}(\mathbf{r}) \quad (46)$$

where the propagator can be computed using the angular spectrum method, $\mathcal{P}_{\Delta z}\{\cdot\} = \mathcal{F}^{-1} \left\{ \exp \left[i 2 \pi \Delta z \sqrt{1/\lambda^2 - |u|^2} \right] \mathcal{F}\{\cdot\} \right\}$. After passing through the object, the far-field diffraction intensity is measured,

$$I_j(\mathbf{u}) = |\mathcal{F}\{\psi_j(\mathbf{r})\}|^2 \quad (47)$$

The reconstruction algorithm can be implemented by extending the sequential 2D alternating projection algorithm to accounting for the multi-slice model. The j th subroutine within the s th iteration starts with the estimated j th exit field $\psi_{j,s}(\mathbf{r})$. Next, the Fourier projection is applied to update the exit field,

$$\psi'_{j,s}(\mathbf{r}) = \prod_{\mathbf{F}}^S [\psi_{j,s}(\mathbf{r})] \quad (48)$$

To update each object slice, the algorithm then recursively applies the “overlap-projection back-propagation” operation until the first slice is reached. The incident field $\phi_{j,s}^{(m)}(\mathbf{r})$ and the exit field $\psi_{j,s}^{(m)}(\mathbf{r})$ for the m th slice is related by

$$\psi_{j,s}^{(m)}(\mathbf{r}) = \phi_{j,s}^{(m)}(\mathbf{r}) g_{j,s}^{(m)}(\mathbf{r}) \quad (49)$$

Eq. (49) takes the same form as Eq. (32), except for the coordinate change. As a result, the overlap projection operation in 2D ptychography can be easily extended here. The modified overlap projection $\prod_{\mathbf{O}}^{S'}$ essentially separates the field scattered from the current slice from the field scattered from all previous slices,

$$\prod_{\mathbf{O}}^{S'} [\psi_{j,s}^{(m)}(\mathbf{r})] : g_{j+1,s}^{(m)}(\mathbf{r}) = g_{j,s}^{(m)}(\mathbf{r}) + \alpha \frac{\phi_{j,s}^{(m)*}(\mathbf{r})}{|\phi_{j,s}^{(m)}(\mathbf{r})|_{\max}^2} [\psi_{j,s}^{(m)'}(\mathbf{r}) - \psi_{j,s}^{(m)}(\mathbf{r})] \quad (50)$$

$$\phi_{j,s}^{(m)'}(\mathbf{r}) = \phi_{j,s}^{(m)}(\mathbf{r}) + \beta \frac{g_{j,s}^{(m)*}(\mathbf{r})}{|g_{j,s}^{(m)}(\mathbf{r})|_{\max}^2} [\psi_{j,s}^{(m)'}(\mathbf{r}) - \psi_{j,s}^{(m)}(\mathbf{r})] \quad (51)$$

The back-propagation operation is then applied to update the exit field of the $(m-1)$ th slice,

$$\psi_{j,s}^{(m-1)'}(\mathbf{r}) = \mathcal{P}_{\Delta z_{m-1}}^{-1} \left\{ \phi_{j,s}^{(m)'}(\mathbf{r}) \right\} \quad (52)$$

At the first slice, the overlap projection of 2D ptychography is applied to update the object slice and the probe functions: $\prod_{\mathbf{O}}^S [\psi_{j,s}^{(1)}(\mathbf{r})]$.

Illumination Angle Scanning-Based Fourier Ptychography

Fourier ptychography (FP) is a computational microscopy technique that achieves gigapixel imaging capability with both wide FOV and high resolution in both phase and amplitude. The working principle of FP can be understood as the implementation of the regular ptychography in the Fourier conjugate planes. The setup involves scanning the illumination across different angles, as illustrated in (Fig. 6).

Consider a thin sample with transmission function $g(\mathbf{r})$. The object is illuminated by a plane wave with the spatial frequency $\mathbf{u}_\ell$, defined by $\exp(i 2 \pi \mathbf{u}_\ell \cdot \mathbf{r})$. After passing through the object, the exit wave is the product of the object and the illumination,

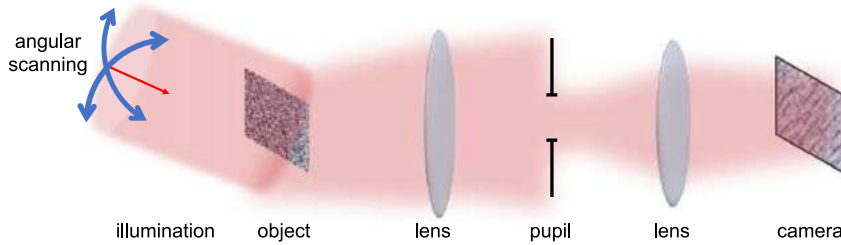


Fig. 6 Schematics of Fourier ptychography.

$g(\mathbf{r}) \exp(i2\pi\mathbf{u}_\ell \cdot \mathbf{r})$. In the Fourier space, this corresponds to a shifted version of the Fourier spectrum of the object, $G(\mathbf{u} - \mathbf{u}_\ell)$, where $G(\mathbf{u}) = \mathcal{F}\{g(\mathbf{r})\}$. This Fourier spectrum is then low-pass filtered by the pupil function, $P(\mathbf{u})$. The final intensity taken by the real-space camera is

$$I_\ell(\mathbf{r}) = |\mathcal{F}\{P(\mathbf{u})O(\mathbf{u} - \mathbf{u}_\ell)\}|^2 \quad (53)$$

The goal of the FP algorithm is to stitch together all the diffraction intensity corresponding to different areas of Fourier space in order to increase the bandwidth/resolution of the final image. The effective numerical aperture (NA) is the sum of the objective NA and illumination NA. Similar to the regular ptychography, phase is necessary to *coherently* combine all the intensity images. The first FPM algorithm adapted the sequential alternating projection algorithm in the regular ptychography. Shifted support constraints from the finite pupil size are enforced in the Fourier domain and the corresponding amplitude constraints (measured images) are applied in the real space, while letting the phase evolve as each image is stepped through sequentially.

Significant progress has been made to extend the various schemes in the regular ptychography to FP. The recovery of the aberration in the pupil function is demonstrated by adapting the probe recovery algorithm. FP mode multiplexing algorithms have been developed to allow reconstruction using much fewer images by simultaneously illuminating the object from multiple mutually incoherent illumination beams when capturing each image, as well as to recover spectrally multiplexed information. FP multi-slice algorithm has been developed to recover 3D object and improve spatial resolution in all three dimensions.

Further Reading

- Batey, D.J., Claus, D., Rodenburg, J.M., 2014. Information multiplexing in ptychography. *Ultramicroscopy* 138, 13–21.
- Chapman, H.N., Barty, A., Marchesini, S., *et al.*, 2006. High-resolution ab initio three-dimensional X-ray diffraction microscopy. *Journal of the Optical Society of America A* 23, 1179–1200.
- Clark, J.N., Huang, X., Harder, R., Robinson, I.K., 2012. High-resolution three-dimensional partially coherent diffraction imaging. *Nature Communications* 3, 993.
- Clark, J.N., Peele, A.G., 2011. Simultaneous sample and spatial coherence characterisation using diffractive imaging. *Applied Physics Letters* 99, 154103.
- Fienup, J.R., 1982. Phase retrieval algorithms: A comparison. *Applied Optics* 21, 2758–2769.
- Guigay, J.P., Langer, M., Boistel, R., Cloetens, P., 2007. Mixed transfer function and transport of intensity approach for phase retrieval in the Fresnel region. *Optics Letters* 32, 1617–1619.
- Jingshan, Z., Claus, R.A., Dauwels, J., Tian, L., Waller, L., 2014. Transport of intensity phase imaging by intensity spectrum fitting of exponentially spaced defocus planes. *Optics Express* 22, 10661–10674.
- Jingshan, Z., Tian, L., Dauwels, J., Waller, L., 2015. Partially coherent phase imaging with simultaneous source recovery. *Biomedical Optics Express* 6, 257–265.
- Maiden, A.M., Humphry, M.J., Rodenburg, J.M., 2012. Ptychographic transmission microscopy in three dimensions using a multi-slice approach. *Journal of the Optical Society of America A* 29, 1606–1614.
- Maiden, A.M., Rodenburg, J.M., 2009. An improved ptychographical phase retrieval algorithm for diffractive imaging. *Ultramicroscopy* 109, 1256–1262.
- Marchesini, S., 2007. Invited article: A unified evaluation of iterative projection algorithms for phase retrieval. *Review of Scientific Instruments* 78, 011301.
- Nugent, K., Gureyev, T., Cookson, D., Paganin, D., Barnea, Z., 1996. Quantitative phase imaging using hard X-rays. *Physical Review Letters* 77, 2961–2964.
- Paxman, R.G., Schulz, T.J., Fienup, J.R., 1992. Joint estimation of object and aberrations by using phase diversity. *Journal of the Optical Society of America A* 9, 1072–1085.
- Pelz, P.M., Guizar-Sicairos, M., Thibault, P., *et al.*, 2014. On-the-fly scans for X-ray ptychography. *Applied Physics Letters* 105, 251101.
- Shechtman, Y., Eldar, Y.C., Cohen, O., *et al.*, 2015. Phase retrieval with application to optical imaging: A contemporary overview. *IEEE Signal Processing magazine* 32, 87–109.
- Teague, M.R., 1983. Deterministic phase retrieval: A Green's function solution. *Journal of the Optical Society of America* 73, 1434–1441.
- Thibault, P., Dierolf, M., Bunk, O., Menzel, A., Pfeiffer, F., 2009. Probe retrieval in ptychographic coherent diffractive imaging. *Ultramicroscopy* 109, 338–343.
- Thibault, P., Menzel, A., 2013. Reconstructing state mixtures from diffraction measurements. *Nature* 494, 68–71.
- Tian, L., Li, X., Ramchandran, K., Waller, L., 2014. Multiplexed coded illumination for Fourier ptychography with an LED array microscope. *Biomedical Optics Express* 5, 2376–2389.
- Tian, L., Waller, L., 2015. 3D intensity and phase imaging from light field measurements in an LED array microscope. *Optica* 2, 104–111.
- Wade, R., 1992. A brief look at imaging and contrast transfer. *Ultramicroscopy* 46, 145–156.
- Waller, L., Tian, L., Barbastathis, G., 2010. Transport of intensity phase-amplitude imaging with higher order intensity derivatives. *Optics Express* 18, 12552–12561.
- Whitehead, L.W., Williams, G.J., Quiney, H.M., *et al.*, 2009. Diffractive imaging using partially coherent X-rays. *Physical Review Letters* 103, 243902.
- Yang, C., Qian, J., Schirotzek, A., Maia, F., Marchesini, S., 2011. Iterative algorithms for Ptychographic phase retrieval. *arXiv* 1105, 5628.
- Yeh, L.-H., Dong, J., Zhong, J., *et al.*, 2015. Experimental robustness of Fourier ptychography phase retrieval algorithms. *Optics Express* 23, 33214–33240.
- Zheng, G., Horstmeyer, R., Yang, C., 2013. Wide-field, high-resolution Fourier ptychographic microscopy. *Nature Photonics* 7, 739–745.
- Zhong, J., Tian, L., Varma, P., Waller, L., 2016. Nonlinear optimization algorithm for partially coherent phase retrieval and source recovery. *IEEE Transactions on Computational Imaging* 2, 310–322.

Phase Space and Imaging

Markus Testorf, Thayer School of Engineering, Hanover, NH, United States

© 2018 Elsevier Ltd. All rights reserved.

Introduction

Phase-space optics (Torre, 2005) represents a framework to model optical systems and signal propagation. It introduces concepts of joint time-frequency transforms, well established for signal processing, to the domain of ray and wave optics. In contrast to many joint transforms specifically designed for signal processing applications, phase-space optics preserves a physical interpretation of the joint space – spatial frequency representation. In close analogy to the phase space of classical and quantum mechanics, the physics of optical signal propagation is encoded in the associated phase-space representation.

Optical system analysis and design usually avoid rigorous modeling of Maxwell's equations. Instead, convenient approximations are in use, which are tailored to the typical constraints of optical systems. These include ray optics and radiometry, Gaussian beam optics, Fourier optics of scalar complex amplitudes, as well as coherence theory. Phase-space optics combines most aspects of the aforementioned into a single signal representation. This is particularly true for paraxial signals and systems, where formal relationships or even the equivalence of models can be established rather straightforwardly.

This article surveys the application of phase-space optics to optical imaging. While phase-space optics may be developed as a consistent replacement of more common models of light propagation, it is generally more useful to consider it as a complementary framework, which can be used to find solutions to specific imaging related problems more easily. This refers in particular to problems involving the analysis of the space-bandwidth requirements of signals and systems. For phase imaging phase-space optics provides a unique perspective, which has motivated new signal recovery methods, namely phase-space tomography. The phase-space interpretation provides a unique platform to compare imaging modalities and to establish performance limits. Furthermore, the phase space of quasi-geometrical optics establishes highly intuitive interpretations for optical problems otherwise burdened with the overhead of mathematical formalism.

Elements of Phase Space Optics

Optical phase space may be introduced conveniently as the configuration space of generalized ray coordinates. This means, in each plane along the optical axis z , any optical ray is characterized by its transverse location x and its angle of propagation θ relative to the optical axis. Each pair of generalized coordinates corresponds to a single point in phase space. It is often convenient to associate the propagation direction of a ray with the spatial frequency $v = \sin(\theta)/\lambda$ of a coherent monochromatic plane wave propagating in the same direction, thereby associating the spatial frequency effectively with a bundle of parallel rays.

Consequently, the phase-space distribution of rays $W(x, v)$ obeys the rules of paraxial ABCD matrix optics (Saleh and Teich, 1991). This is, for a paraxial optical system, the input ray coordinates (x_{in}, v_{in}) are related to the output coordinates (x_{out}, v_{out}) by the system's ray transfer matrix,

$$\begin{pmatrix} x_{out} \\ v_{out} \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x_{in} \\ v_{in} \end{pmatrix} \quad (1)$$

with the shape of the phase-space distribution transforming accordingly,

$$W_{out}(x, v) = W_{in}(Dx - Bv, -Cx + Av) \quad (2)$$

If the input and the output plane of the system are located in a medium of the same refractive index, the determinate of the ray-transfer matrix is always equal to one, i.e., only three of the four matrix coefficients are independent.

For three-dimensional geometries, both x and v are two-dimensional vectors. The phase-space distribution $W(x, v)$ may be recognized as the radiance, i.e., the power directed along v and radiating from, or being received by a point x .

For partially coherent optical signals the concept of radiance may be generalized and the phase-space distribution is defined as the Fourier transform of the cross spectral density $J(x_1, x_2)$ (see chapters 1, 2 and 7 of Testorf et al., 2009),

$$W(x, v) = \int_{-\infty}^{\infty} J(x + x'/2, x - x'/2) \exp(-i2\pi x'v) dx' \quad (3)$$

For the ideal case of incoherent signals, the phase space distribution is given by the irradiance distribution,

$$W(x, v) = I(x) \quad (4)$$

For coherent wavefronts, the mutual intensity factorizes and the phase-space distribution becomes the Wigner distribution function (WDF) of the complex amplitude $u(x)$,

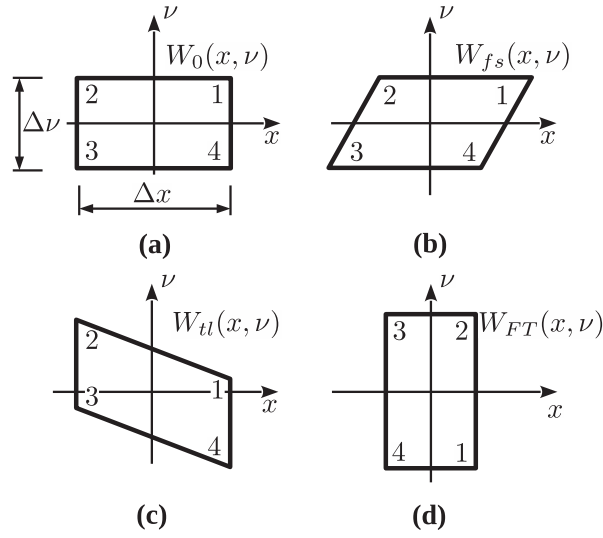


Fig. 1 Phase-space diagrams of ray transformations: (a) Phase-space diagram of a generic signal with essential support Δx and essential bandwidth $\Delta \nu$. (b) Phase-space distribution of the signal in (a) after Fresnel diffraction, (c) transmission through a thin paraxial lens, and (d) after Fourier transformation of the signal. Corners of the phase-space volume are numbered to depict the orientation of the transformed volume.

$$W(x, \nu) = \int_{-\infty}^{\infty} u\left(x + \frac{x'}{2}\right) u^*\left(x - \frac{x'}{2}\right) \exp(-i2\pi \nu x') dx' \quad (5)$$

As long as diffracting system apertures can be ignored, the ABCD transfer matrix equally applies to all definitions of $W(x, \nu)$. This implies, that the dynamics of coherent and partially coherent signals can be associated with geometrical transformations of phase space. For instance, free space propagation corresponds to a shear of phase space parallel to the spatial axis, and the transfer matrix of Fresnel diffraction reads

$$\mathbf{T}_f = \begin{pmatrix} 1 & \lambda z \\ 0 & 1 \end{pmatrix} \quad (6)$$

where z is the propagation distance. A thin parabolic lens of focal length f is given by

$$\mathbf{T}_l = \begin{pmatrix} 1 & 0 \\ -\frac{1}{\lambda f} & 1 \end{pmatrix} \quad (7)$$

which is equivalent to a chirp modulation of the complex amplitude.

A Fourier transform step is equivalent to a rotation of phase space by 90° , and consequently, a 2-f system corresponds to a combination of 90° rotation and a magnification step. Rotation with arbitrary angles correspond to fractional Fourier transforms of the signal (Ozaktas *et al.*, 2001).

Fig. 1 depicts a set of the generic phase-space transformations in the form of phase-space diagrams. Signals are schematically represented by a finite phase-space volume with extensions equivalent to the essential signal support and bandwidth. This identifies the phase-space volume as a measure of the space-bandwidth product of the signal.

In the context of optical imaging it is important to highlight a number of properties of the WDF. The WDF is a real valued function. It is symmetric with respect to the two variables x and ν , and we find the alternative definition,

$$W(x, \nu) = \int_{-\infty}^{\infty} \tilde{u}\left(\nu + \frac{\nu'}{2}\right) \tilde{u}^*\left(\nu - \frac{\nu'}{2}\right) \exp(i2\pi \nu' x) d\nu' \quad (8)$$

with $\tilde{u}(\nu)$ being the Fourier transform of the signal

$$\tilde{u}(\nu) = \int_{-\infty}^{\infty} u(x) \exp(-i2\pi \nu x) dx \quad (9)$$

was used. The WDF is a complete signal representation in so far as the complex amplitude can be recovered from the WDF except for a constant phase factor. Intensity and power spectrum of the complex signal are calculated as projections of the WDF,

$$\int_{-\infty}^{\infty} W(x, \nu) d\nu = |u(x)|^2 \quad \text{and} \quad \int_{-\infty}^{\infty} W(x, \nu) dx = |\tilde{u}(\nu)|^2 \quad (10)$$

The WDF is a bilinear transformation of $u(x)$. This means, the sum of two signals will be transformed into the respective WDFs, with additional interference or cross terms added. For instance, a signal consisting of two plane waves, $u_{1,2}(x) = \exp(i2\pi v_1 x) + \exp(i2\pi v_2 x)$, is transformed into

$$W_{1,2}(x, v) = \delta(v - v_1) + \delta(v - v_2) + 2\cos[2\pi(v_1 - v_2)]\delta\left(v - \frac{v_1 + v_2}{2}\right) \quad (11)$$

The WDF contains a frequency, which is located halfway between the two plane wave frequencies, and which is modulated by a cosine function. The modulation frequency of the interference term reflects the spacing between the unmodulated frequency terms. This interference term effectively encodes the mutual coherence of the two plane waves.

The convolution of two signals,

$$u_c(x) = u_1(x) * u_2(x) \quad (12)$$

is transformed into a convolution of the respective WDFs with respect to x and a multiplication with respect to v ,

$$W_c(x, v) = W_1(x, v) *_x W_2(x, v) \quad (13)$$

Similarly, a multiplication of two signals, i.e., a convolution in the frequency domain, is transformed into a convolution of the WDFs with respect to v ,

$$W_M(x, v) = W_1(x, v) *_v W_2(x, v) \quad (14)$$

For slowly varying phase functions, $u(x) = \exp(i\phi(x))$, the WDF can be approximated by the instantaneous (or local) frequency of the signal,

$$W_u(x, v) \approx \delta\left(v - \frac{1}{2\pi} \frac{d\phi(x)}{dx}\right) \quad (15)$$

which may be interpreted as the quasi-geometrical approximation of WDF. For quadratic phase functions, Eq. (15) is the exact expression for the WDF, which is equivalent to an inclined line in phase space.

The WDF is related to the ambiguity function (AF) (chapter 2 of Testorf *et al.*, 2009) via a double Fourier transformation,

$$A(\bar{v}, \bar{x}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W(x, v) \exp[-i2\pi(\bar{v}x - \bar{x}v)] dx dv \quad (16)$$

Many of the properties of the WDF carry over to the AF domain, namely the ray matrix transformation to relate input and output AF for quadratic phase systems.

Phase-Space Interpretation of Image Formation

In classical geometrical optics, perfect imaging is accomplished, if all rays originating from one point in the object plane intersect in a single point in the imaging plane. The spatial coordinates of object and image points are related by a constant magnification factor, i.e.,

$$x_{out} = m x_{in} \quad (17)$$

Negative magnification refers to an inverted image.

In phase space, the set of all rays crossing through a single point in space define a line parallel to the frequency axis (Fig. 2). This implies that perfect imagery is accomplished with optical systems, which map vertical lines in phase space to vertical lines with a constant scaling factor in x .

This implies a ray transfer matrix with $A=m$ and $B=0$ and $D=1/m$. The forth parameter C can be chosen freely to satisfy the imaging condition. This matrix can be interpreted as a magnification step followed by a chirp modulation, i.e.,

$$\begin{pmatrix} m & 0 \\ C & \frac{1}{m} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \frac{C}{m} & 1 \end{pmatrix} \begin{pmatrix} m & 0 \\ 0 & \frac{1}{m} \end{pmatrix} \quad (18)$$

While coefficient $A=m$ represents the spatial magnification factor of the imaging system, $D=1/m$ is the angular magnification factor.

This is exemplified by a single lens system, consisting of free space propagation over a distance a , a lens of focal length f , followed by propagation over a distance b , with the additional condition $1/f = 1/a + 1/b$. The associated transfer matrix reads

$$\mathbf{T}_{SLI} = \begin{pmatrix} 1 & \lambda b \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\frac{1}{\lambda f} & 1 \end{pmatrix} \begin{pmatrix} 1 & \lambda a \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} -\frac{b}{a} & 0 \\ -\frac{1}{\lambda f} & -\frac{a}{b} \end{pmatrix} \quad (19)$$

The case $C=0$ describes an afocal imaging system. A parallel bundle of input rays is transformed into a parallel bundle of output rays. The 4-f system and the closely related Kepler telescope are particular examples of afocal systems.

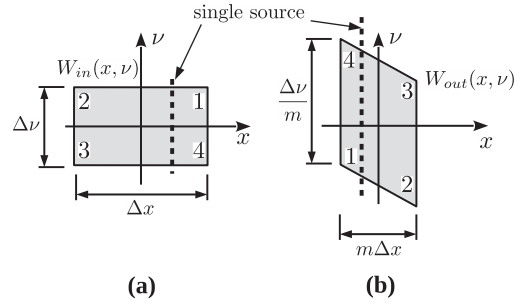


Fig. 2 Single lens imaging: (a) Phase-space diagram of input signal; (b) phase-space distribution in the image plans.

For $C \neq 0$ the chirp modulation corresponds to a shear of the phase-space distribution along the frequency axis (**Fig. 2(b)**). For the ideal case of a strictly incoherent object, with $W(x, \nu) = I(x)$, the phase space distribution is shift-invariant with respect to frequency and the vertical shear is not encoded in the phase space of the image plane. For the more general case of partially coherent or coherent signals, the shear of the phase space distribution corresponds to a modulation of the wavefront with a quadratic phase function.

For a wave optical analysis and finite lens apertures, the WDF of the aperture is convolved in ν with the signal's WDF in the aperture plane and further propagated to the output plane.

Imaging Aberrations

The close relationship between geometrical optics and phase space extends beyond the paraxial regime, and the phase-space formalism inherently is an integral part of geometrical optics theory (Wolf, 2004). This includes the analysis of optical aberrations.

One approach to understand imaging aberrations in the context of optical phase space is the interpretation of geometrical aberrations as deviations from the linear transformation, which connects input and output of the imaging systems. This means, the generalized coordinates of the output ray are no longer linear functions of the input coordinates expressed by a ray transfer matrix with constant coefficients. Instead, the matrix coefficients are functions of the input ray coordinates and the phase space distribution is distorted in both x and ν coordinate. As a consequence, the image, i.e., the projection of the phase space distribution along ν in the output plane of the system, is not longer a perfect copy of the object.

This nonlinear geometrical transform of phase space can be reversed, if the phase space distribution is known. This forms the basis for aberration correction with lightfield camera systems (see Section Lightfield Imaging).

The interpretation of aberrations as a geometrical distortion of phase space can be translated to the analysis of wavefront aberrations and their influence on the WDF of the optical signal. Wave aberrations may be interpreted as a phase-only modulation of the wavefront in the exit pupil of the imaging system. The phase $\phi(x)$ can be expressed as a power series and for a smooth aberration function the WDF of the modulating function can be approximated by the local frequency (Lohmann et al., 1983),

$$W_{ab}(x, \nu) \approx \delta \left[\nu - \frac{1}{2\pi} \frac{\phi(x)}{dx} \right] \quad (20)$$

The aberrated phase space distribution is the result of a convolution of $W_{ab}(x, \nu)$ with the WDF of the ideal wavefront in ν in the plane of the exit pupil. This is identical with a local shift of the WDF along the frequency axis and proportional to the local frequency.

In the image plane, the WDF of the exit pupil appears essentially rotated due to the inherent Fourier relationship between exit pupil and image plane, and the aberrations are manifest as frequency dependent shifts along the x axis. Consequently, image points derived as the projection of the output WDF are blurred compared to the non-aberrated image point.

The Self-Imaging Condition

The phase space analysis of classical geometric optical imaging suggests generalizations imaging. The point-to-point mapping of optical rays and the associated mapping of vertical lines in phase space implies a mapping of the irradiance distribution, i.e., the imaging system is designed to reproduce the vertical projection of phase space. For incoherent imaging this reproduces all information available about the object.

For imaging applications based on coherent optical signals, it is equally possible to associate imaging with a recovery of the complex amplitude, equivalent to a full recovery of the object's phase space distribution. Furthermore, neither the recovery of the actual phase space distribution, nor its projection needs to be the result of a spatial point-to-point mapping of the signal. Instead, each image point may be assembled as an interference pattern of signals originating from more than one single object point.

The latter cannot be ensured to work for arbitrary objects, but rather limits imaging for a particular system to a subset of all possible objects.

One example of this type of imaging is the Talbot effect, also known as coherent self-imaging. In its most generic form, self-imaging is associated with Fresnel diffraction of periodic object waves. Images, i.e., replica of the complex amplitude of the object are observed at multiples of the Talbot distance,

$$z_T = \frac{2p^2}{\lambda} \quad (21)$$

where p is the transverse period of the object wave.

In phase space the coherent periodic signal corresponds to a WDF, which is periodic in x and discrete in frequency ν (Fig. 3(a)). In particular, for signals of the form,

$$u(x) = u(x - p) = \sum_{n=-\infty}^{\infty} a_n \exp(i2\pi x n/p) \quad (22)$$

the WDF reads,

$$W(x, \nu) = W(x - p, \nu) = \sum_{n, n'=-\infty}^{\infty} a_n a_{n'}^* \exp(i2\pi x(n - n')/p) \delta\left(\nu - \frac{n + n'}{2p}\right) \quad (23)$$

Due to cross-terms the spacing of the discrete frequencies is $\Delta\nu = 1/(2p)$.

Fresnel diffraction corresponds to a linear shear (Fig. 3(b)). The sheared WDF is identical to the initial distribution, when the lowest non-zero frequency component is shifted by p , or

$$\lambda z_T \frac{1}{2p} = p \quad (24)$$

which effectively provides a geometrical derivation of the Talbot distance (see chapter 10 of Testorf *et al.*, 2009).

The self-imaging principle can easily be generalized to account for image magnification, and other optical systems, which restore the WDF of particular types of signals (Testorf and Hennelly, 2016). Most interestingly, the phase-space interpretation highlights the duality between operation in the spatial and the spatial frequency domain. This means, for instance, the Talbot effect has a dual frequency counterpart, which is equivalent to modulating a discrete signal with a quadratic phase function. In phase-space, space and spatial frequency axis are merely two particular orientations and self-imaging can be further generalized to include discrete chips which are sheared along their axis of orientation in phase space.

Holographic Imaging

The phase-space formulation of holographic imaging provides an intuitive interpretation of critical properties, namely the space-bandwidth product requirements of the recording medium and conditions for a unique recovery of the wavefront during the reconstruction step (Lohmann *et al.*, 2004).

The phase-space analysis of the holographic principle can be demonstrated by assuming a wavefront in the recording plane composed of the complex amplitude of the object wave $u(x)$ and a plane off-axis reference wave. Fig. 4(a) schematically depicts the object wave O and the off-axis plane wave R in the recording plane. The reference wave corresponds to a line parallel to the x axis. The extensions in space Δx and frequency $\Delta\nu$ define the space-bandwidth product (SBP) of the object wave. The spatial frequency of the reference wave is ν_R . The phase-space diagram in Fig. 4 does not contain any cross-terms, since they are inconsequential for the interpretation of the holographic principle that follows.

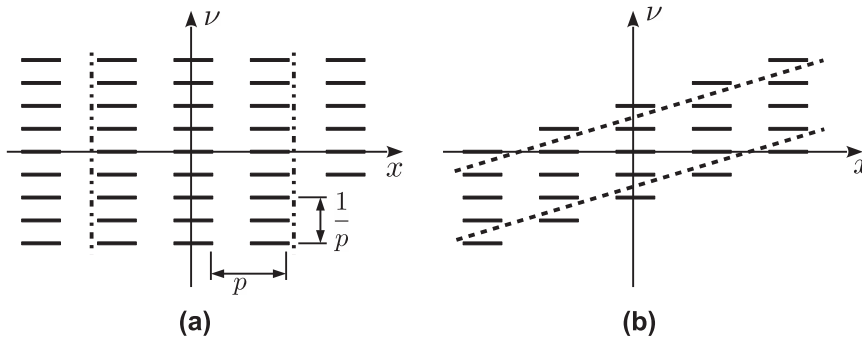


Fig. 3 Selfimaging in phase-space: (a) Phase-space diagram of periodic signal. The diagram depicts a finite number of periods and harmonics only. The vertical dashed lines mark arbitrary locations x ; (b) phase-space distribution after Fresnel diffraction over a single Talbot length.

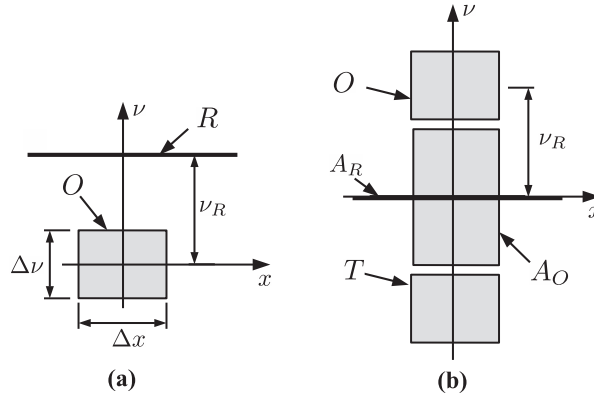


Fig. 4 Holography in phase space: (a) Phase-space diagram of the signal O and the plane reference wave R . (b) Phase-space distribution after the recording. The hologram is composed of the object O , the autocorrelation of object A_O and reference A_R , as well as the twin image T .

The holographic recording stores the squared magnitude of the complex amplitude, i.e., the transmission function of the recorded hologram reads

$$t(x) = |u(x) + \exp(i2\pi v_R x)|^2 = (1 + |u(x)|^2) + u^*(x)\exp(i2\pi v_R x) + u(x)\exp(-i2\pi v_R x) \quad (25)$$

with last term containing the object wave.

In phase space, the squared modulus of a signal is translated into the autocorrelation function of the WDF along v . The autocorrelation of object wave A_O and reference A_R are centered at the origin in phase space (Fig. 4(b)). Apart from the object wave O it is possible to observe the twin image T . All terms occupy separate volumes and do not overlap. This is necessary to recover and isolated the object term by illuminating the hologram with the reference wave. Estimating the bandwidth of the autocorrelation term to be twice the bandwidth of the object, the minimum frequency of the reference wave has to be

$$v_R \geq \frac{3}{2} \Delta v \quad (26)$$

The bandwidth of the hologram equals $2v_r + \Delta v$, which is at least four times the bandwidth of the hologram.

Fig. 4(b) can be interpreted as the phase-space diagram of image plane holography or Fourier plane holography. For the latter case, object wave and reference are recorded after a Fourier transform step equivalent to a 90° rotation of phase space. For Fresnel plane holography, the object wave is propagated in free space before it is recorded. The corresponding phase space operation is a horizontal shear, and it is possible to derive the SBP requirements of the recording medium from a geometric interpretation in phase space (Lohmann *et al.*, 2004).

Instead of analysing holography exclusively in the WDF domain, it is possible to use the AF (Situ and Sheridan, 2007). The AF of the complex amplitude may be interpreted as a generalized autocorrelation function. This means, the phase-space diagram of the AF of the complex amplitude resembles the diagram of the WDF of the recorded signal in Fig. 4(b). The actual transmission function of the hologram can be identified as the Fourier transform of the cross-section of the AF along its frequency axis.

Lightfield Imaging

The lightfield is closely related to the phase space of geometrical optics. Instead of parameterizing rays in each plane along the optical axis by their transverse location and propagation direction, the lightfield defines each ray by its spatial coordinates in two planes. While encoding the ray parameters in a different way, the lightfield and geometrical optical phase space contain the same information about the optical signal. Lightfield imaging is aimed at sampling the lightfield (or phase space) of the signal directly, rather than its projections.

This may be accomplished with a lightfield (or plenoptic) camera. The most generic implementation of the plenoptic camera is based on Lippmann's concept of integral photography. The detector array in the image plane of a camera system is replaced with a lenslet array and the detector, now located in the focal plane of the lens array (Fig. 5(a)). The aperture Δx of each lenslet defines the spatial resolution of the image as well as the spatial resolution with which phase space is sampled. The propagation direction of the ray is encoded as lateral displacement in the detector plane. Consequently, with the framework of geometrical optics, the angular resolution is determined by the pixel size Δx_θ , with an angular resolution $\Delta\theta \approx \Delta x_\theta / f$ and the resolution in frequency estimated as $\Delta v \approx \Delta x_\theta / (\lambda f)$.

This means, each detector pixel corresponds to a finite area in phase-space. The center of this finite volume in phase space indicates the position of the microlens. Each microlens projects the signal onto a subset of detector elements each in turn encoding a location along the frequency direction.

The phase-space diagram of the focal plane array is rendered in Fig. 5(b). Each ellipse schematically depicts the footprint of a single detector element. The recorded signal is the average of the signal's phase-space distribution over the volume covered by a

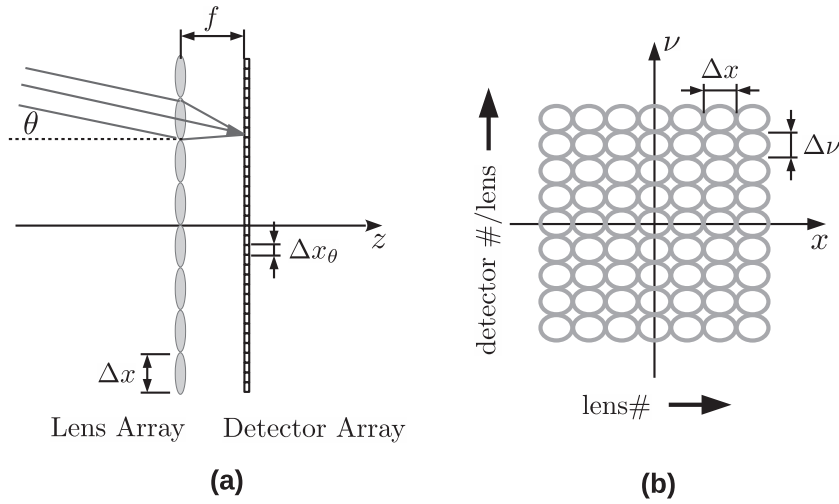


Fig. 5 Lightfield detector: (a) a focal plane array can be used to measure the radiance distribution. (b) The phase-space diagram of the focal plane array. Each ellipse corresponds to the phase-space volume recorded by a single detector element.

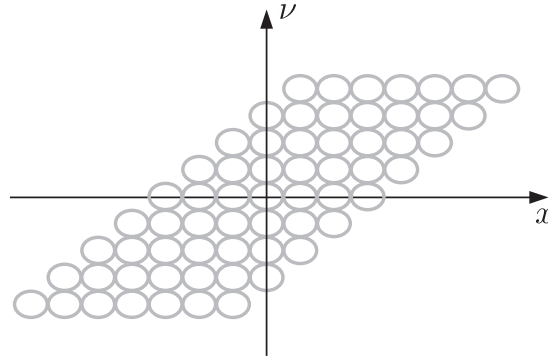


Fig. 6 Phase-space diagram of the focal plane array for correcting defocus: Each detector in phase space is associated with the location the in-focus phase-space distribution would assume in an out-of-focus plane. The in-focus image can be recovered by numerically reassigning the position of the detected phase-space locations before rendering the image from the frequency projection.

detector element. This means, as long as the phase-space distribution changes slowly on the scale of the detector's phase-space coverage, the focal plane array can be used as a phase-space detector without loss of important information. The number of detectors is equivalent to the SBP of the camera, i.e., comparing this with the SBP of the signal allows an estimation of how well the lightfield camera is recording the signal information.

Direct access to the phase-space distribution allows one to transform phase space before rendering an image by projecting along the frequency axis. This permits a change of perspective, refocusing, as well as the compensation of aberrations after recording a single lightfield. For instance, to correct defocus the signal associated with a detector is assigned to a point in phase space, which reflects the geometrical transform of phase space associated with free space propagation (**Fig. 6**). The defocus blur can be removed by reassigning the phase-space coordinates to the locations of the in-focus plane in **Fig. 5(b)**, which is equivalent to rendering the image from projections at the shear angle.

The lightfield concept can also be used to design three-dimensional displays for autostereoscopic viewing. This can be understood as reversing the recording process, i.e., each pixel of the detector is replaced with a point source. This creates a virtual image, which is viewed through the lenslet array. Each of the point sources encodes one sample of the radiance distribution of the reconstructed optical signal.

The lightfield inherently assumes a geometrical optics model of imaging with a phase-space distribution that is changing slowly. In this context, the focal plane array provides sufficient information to determine the radiance distribution from the recorded samples. If diffraction effects have to be accounted for the finite phase-space volume of a single detector does not provide sufficient resolution of phase space and it is necessary to recover the actual WDF of the optical signal. In general, it is necessary to resort to a synthesis of the phase-space distribution from information acquired with multiple recordings in different diffraction planes.

Phase-Space Tomography and Phase Retrieval

Every measurement in the optical regime can be interpreted as a power reading written as a projection

$$P = |\langle u(x), h(x) \rangle|^2 \quad (27)$$

where the complex amplitude $u(x)$ is detected with an optical instrument $h(x)$. This formalism, for instance, describes the coupling of the coherent wavefront $u(x)$ incident on an optical fiber, with $h(x)$ describing one particular optical mode. Eq. (27) then refers to the power P coupled into the mode and transported to a square law detector. In phase space, this overlap integral becomes

$$P = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W_u(x, v) W_h(x, v) dx dv \quad (28)$$

where $W_u(x, v)$ and $W_h(x, v)$ are the normalized WDFs of the signal and the instrument, respectively.

The ideal generic detector may be modeled as a point detector, located at x_d , which receives optical power independent of the incident angle,

$$W_{pd}(x, v) = \delta(x - x_d) \quad (29)$$

The detected signal is the overlap integral,

$$P = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W_u(x, v) W_{pd}(x, v) dx dv = \int_{-\infty}^{\infty} W_u(x_d, v) dv \quad (30)$$

which is consistent with the projection theorem of the WDF in Eq. (10). The signal power as a function of x is obtained with a continuum of point detectors.

This also means, the focal plane array of the lightfield camera effectively provides samples of a low pass filtered WDF only. It is possible, however, to add additional measurements with different instruments $W_h(x, v)$ and recover the signal WDF from the set of all measurements.

In particular, it is possible to envision power recordings obtained after passing the wavefront through a set of fractional Fourier transform systems. The fractional Fourier transformation generalizes the Fourier transforms in that it implements rotations of phase space at arbitrary angles. The rotation angle is proportional to the order of the fractional Fourier transformation.

Measuring power after fractional Fourier transforming the signal corresponds to the recording of projections of the WDF of the original signal along the rotation angle of the fractional Fourier transform. This means, a set of measurements covering all projection angles can be used for tomographic reconstruction of the WDF. Invoking the projections slice theorem of tomography identifies the projections at each angle as one slice of the AF of the signal. From knowledge of the AF the WDF and the signal can be estimated via an inverse Fourier transform.

Fractional Fourier transforms are not necessarily required to obtain the projections for tomographic reconstruction. For instance, Fresnel diffraction patterns may be recorded instead. Fig. 7 demonstrates, how projections along vertical lines $\overline{P_1 P_2}$ of the rotated WDF can be found after a horizontal shear, equivalent to free space propagation. While mapping recordings in Fresnel diffraction planes requires numerical scaling of the recording and presents inherent difficulties to obtain the equivalent of rotation angles close to 90° , it is typically easier to move the detector array along the optical axis, rather than implementing a large set of fraction Fourier transform systems.

This approach, known as phase-space tomography (Smithey *et al.*, 1993), is remarkable in that it provides a linear inverse algorithm for phase retrieval. This advantage is partially counterbalanced by the number of measurements required for the reconstruction of the WDF. For coherent wavefronts, a signal which is represented by N samples, is transformed into nominally $4N^2$ samples of the WDF, which includes the interference terms. Without additional prior knowledge, the linear retrieval then requires the same order of measurements to uniquely recover the WDF.

For partially coherent optical signals the number of samples to represent the phase-space distribution is equivalent to the number of sampling points to describe the mutual intensity. This means, in this case, phase-space tomography fully represents the complexity of the inverse problem.

For the recovery of two-dimensional images the projection slice theorem calls for anamorphic fractional Fourier transform systems, where the order of the fractional Fourier transform system can be changed independently for the two transverse coordinates. This means, a four-dimensional phase space has to be recovered to determine a two dimensional image. For coherent imaging, however, it is possible to reduce the experimental burden and recover the 2-D image as a set of 1-D slices.

The phase-space interpretation of tomographic signal recovery can be extended to other forms of phase-retrieval, most notably deterministic phase retrieval methods (Semichaevsky and Testorf, 2004) and pytochraphic imaging (Lubk and Röder, 2015). For coherent signals, it is possible to reduce the number of power measurements by surrendering the linearity of the reconstruction algorithm. For instance, deterministic phase retrieval from two closely located Fresnel diffraction planes are inherently based on the recovery the WDF from its local first order moments. This provides access to the AF and it first derivative along the frequency axis in turn providing magnitude and phase of the signal. The phase-space interpretation then provides a way to compare underlying principles of various phase retrieval methods.

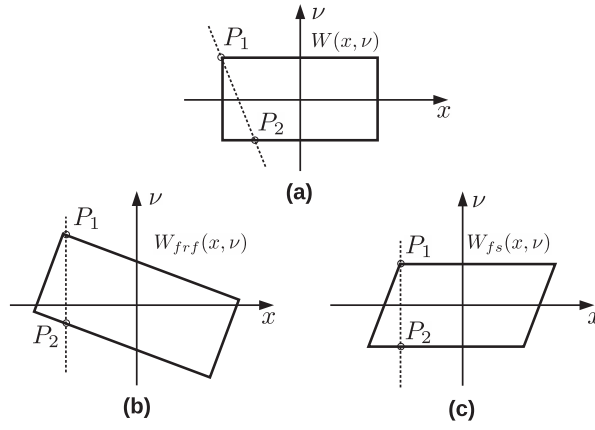


Fig. 7 Equivalence of Fresnel diffraction and fractional Fourier transform. (a) Phase-space diagram of the signal. (b) Signal after fractional Fourier transform. The line $\overline{P_1 P_2}$ describes the projection associated with the signal detected in a fraction Fourier plane. (c) The phase-space after Fresnel diffraction. The horizontal shear can be selected to provide the same vertical projections as observed for the rotated phase-space distribution.

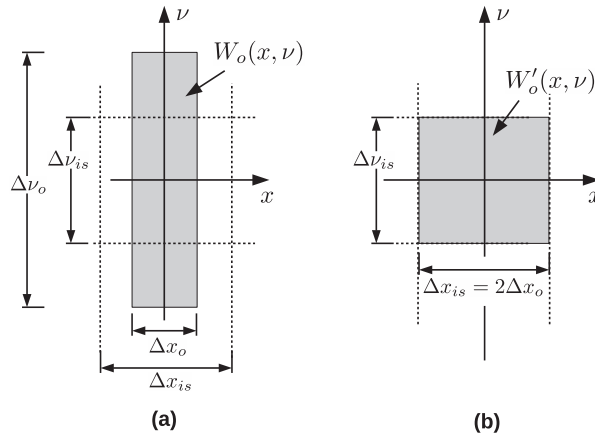


Fig. 8 Phase-space diagram of superresolution imaging: (a) The object bandwidth $\Delta\nu_o$ is twice the bandwidth supported by the optical system $\Delta\nu_{is}$. (b) Adapting the phase-space volume to the constraints of the imaging system allows transmission through the optical system without loss of information.

Superresolution Imaging

Superresolution imaging describes mathematical algorithms and imaging systems, which provide optical resolution superior to the nominal value associated with the size of the clear aperture of the system (Zalevsky and Mendlovic, 2004).

The mechanisms to achieve superresolution may be understood as a consequence of Shannon's information capacity theorem. Interpreting the imaging system as a continuous Gaussian channel of bandwidth $\Delta\nu_{is}$, field of view Δx_{is} and signal-to-noise ratio SNR, we expect to resolve

$$C = \Delta x_{is} \Delta \nu_{is} \log_2(1 + \text{SNR}) \quad (31)$$

bits of information from the image. If input signals to the system do not occupy the entire information capacity in terms of spatial support, or signal-to-noise ratio, it is possible to reassign information, which is encoded by the signal bandwidth, to either the signal extension or the SNR and transmit signals with a bandwidth exceeding the supported bandwidth of the system.

The phase-space interpretation is useful, in particular, to explore the trade-off between signal extension and bandwidth (Zalevsky et al., 2000). The generic operation of the this type of superresolution system is shown in Fig. 8. The input signal has bandwidth $\Delta\nu_o$ and a lateral extension Δx_o . The imaging systems only supports half the signal's bandwidth, i.e. $\Delta\nu_{is} = \Delta\nu_o/2$, but the field of view Δx_{is} is twice the object extension. Then, a coding step can be implemented to transform the signal into the transmitted signal with half the original bandwidth, but twice the original signal extension. This type of bandwidth compression may be accomplished with quadratic phase systems which leave the phase-space volume in tact.

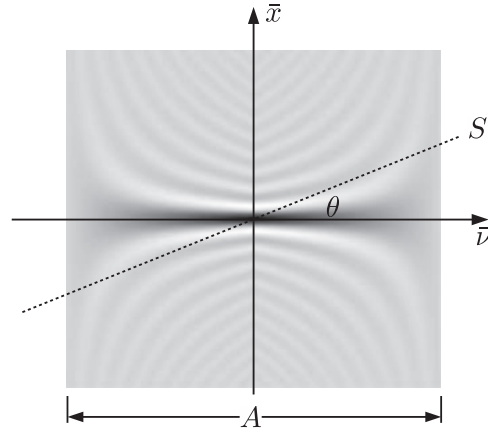


Fig. 9 Ambiguity function of the diffraction limited rectangular aperture: Cross-sections C correspond to the OTF in out-of-focus planes. The angle θ is related to the degree of defocus.

After passing the modified signal through the imaging system a decoding step is required to restore the original signal. Since the bandwidth of the transmitted signal, and consequently its resolution, is twice the nominal bandwidth of the system, this type of system effectively achieves superresolution imaging.

In order to recover a superresolved image of the object, the phase-space distribution at the output plane should not overlap with any other signal component.

Coding and decoding step may fundamentally differ depending on the imaging application. For instance, structured illumination sources may be used to move high frequencies of the object to the transmitted baseband, while numerical computation is used to assemble the image.

Pupil Design for Extended Depth of Field

The phase-space interpretation of wave aberrations can be combined with the concept of phase-space tomography to design imaging systems with increased depth of focus (see chapter 5 of Testorf *et al.*, 2009). Conceptually, a defocus error can be described as the response of a linear system (Fresnel diffraction), which would suggest a removal of the defocus error by applying an inverse filter. However, the optical transfer function (OTF) for out-of-focus planes contains zeros, and the application of an inverse filter will typically result in images of low quality.

Alternatively, it is possible to design a lens function, for which the OTF never assumes the optimum shape of a diffraction limited imaging system, but which contains no zeros and a shape only depending weakly on the position of the object along the optical axis. This means, it is possible to apply an inverse filter numerically to the recorded low quality image to approximate more closely a diffraction limited OTF over a range of object planes superior to the depth of field predicted by the f -number of the lens. Originally, a cubic phase plate was suggested as a filter function to achieve this effect (Dowski and Cathey, 1995). Other families of filter functions were shown to provide a systems response suitable to extending the depth of field.

Recording defocused objects corresponds to the projections of the in-focus WDF after shearing relative to the x axis. This is equivalent to projections of the WDF at oblique angles (compare Section Phase-Space Tomography and Phase Retrieval). The Fourier transform of these projections are equivalent to slices of the in-focus AF at different angles. If the object is a perfect point source, the image represents the point-spread function of the system. The corresponding AF of the in-focus image, in Fig. 9, displays the point-spread function along the $\bar{x}$ axis and the in-focus OTF along the $\bar{v}$ axis. Slices S represent the OTF at a plane which is out-of focus by a distance

$$\Delta z = \frac{\tan(\theta)}{\lambda} \quad (32)$$

For phase-space tomography each slice corresponds to information about the AF, which is used to reconstruct the signal. Here, the same concept can be used to evaluate pupil functions. The AF represents a simultaneous display of all out-of-focus OTFs for a particular object plane. A filter function suitable for extending the depth-of-field corresponds to an AF, with cross-sections not changing over the range of angles for which the defocus error will be removed.

References

- Dowski, E.R., Cathey, W.T., 1995. Extended depth of field through wave-front coding. *Appl. Opt.* 34 (11), 1859–1866.
 Lohmann, A.W., Ojeda-Castañeda, J., Streibl, N., 1983. The influence of wave aberrations on the Wigner distribution. *Optica Applicata* 13 (4), 465–471.

- Lohmann, A.W., Testorf, M., Ojeda-Castañeda, J., 2004. The art and science of holography: A tribute to Emmett Leith and Yuri Denisyuk. In: Caulfield, H.J. (Ed.), *Holography and the Wigner Function*. SPIE Press, pp. 129–144. (chapter **Holography and the Wigner Function**).
- Lubk, A., Röder, F., 2015. Phase-space foundations of electron holography. *Phys. Rev. A* 92,033844.
- Ozaktas, H.M., Zalevsky, Z., Kutay, M.A., 2001. *The Fractional Fourier Transform*. Chichester: John Wiley & Sons.
- Saleh, B.E.A., Teich, M.C., 1991. *Fundamentals of Photonics*. New York: John Wiley & Sons.
- Semichaevsky, A., Testorf, M., 2004. Phase-space interpretation of deterministic phase retrieval. *J. Opt. Soc. Am. A* 21 (11), 2173–2179.
- Situ, G., Sheridan, J.T., 2007. Holography: An interpretation from the phase-space point of view. *Opt. Lett.* 32 (24), 3492–3494.
- Smithey, D.T., Beck, M., Raymer, M.G., Faridani, A., 1993. Measurement of the Wigner distribution and the density matrix of light mode using optical homodyne tomography: Application to squeezed states and the vacuum. *Phys. Rev. Lett.* 70,1244–1247.
- Testorf, M., Hennelly, B., 2016. Self-imaging and discrete paraxial optics. In: Healy, J.J., Kutay, M.A., Ozaktas, H.M., Sheridan, J.T. (Eds.), *Linear Canonical Transforms – Theory and Applications*. Springer, pp. 257–292.
- Testorf, M., Hennelly, B., Ojeda-Castañeda, J., 2009. *Phase-Space Optics: Fundamentals and Applications*. McGraw-Hill.
- Torre, A., 2005. *Linear Ray and Wave Optics in Phase Space*. Amsterdam: Elsevier.
- Wolf, K.B., 2004. *Geometric Optics on Phase Space*. Springer.
- Zalevsky, Z., Mendlovic, D., 2004. *Optical Superresolution*. Springer.
- Zalevsky, Z., Mendlovic, D., Lohmann, A.W., 2000. Understanding superresolution in Wigner space. *J. Opt. Soc. Am. A* 17 (12), 2422–2430.

Information Theory in Imaging

FO Huck and CL Fales, NASA Langley Research Center, Hampton, VA, USA

© 2005 Elsevier Ltd. All rights reserved.

Introduction

Information theory is an intuitively attractive mathematical approach for combining observation with communication, or sensing with processing. So far, however, this approach has remained largely in the realm of signal processing and transmission, as first formalized by Shannon and Wiener over 50 years ago. Recent extensions of this approach to imaging have mostly addressed the interpretation of images, without explicitly accounting for the critical limiting factors that constrain image gathering and restoration. Yet, to be useful in the characterization of imaging, information theory must account for these constraints and lead to a close correlation between predicted and actual performance. The characterization that is outlined here traces the rate of information from the scene to the observer; and it pairs this rate, first, with the theoretical minimum data rate to assess the efficiency with which information can be transmitted, and, second, with the maximum-realizable fidelity to assess the quality with which images can be restored. Further details, especially for the mathematical development of figures of merit and for the design of imaging systems, can be found in the Further Reading.

Model of Imaging

Description

Modern imaging systems usually have the following three stages, as depicted in Fig. 1:

1. *Image gathering*, to capture the radiance field that is either reflected or emitted by the scene and transform this field into a digital signal;
2. *Signal coding*, to encode the acquired signal for the efficient transmission and/or storage of data and decode the received signal for the restoration of images. The encoding may be either lossless (without loss of information) or lossy (with some loss of information), if a higher compression is required;
3. *Image restoration*, to produce a digital representation of the scene and transform this representation into a continuous image for the observer.

Human vision may be included in this model as a fourth stage, to characterize imaging systems in terms of the information rate that the eye of the observer conveys from the displayed image to the higher levels of the brain.

These stages are similar to each other in one respect: each contains one or more transfer functions, either continuous or discrete, followed by one or more sources of noise, also either continuous or discrete. The continuous transfer function of the image-gathering device (and of the human eye) is followed by sampling that transforms the captured radiance field into a discrete signal with analog magnitudes. In most devices the photodetection mechanism conveys this signal serially into an analog-to-digital (A/D) converter to produce a digital signal. However, analog signal processing in a parallel structure, akin to that in human vision, has many advantages that are increasingly emulated by neural networks.

This model of imaging differs from the classical model of communication that Shannon and Wiener addressed in two fundamental ways:

1. *The continuous-to-discrete transformation in image gathering.* Whereas communication is constrained critically only by bandwidth and noise, image gathering is constrained also by the compromise between blurring and aliasing, due to limitations in the response of optical apertures and in the sampling density of photodetection mechanisms. This additional constraint requires the rigorous treatment of insufficiently sampled signals throughout the imaging system.
2. *The sequence of image gathering followed by signal coding.* Whereas the characterization of communication addresses the perturbations that occur at the encoder and decoder or during transmission, the characterization of imaging systems must also address the perturbations in the image-gathering process prior to coding. These additional perturbations require a clear

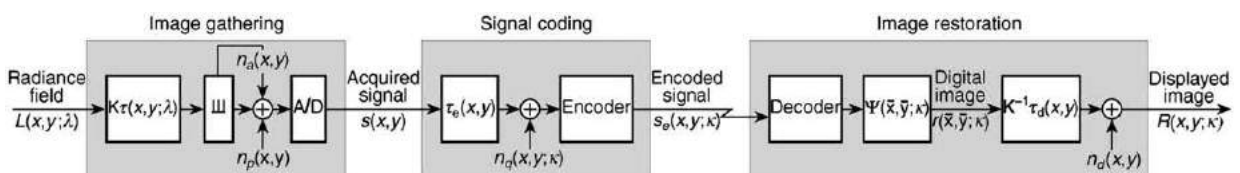


Fig. 1 Model of digital imaging systems.

distinction between the information rate of the encoded signal and the associated theoretical minimum data rate or, more generally, between information and entropy.

The mathematical model of imaging that is outlined here accounts for the appropriate deterministic and stochastic (random) properties of natural scenes and for the critical limiting factors that constrain the performance of imaging systems. The corresponding formulations of the figures of merit are sufficiently general to account for a wide range of radiance field properties, photodetection mechanisms, and signal processing structures. However, the illustrations emphasize: (a) the salient properties of natural scenes that dominate the captured radiance field in the visible and near-infrared region under normal daylight conditions; and (b) the photodetector arrays that are used most frequently as the photodetection mechanism of image-gathering devices.

Radiance Field

The radiance field $L(x, y; \lambda)$, reflected by natural scenes, may be modeled approximately as

$$L(x, y; \lambda) = \frac{1}{\pi} I(\lambda) \rho(\lambda) \rho(x, y) \quad (1)$$

where $I(\lambda)$ and $\rho(\lambda)$, respectively, represent the spectral irradiance and reflectance properties as a function of the wavelength λ , and $\rho(x, y)$ represents the random reflectance properties as a function of the spatial coordinates (x, y) . Fig. 2 illustrates a model of $\rho(x, y)$ for which the power spectral density (PSD) $\hat{\Phi}_\rho(v, \omega)$ corresponds closely to that typical of natural scenes, as given by

$$\hat{\Phi}_\rho(v, \omega) = \frac{2\pi\mu_\rho^2\sigma_\rho^2}{[1 + (2\pi\mu_\rho\zeta)^2]^{3/2}} \quad (2)$$

where $\zeta^2 = v^2 + \omega^2$ and (v, ω) are the spatial frequency coordinates of the Fourier-transform domain (see Fig. 3). This model consists of random polygons, whose boundaries are distributed according to Poisson probability with a mean separation of μ_ρ , and whose magnitudes are distributed according to independent zero-mean Gaussian statistics of variance σ_ρ^2 . The constraints associated with Eq. (1) are satisfied by letting the mean reflectance $\bar{\rho} = 0.5$ and the standard deviation $\sigma_\rho \leq 0.2$.

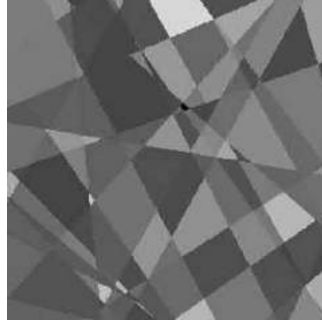


Fig. 2 Model of reflectance $\rho(x, y)$.

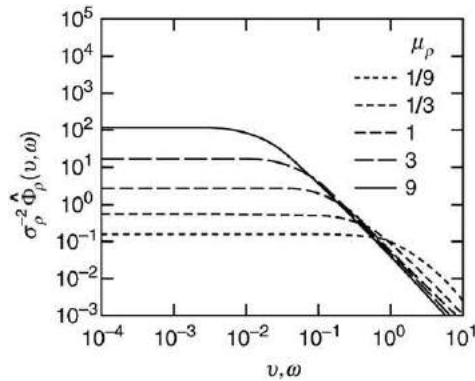


Fig. 3 Normalized PSD of the reflectance $\rho(x, y)$.

Image Gathering

Fig. 4 depicts the basic optical geometry of an image-gathering device, in which the objective lens forms an image of the captured radiance field across the photodetector array, and the array transforms this image into a discrete signal. The sampling lattice (X, Y) , which determines the resolution of this device, is formed in angular units by the separation (X_p, Y_p) of the photodetector-array apertures and the distance ℓ_p as given for one of the dimensions by

$$X = 2 \tan^{-1} \frac{X_p}{2\ell_p} \approx \frac{X_p}{\ell_p} \text{ rad} = \frac{57.3 X_p}{\ell_p} \text{ deg} \quad (3)$$

Normally, the sampling lattice is square, as shown in **Fig. 5**, so that $X_p = Y_p$. Then the corresponding area of the photodetector apertures is $A_p = \gamma_p^2$, and the solid angle Ω of the instantaneous field of view (IFOV) is $\Omega \approx A_p / \ell_p^2$ ster.

The sampling lattice projected onto the scene at the distance ℓ_o has the dimensions $(X_o, Y_o) \approx (X\ell_o, Y\ell_o)$. Because these dimensions represent the finest spatial detail that normally can be restored, it is convenient to measure the mean spatial detail of the scene relative to the sampling lattice, as given by $\mu = \mu_p(X_o, Y_o)^{-1/2}$, which reduces to $\mu = \mu_p X_o^{-1}$ for $X = Y$. Hence, $\mu = 1$ indicates that the sampling interval is equal to the mean spatial detail.

The image-gathering process that transforms the continuous radiance field $L(x, y; \lambda)$ into the discrete signal $s(\mathbf{x}, \mathbf{y})$ may be modeled by the expression

$$s(\mathbf{x}, \mathbf{y}) = K \rho(\mathbf{x}, \mathbf{y}) * \tau(\mathbf{x}, \mathbf{y}) + n_p(\mathbf{x}, \mathbf{y}) \quad (4)$$

where the discrete coordinates $(\mathbf{x}, \mathbf{y}) = (mX, nY)$, and the symbol $*$ denotes continuous spatial convolution. This model accounts, in the simplest form, for the reflectance-to-signal conversion gain K , the reflectance $\rho(x, y)$ of the scene, the spatial response (SR) $\tau(x, y)$ of the image-gathering device, the continuous-to-discrete transformation by the photodetector-array lattice, and the discrete noise $n_p(\mathbf{x}, \mathbf{y})$ of the photodetectors. The discrete coordinates are defined by the sampling function

$$\mathbb{I} \equiv \mathbb{I}(\mathbf{x}, \mathbf{y}) = XY \sum_{\mathbf{x}, \mathbf{y}} \delta(\mathbf{x} - \mathbf{x}, \mathbf{y} - \mathbf{y}) \quad (5)$$

where $\delta(x, y)$ is the Dirac delta function; and the gain is given by

$$K = \frac{1}{\pi} A_\ell \Omega \int I(\lambda) \rho(\lambda) r(\lambda) d\lambda \quad (6)$$

where $A_\ell = \pi D^2/4$ is the area of the objective lens with diameter D , and $r(\lambda)$ is the responsivity of the photodetectors. This gain relates the magnitude of the signal $s(\mathbf{x}, \mathbf{y})$ directly to that of the reflectance $\rho(x, y)$; and, together with the variances σ_ρ^2 and σ_p^2 of the reflectance and the photodetector noise, respectively, it also determines the rms-signal-to-rms-noise ratio (SNR) $K\sigma_\rho/\sigma_p$.

The DFT of $s(\mathbf{x}, \mathbf{y})$ is the periodic spectrum $\tilde{s}(\mathbf{v}, \omega)$ given by

$$\tilde{s}(\mathbf{v}, \omega) = K \hat{\rho}(\mathbf{v}, \omega) \hat{\tau}(\mathbf{v}, \omega) * \mathbb{I} + \tilde{n}_p(\mathbf{v}, \omega) \quad (7)$$

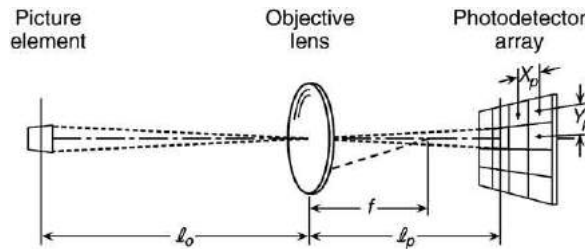


Fig. 4 Optical geometry of image-gathering device.

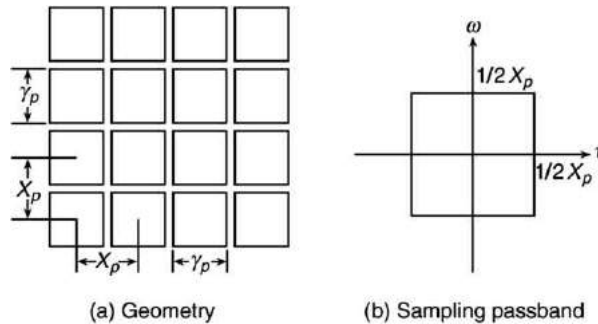


Fig. 5 Photodetector-array geometry and sampling passband.

where $\hat{\rho}(v, \omega)$ is the spectrum of the reflectance, $\hat{\tau}(v, \omega)$ is the spatial frequency response (SFR) of the image-gathering device, and $\hat{n}_p(v, \omega)$ is the spectrum of the photodetector noise. The symbol $\hat{\mathbb{I}}$ is the Fourier transform of the sampling function $\mathbb{I}$. It accounts for the sidebands generated by the sampling process and is given by

$$\hat{\mathbb{I}} \equiv \hat{\mathbb{I}}(v, \omega) = \sum_{v, \omega} \delta(v - v, \omega - \omega) \quad (8)$$

where $(v, \omega) = (m/X, n/Y)$. The corresponding sampling passband is

$$\hat{B} = \left\{ (v, \omega); |v| < \frac{1}{2X}, |\omega| < \frac{1}{2Y} \right\} \quad (9)$$

The above expression for $\hat{s}(v, \omega)$ can be rewritten more conveniently as

$$\hat{s}(v, \omega) = K\hat{\rho}(v, \omega)\hat{\tau}(v, \omega) + \hat{n}(v, \omega) \quad (10)$$

where

$$\hat{n}(v, \omega) = \hat{n}_a(v, \omega) + \hat{n}_p(v, \omega) \quad (11)$$

is the spectrum of the total noise in the acquired signal. The first term $\hat{n}_a(v, \omega)$ represents the aliased signal components caused by insufficient sampling, as given by

$$\hat{n}_a(v, \omega) = K\hat{\rho}(v, \omega)\hat{\tau}(v, \omega) * \hat{\mathbb{I}}_s \quad (12)$$

where

$$\hat{\mathbb{I}}_s \equiv \hat{\mathbb{I}}_s(v, \omega) = \sum_{v, \omega \neq (0,0)} \delta(v - v, \omega - \omega) \quad (13)$$

If, for spatial coordinates at either the scene or the photodetector array, the unit of (x, y) is meters, then the unit of (v, ω) is cycles m^{-1} ; and if, for angular coordinates centered at the objective lens, the unit of (x, y) is radians, then the unit of (v, ω) is cycles rad^{-1} . Either way, it is convenient to normalize the sampling intervals to unity (i.e., $X=Y=1$). Then the area of the sampling passband is $|\hat{B}| = 1$, the unit of (x, y) is samples, and the unit of (v, ω) is cycles $sample^{-1}$.

The SFR $\hat{\tau}(v, \omega)$ of the image-gathering device is

$$\hat{\tau}(v, \omega) = \hat{\tau}_\ell(v, \omega)\hat{\tau}_p(v, \omega) \quad (14)$$

where $\hat{\tau}_\ell(v, \omega)$ and $\hat{\tau}_p(v, \omega)$ are the SFRs of the objective lens and photodetector apertures, respectively. This SFR may be modeled approximately by the circularly symmetric Gaussian shape

$$\hat{\tau}(v, \omega) = \exp[-(\zeta/\zeta_c)^2] \quad (15)$$

where, as shown in Fig. 6, the optical-response index ζ_c controls the trade-off between the blurring that is caused by the decrease of the SFR inside the sampling passband and the aliasing that is caused by the extension of the SFR beyond this passband.

Signal Coding

Signal coding may be divided into two stages, as depicted in Fig. 1. The first stage, which is represented by the digital operator $\tau_c(x, y)$ and the quantization noise $n_q(x, y; \kappa)$, accounts for manipulations of the acquired signal that seek to extract those features of the captured radiance field considered to be perceptually most significant, which usually entails some loss of information. The second stage, which is represented by the encoder, accounts for manipulations that seek to reduce redundancies in the acquired

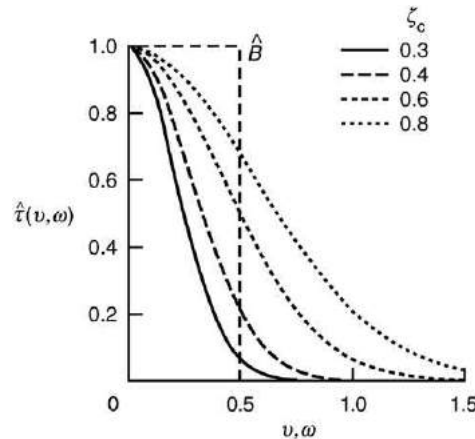


Fig. 6 SFRs $\hat{\tau}(v, \omega)$ of the image-gathering device relative to the sampling passband $\hat{B}$ for unit sampling intervals (i.e., $X=Y=1$).

signal, without loss of information. The resultant encoded signal $s_e(\mathbf{x}, \mathbf{y}; \boldsymbol{\kappa})$ may be expressed in the same form as Eq. (4) for the acquired signal $s(\mathbf{x}, \mathbf{y})$ as

$$s_e(\mathbf{x}, \mathbf{y}; \boldsymbol{\kappa}) = s(\mathbf{x}, \mathbf{y}) \otimes \tau_e(\mathbf{x}, \mathbf{y}) + n_q(\mathbf{x}, \mathbf{y}; \boldsymbol{\kappa}) \quad (16)$$

where the symbol $\otimes$, denotes discrete spatial convolution, and $n_q(\mathbf{x}, \mathbf{y}; \boldsymbol{\kappa})$ is the quantization noise. If the data transmission and/or storage are error-free, as is assumed here, then the signal that leaves the decoder remains identical to the signal that enters the encoder.

The DFT of $s_e(\mathbf{x}, \mathbf{y}; \boldsymbol{\kappa})$ is the periodic spectrum $\tilde{s}_e(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa})$ given in the same form as Eq. (10) by

$$\tilde{s}_e(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) = K\hat{\rho}(\mathbf{v}, \boldsymbol{\omega})\hat{\Gamma}_e(\mathbf{v}, \boldsymbol{\omega}) + \hat{n}_e(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) \quad (17)$$

where

$$\hat{\Gamma}_e(\mathbf{v}, \boldsymbol{\omega}) = \hat{\tau}(\mathbf{v}, \boldsymbol{\omega})\hat{\tau}_e(\mathbf{v}, \boldsymbol{\omega}) \quad (18)$$

is the throughput SFR from image gathering to transmission, and

$$\hat{n}_e(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) = \hat{n}(\mathbf{v}, \boldsymbol{\omega})\hat{\tau}_e(\mathbf{v}, \boldsymbol{\omega}) + \tilde{n}_q(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) \quad (19)$$

is the accumulated noise in the encoded signal. For $\log \boldsymbol{\kappa}$ -bit quantization, where $\log \boldsymbol{\kappa}$ is the number of uniformly spaced quantization levels, $\tilde{n}_q(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa})$ may be treated (for ≥ 4) as independent Gaussian noise with the PSD

$$\tilde{\Phi}_q(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) = \sigma_q^2 = (\sigma_{s_e}/\boldsymbol{\kappa})^2 \quad (20)$$

where $\sigma_{s_e}^2$ is the variance given approximately by

$$\sigma_{s_e}^2 \approx K^2 \int_{\hat{B}} \hat{\Phi}_\rho(\mathbf{v}, \boldsymbol{\omega}) |\hat{\Gamma}_e(\mathbf{v}, \boldsymbol{\omega})|^2 d\mathbf{v} d\boldsymbol{\omega} \quad (21)$$

Image Restoration

Image restoration, as depicted in Fig. 1, transforms the decoded signal $s_e(\mathbf{x}, \mathbf{y}; \boldsymbol{\kappa})$ into the digital image $r(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \boldsymbol{\kappa})$ as modeled by the expression

$$r(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \boldsymbol{\kappa}) = \overline{\overline{\overline{s_e}}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\kappa}) \otimes \Psi(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \boldsymbol{\kappa}) \quad (22)$$

The operation indicated by $\overline{\overline{\overline{s_e}}}(\mathbf{x}, \mathbf{y}; \boldsymbol{\kappa})$ (called zero padding) prepares the decoded signal for image restoration with interpolation, and the operator $\Psi(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \boldsymbol{\kappa})$ produces the digital image $r(\bar{\mathbf{x}}, \bar{\mathbf{y}}; \boldsymbol{\kappa})$ on the interpolation lattice $\overline{\overline{\overline{\cdot}}}$ that is Z times denser than the sampling lattice $\overline{\overline{\cdot}}$. If the interpolation in the image restoration is sufficiently dense (normally, for $Z \geq 4$), then the blurring and raster effects of the image-display process are suppressed entirely. This condition leads not only to the best realizable image quality but also to the simplest expression for the continuous displayed image, as given in the same form as Eqs. (10) and (17) by

$$\hat{R}(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) = \hat{\rho}(\mathbf{v}, \boldsymbol{\omega})\hat{\Gamma}_r(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) + \hat{n}_r(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) \quad (23)$$

where

$$\hat{\Gamma}_r(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) = \hat{\tau}(\mathbf{v}, \boldsymbol{\omega})\hat{\tau}_e(\mathbf{v}, \boldsymbol{\omega})\hat{\Psi}(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) \quad (24)$$

is the throughput SFR from image gathering to display, and

$$\hat{n}_r(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) = K^{-1}\hat{n}_e(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa})\hat{\Psi}(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) = \hat{n}_{ar}(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) + \hat{n}_{pr}(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) + \hat{n}_{qr}(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) \quad (25)$$

is the accumulated noise in the restored image. This noise may be expressed explicitly by substituting the noise components given above. The noise due to the granularity in image-display mediums is not accounted for here.

The Wiener filter, that minimizes the mean-square error due to the perturbations in image gathering and coding, is given by

$$\hat{\Psi}(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) = \frac{\hat{\Phi}'_\rho(\mathbf{v}, \boldsymbol{\omega})\hat{\Gamma}_e^*(\mathbf{v}, \boldsymbol{\omega})}{\hat{\Phi}'_\rho(\mathbf{v}, \boldsymbol{\omega})|\hat{\Gamma}_e(\mathbf{v}, \boldsymbol{\omega})|^2 + \hat{\Phi}'_{n_e}(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa})} \quad (26)$$

where $\hat{\Phi}'_\rho(\mathbf{v}, \boldsymbol{\omega}) = \sigma_\rho^{-2}\hat{\Phi}_\rho(\mathbf{v}, \boldsymbol{\omega})$ is the normalized PSD of the radiance field and

$$\hat{\Phi}'_{n_e}(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) = (K\sigma_\rho)^{-2}[\hat{\Phi}_n(\mathbf{v}, \boldsymbol{\omega})|\hat{\tau}_e(\mathbf{v}, \boldsymbol{\omega})|^2 + \tilde{\Phi}_q(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa})]$$

is the normalized PSD of the accumulated noise in the encoded signal. If the photodetector noise is white so that the PSD $\tilde{\Phi}_p(\mathbf{v}, \boldsymbol{\omega})$ is equal to its variance σ_p^2 and the PSD $\tilde{\Phi}_q(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa})$ is equal to its variance σ_q^2 , then the PSD of the noise simplifies to

$$\hat{\Phi}'_{n_e}(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) = \hat{\Phi}'_{n_e}(\mathbf{v}, \boldsymbol{\omega}) + \hat{\Phi}'_q(\mathbf{v}, \boldsymbol{\omega}; \boldsymbol{\kappa}) \quad (27)$$

where

$$\hat{\Phi}'_{n_e}(\mathbf{v}, \boldsymbol{\omega}) = [\hat{\Phi}'_\rho(\mathbf{v}, \boldsymbol{\omega})\hat{\tau}(\mathbf{v}, \boldsymbol{\omega})]^2 * \overline{\overline{\overline{\delta}}} + (K\sigma_\rho/\sigma_\rho)^{-2}|\hat{\tau}_e(\mathbf{v}, \boldsymbol{\omega})|^2 \quad (28)$$

and

$$\hat{\Phi}'_q(v, \omega; \kappa) = (K\sigma_\rho/\sigma_{s_e})^{-2}\kappa^{-2} \quad (29)$$

These simplifying conditions permit the Wiener filter given above and the figures of merit given below to be expressed as a function of the SNR $K\sigma_\rho/\sigma_p$, which can be specified without accounting explicitly for either the gain K or the variances σ_ρ^2 and σ_p^2 .

Figures of Merit

Information Rate $\mathcal{H}$

The information rate $\mathcal{H}$ (in bits sample⁻¹) of the encoded signal $s_e(\mathbf{x}, \mathbf{y}; \kappa)$ is defined by

$$\mathcal{H} = \mathcal{E}[\tilde{s}_e(v, \omega; \kappa)] - \mathcal{E}[\tilde{s}_e(v, \omega; \kappa) | \hat{\rho}(v, \omega) : (v, \omega) \in \hat{B}] \quad (30)$$

where the first term, $\mathcal{E}[\cdot]$ represents the entropy of $\tilde{s}_e(v, \omega; \kappa)$ and the second term, $\mathcal{E}[\cdot | \cdot]$ represents the conditional entropy of this signal when the reflectance spectrum $\hat{\rho}(v, \omega)$ within the sampling passband $\hat{B}$ is known. For the assumptions that have been made about image gathering and coding, the conditional entropy becomes the entropy of the accumulated noise $\hat{n}_e = (v, \omega; \kappa)$, so that $\mathcal{H}$ simplifies to

$$\mathcal{H} = \mathcal{E}[\tilde{s}_e(v, \omega; \kappa)] - \mathcal{E}[\hat{n}_e(v, \omega; \kappa) : (v, \omega) \in \hat{B}] = \frac{1}{2} \int_{\hat{B}} \int \log \left[1 + \frac{\hat{\Phi}'_\rho(v, \omega) |\hat{\Gamma}_e(v, \omega)|^2}{\hat{\Phi}'_{n_e}(v, \omega; \kappa)} \right] dv d\omega \quad (31)$$

The theoretical upper bound of $\mathcal{H}$ is Shannon's channel capacity $\mathcal{C}$ for a bandwidth-limited system, with white photodetector noise and an average power limitation (here the variance σ_ρ^2), as given by

$$\mathcal{C} = \frac{1}{2} \log[1 + (K\sigma_\rho/\sigma_p)^2] \quad (32)$$

$\mathcal{H}$ can reach $\mathcal{C}$ only when

$$\begin{aligned} \hat{\Phi}_\rho(v, \omega) &= \begin{cases} \sigma_\rho^2, & (v, \omega) \in \hat{B} \\ 0, & \text{elsewhere} \end{cases} \\ \hat{\tau}(v, \omega) &= \begin{cases} 1, & (v, \omega) \in \hat{B} \\ 0, & \text{elsewhere} \end{cases} \end{aligned} \quad (33)$$

However, neither of these two conditions can occur in practice because both the PSD $\hat{\Phi}_\rho(v, \omega)$ of natural scenes and the SFR $\hat{\tau}(v, \omega)$ of image-gathering devices decrease gradually with increasing spatial frequency, as depicted in **Figs. 3** and **6**, respectively. Hence, $\mathcal{H}$ can never reach $\mathcal{C}$.

Theoretical Minimum Data Rate $\mathcal{E}$

The theoretical minimum data rate $\mathcal{E}$ (also in bits sample⁻¹) that is required to convey the information rate $\mathcal{H}$ is defined by

$$\mathcal{E} = \mathcal{E}[\tilde{s}_e(v, \omega; \kappa)] - \mathcal{E}[\tilde{s}_e(v, \omega; \kappa) | \tilde{s}(v, \omega)] \quad (34)$$

where the second term, $\mathcal{E}[\cdot | \cdot]$ is the uncertainty of $\tilde{s}_e(v, \omega; \kappa)$ when $\tilde{s}(v, \omega)$ is known. For the assumptions that have been made about the coding process, the conditional entropy becomes the entropy $\mathcal{E}[\hat{n}_q(v, \omega; \kappa)]$ of the quantization noise, so that $\mathcal{E}$ simplifies to

$$\mathcal{E} = \mathcal{E}[\tilde{s}_e(v, \omega; \kappa)] - \mathcal{E}[\hat{n}_q(v, \omega; \kappa)] = \frac{1}{2} \int_{\hat{B}} \int \log \left[1 + \frac{\hat{\Phi}'_\rho(v, \omega) |\hat{\Gamma}_e(v, \omega)|^2 + \hat{\Phi}'_{n_e}(v, \omega)}{\hat{\Phi}'_q(v, \omega; \kappa)} \right] dv d\omega \quad (35)$$

This expression for $\mathcal{E}$ represents the entropy of completely decorrelated data.

Information Efficiency $\mathcal{H}/\mathcal{E}$

The information efficiency of completely decorrelated data is defined by the ratio $\mathcal{H}/\mathcal{E}$. This ratio reaches its upper bound $\mathcal{H}/\mathcal{E} = 1$ when $\hat{\Phi}'_{n_e}(v, \omega) \ll \hat{\Phi}'_\rho(v, \omega) |\hat{\Gamma}_e(v, \omega)|^2$ and $\hat{\Phi}'_{n_e}(v, \omega) \ll \hat{\Phi}'_\rho(v, \omega; \kappa)$ because then both **Eq. (31)** for $\mathcal{H}$ and **Eq. (35)** for $\mathcal{E}$ reduce to the same expression

$$\mathcal{H} = \mathcal{E} = \frac{1}{2} \int_{\hat{B}} \int \log \left[1 + \frac{\hat{\Phi}'_\rho(v, \omega) |\hat{\Gamma}_e(v, \omega)|^2}{\hat{\Phi}'_q(v, \omega; \kappa)} \right] dv d\omega \quad (36)$$

This upper bound can be approached with a small loss in $\mathcal{H}$ only when the aliasing and photodetector noise are small. Hence, the electro-optical design that optimizes $\mathcal{H}$ also optimizes $\mathcal{H}/\mathcal{E}$. However, a compromise always remains between $\mathcal{H}$ and $\mathcal{H}/\mathcal{E}$ in the selection of the number of quantization levels: $\mathcal{H}$ favors fine quantization, whereas $\mathcal{H}/\mathcal{E}$ favors coarse quantization.

Maximum-Realizable Fidelity $\mathcal{F}$

The maximum-realizable fidelity $\mathcal{F}$ of the images restored with the Wiener filter $\hat{\Psi}(v, \omega; \kappa)$ can be expressed as a function of the throughput SFR $\hat{\Gamma}_r(v, \omega; \kappa)$ given by Eq. (24) as

$$\mathcal{F} = \int \int \hat{\Phi}'_p(v, \omega) \hat{\Gamma}_r(v, \omega; \kappa) dv d\omega \quad (37)$$

or, equivalently, as a function of the integrand $\hat{\mathcal{H}}(v, \omega; \kappa)$ of the information rate $\mathcal{H}$ given by Eq. (31) as

$$\mathcal{F} = \int \int \hat{\Phi}'_p(v, \omega) [1 - 2^{-\hat{\mathcal{H}}(v, \omega; \kappa)}] dv d\omega \quad (38)$$

This relationship between fidelity and information rate can be extended to other goals. For example, the enhancement of spatial features (e.g., edges and boundaries), for robotic vision, can be accommodated by letting $\tilde{\epsilon}_e(v, \omega)$ be the desired feature-enhancement filter.

Performance and Design Trade-offs

Electro-optical Design

Fig. 7 presents the information rate $\mathcal{H}$ (for lossless encoding) as a function of the optical-response index ζ_c and the mean spatial detail μ . The information rate $\mathcal{H}$ versus ζ_c is given for several SNRs $K\sigma_p/\sigma_p$ and $\mu=1$, and $\mathcal{H}$ versus μ is given for the three informationally optimized designs specified in Table 1.

The curves show that $\mathcal{H}$ can reach its highest possible value only when both of the following conditions are met:

1. The relationship between the SFR $\hat{\tau}(v, \omega)$ and the sampling passband $\hat{B}$ is chosen to optimize $\mathcal{H}$ for the available SNR $K\sigma_p/\sigma_p$. This design is said to be informationally optimized.

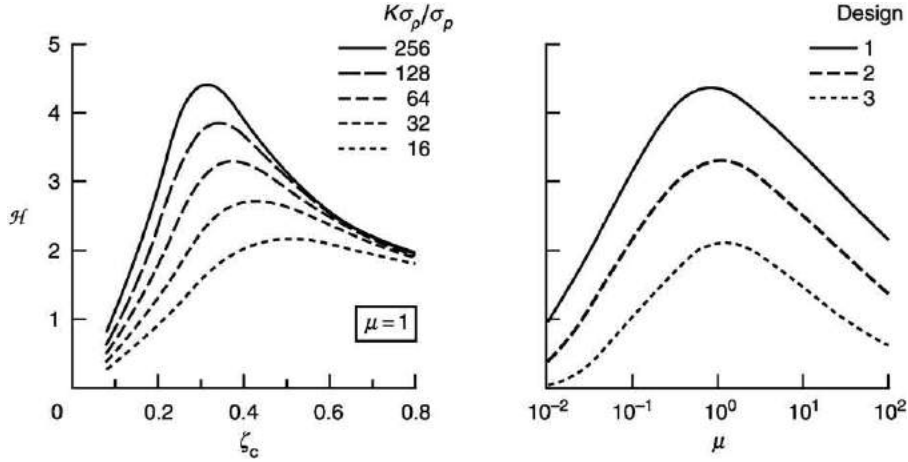


Fig. 7 Information rate $\mathcal{H}$ versus the optical-response index ζ_c and the mean spatial detail μ (relative to the sampling interval). The designs are specified in Table 1, and the SFRs $\hat{\tau}(v, \omega)$ for ζ_c are shown in Fig. 6.

Table 1 Informationally optimized designs

Design	$K\sigma_p/\sigma_p$	ζ_c	$\mathcal{H}^*$
1	256	0.3	4.4
2	64	0.4	3.3
3	16	0.5	2.2

* $\mu=1$.

2. The relationship between the sampling passband $\hat{B}$ and the PSD $\hat{\Phi}_\rho(v, \omega)$ is chosen to optimize $\mathcal{H}$. For the PSDs $\hat{\Phi}_\rho(v, \omega)$ typical of natural scenes, the best match occurs when the sampling interval is near the mean spatial detail of the scene (i.e., when $\mu \approx 1$).

These conditions are consistent with those for which the information rate $\mathcal{H}$ reaches Shannon's channel capacity $\mathcal{C}$. However, blurring and aliasing constrain $\mathcal{H}$ to values that are smaller than $\mathcal{C}$ by a factor of nearly two (i.e., $\mathcal{H}\mathcal{C}/2$). Hence, these conditions replace the role of white and band-limited signals in classical communication theory, to establish an upper bound on the information rate in imaging systems.

The above two conditions appeal intuitively when the problem of image restoration is considered:

1. The result that the SFR $\hat{\tau}(v, \omega)$ that optimizes $\mathcal{H}$ depends on the available SNR is consistent with the observation that, in one extreme, when the SNR is low, substantial blurring should be avoided because the photodetector noise would constrain the enhancement of fine spatial detail even if aliasing were negligible. In the other extreme, when the SNR is high, substantial aliasing should be avoided so that this enhancement is relieved from any constraints except those that the sampling passband inevitably imposes.
2. The result that $\mathcal{H}$ reaches its highest value when the sampling interval is near the mean spatial detail of the scene is consistent with the observation that, ordinarily, it would not be possible to restore spatial detail that is finer than the sampling interval, whereas it would be possible to restore much coarser detail from fewer samples.

Information Rate and Data Rate

Fig. 8 presents an information-entropy $\mathcal{H}(\mathcal{E})$ plot that characterizes the relationship between the information rate $\mathcal{H}$ of the encoded signal and the associated theoretical minimum data rate, or entropy, $\mathcal{E}$ for η -bit quantization. The encoder SFR $\hat{\tau}_e(v, \omega)$ is assumed to be unity. The plot illustrates the trade-off between $\mathcal{H}$ and $\mathcal{E}$ in the selection of the number of quantization levels for encoding the acquired signal. It shows, in particular, that the design that realizes the highest information rate $\mathcal{H}$ also enables the encoder to realize the highest information efficiency $\mathcal{H}/\mathcal{E}$. This result appeals intuitively, because the perturbations due to aliasing and photodetector noise that interfere with the image restoration can also be expected to do so with the signal decorrelation.

Throughput SFR

Fig. 9 illustrates the SFRs $\hat{\tau}(v, \omega)$ and $\hat{\Psi}(v, \omega; \kappa)$ of the image-gathering device and Wiener filter, respectively, and of their product, the throughput SFR $\hat{\Gamma}_r(v, \omega; \kappa)$ for the three designs specified in Table 1. As the accumulated noise becomes negligible, $\hat{\Gamma}_r(v, \omega; \kappa)$ approaches the ideal SFR of the classical communication channel, which is unity inside the sampling passband and zero outside.

Information Rate, Fidelity and Robustness

The PSD $\hat{\Phi}_\rho(v, \omega)$ of a scene is seldom known when images are restored. Moreover, even if the PSD were known to be the one given by Eq. (2), then the mean spatial detail μ would usually remain uncertain because it depends, not only on the properties of the scene, but also on the angular resolution and viewing distance of the image-gathering device. In practice, therefore, image restorations have to rely on estimates of the statistical properties of scenes. The tolerance of the quality of the restored image to errors in these estimates is referred to as the robustness of the image restoration.

Fig. 10 presents the information rate $\mathcal{H}$ and the corresponding maximum-realizable fidelity $\mathcal{F}$ for matched and mismatched Wiener restorations. These restorations represent the digital image $R(\bar{x}, \bar{y}; \kappa)$ or, equivalently, the continuous image $R(x, y; \kappa)$ displayed on a noise-free medium. The matched restorations use the correct value of μ , whereas the mismatched restorations use wrong estimates of μ . These curves reveal that the informationally optimized designs produce the highest fidelity and robustness, and that both improve with increasing $\mathcal{H}$. Moreover, these curves reveal that, ordinarily, one cannot go far wrong in Wiener restorations by assuming that the mean spatial detail is equal to the sampling interval (i.e., $\mu = 1$). This observation appeals

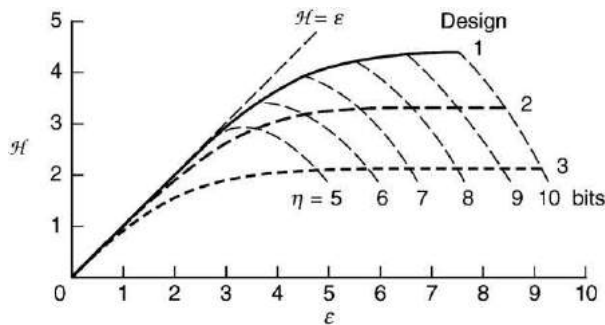


Fig. 8 The information-entropy $\mathcal{H}(\mathcal{E})$ plot of the information rate $\mathcal{H}$ versus the associated theoretical minimum data rate $\mathcal{E}$ for η -bit quantization. The designs are specified in Table 1.

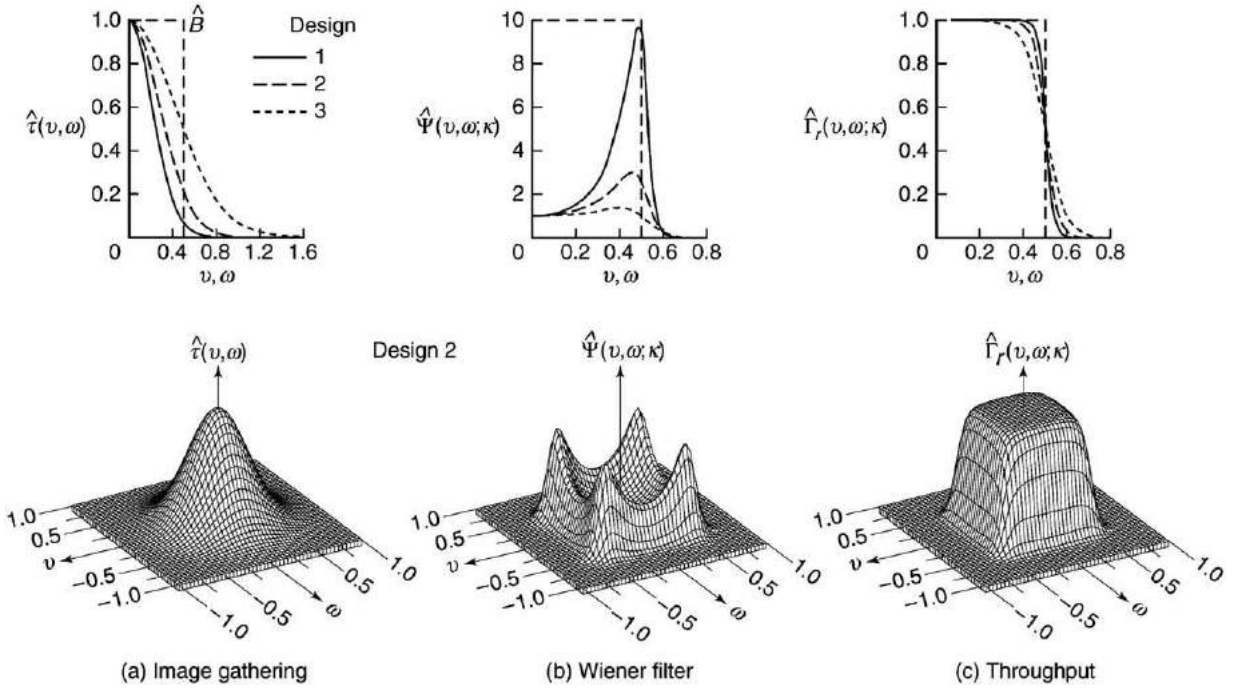


Fig. 9 SFRs of image gathering, restoration, and throughput. The designs are specified in Table 1.

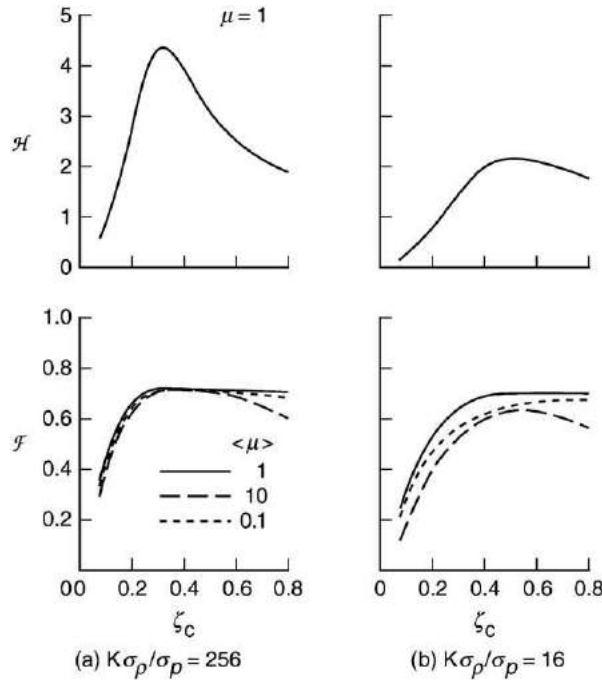


Fig. 10 Information rate $\mathcal{H}$ and maximum-realizable fidelity $\mathcal{F}$ versus the optical-response index ζ_c for two SNRs $K\sigma_p/\sigma_p$. $\mathcal{F}$ is given for the matched ($\langle\mu\rangle=1$) and two mismatched Wiener restorations.

intuitively because spatial detail that is much smaller cannot be resolved, and detail that is larger is not degraded substantially by blurring and aliasing.

Information Rate and Visual Quality

Figs. 11 and 12 present Wiener restorations and their three components, as given by Eq. (23), where $\rho(x, y) * \Gamma_r(x, y; \kappa)$ accounts for blurring, $n_{ar}(x, y; \kappa)$ accounts for aliasing, and $n_{pr}(x, y; \kappa)$ accounts for photodetector noise. The contrast of the illustrations of the

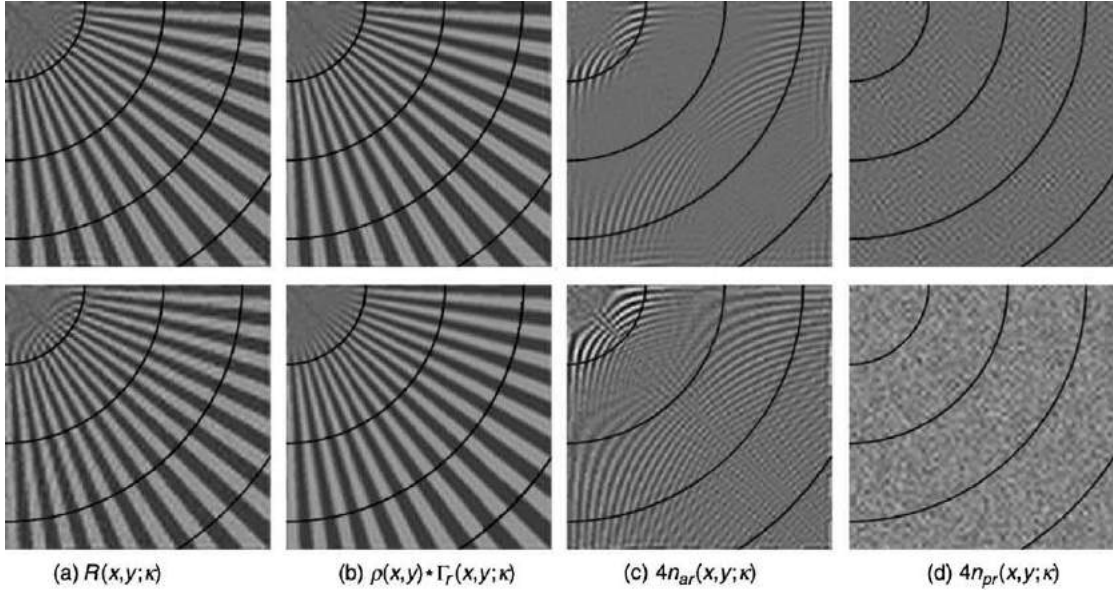


Fig. 11 Images restored with the Wiener filter and their three components for designs 1 (top) and 3 (bottom).

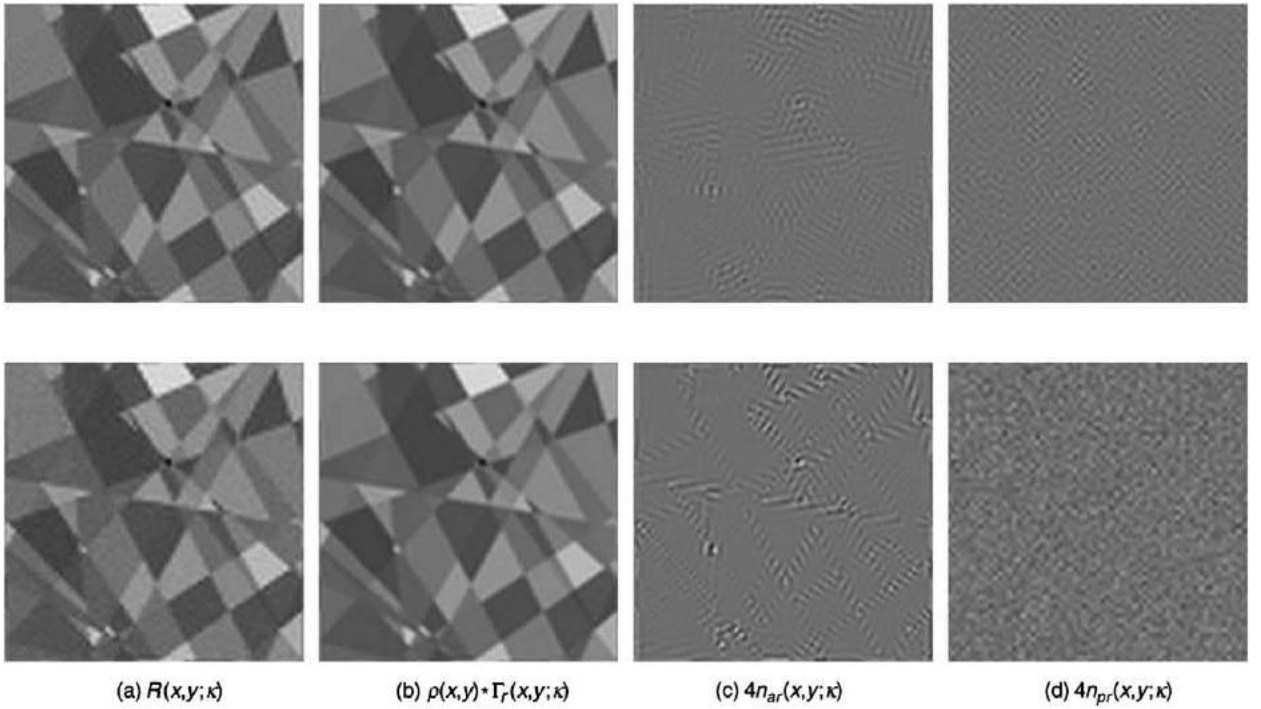


Fig. 12 Images restored with the Wiener filter and their three components for designs 1 (top) and 3 (bottom).

two noise components has been increased by a factor of four to improve their visibility. The extraneous structure that aliasing produces in the images of the resolution wedges is commonly referred to as Moiré pattern. The corresponding distortion in the images of the random polygons emerges more subtly as jagged (or staircase) edges.

The images are formed from a set of 64×64 samples with a $Z=4$ interpolation and have a 4×4 cm format. If these images were displayed in a smaller format, then the jagged edges could not be resolved by the observer and would appear to be the result of blurring or photodetector noise instead of aliasing. Aliasing has, therefore, often been overlooked as a significant source of image degradation.

The Wiener restoration produces images in which fine spatial detail near the limit of resolution normally has a high contrast approaching that in the scene. However, these images also tend to contain visually annoying defects, due to the accumulated noise

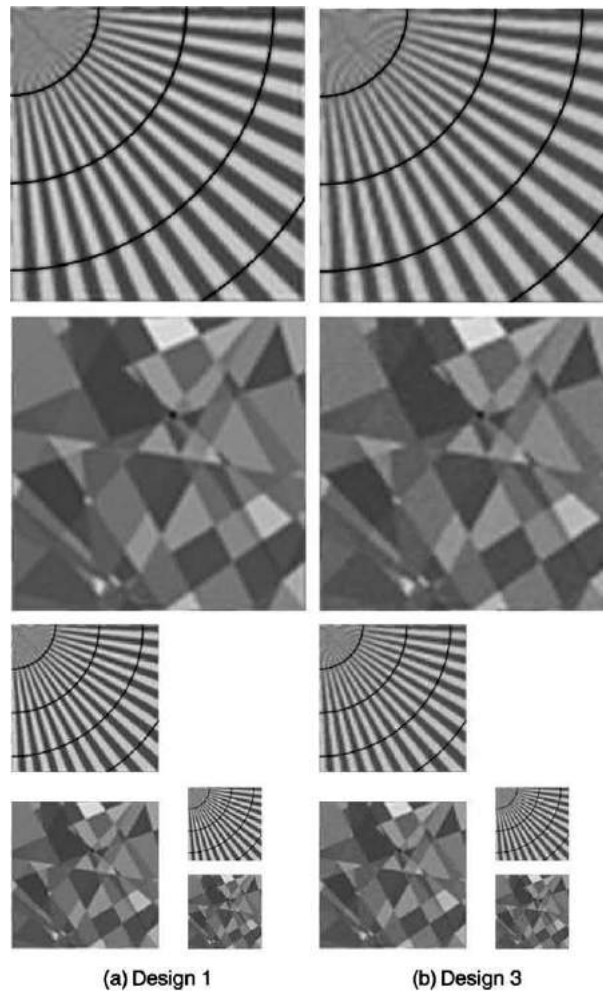


Fig. 13 Images restored with the Wiener filter and enhanced for visual quality. The large, medium, and small formats contain 16, 32 and 64 pixels/cm, respectively.

and the ringing near sharp edges (Gibbs phenomenon). Hence, it is often desirable to combine this restoration with an enhancement filter that gives the user some control over the trade-off among fidelity, sharpness and clarity.

Fig. 13 presents images for which the Wiener restorations given in **Figs. 11** and **12** are enhanced for visual quality by suppressing some of the defects due to aliasing and ringing. This enhancement improves clarity at the cost of a small loss in sharpness. The images are produced again from 64×64 pixels, but they are displayed in three different formats. The smallest format, with a density of 64 pixels/cm, corresponds to a display of 256×256 pixels in a 4×4 cm area, which is typical of the images found in the prevalent digital image-processing literature. As can be observed, this small format leads to images that are indistinguishable from each other because they hide many of the distortions that larger formats clearly reveal.

In general, if the images are displayed in a format that is large enough to reveal the finest spatial detail near the limit of resolution of the imaging system, then distortions due to the perturbations in this system become also visible. When no effort is spared to reduce these distortions without a perceptual loss of resolution and sharpness, then it becomes strongly evident that the best visual quality with which images can be restored improves with increasing information rate, even after the fidelity has essentially reached its upper bound. This improvement continues until it is gradually ended by the unavoidable compromise among sharpness, aliasing, and ringing as well as by the granularity of the image display.

Concluding Remarks

The performance of image-gathering devices and the human eye is constrained by the same critical limiting factors. When these factors are accounted for properly, then it emerges that the design of the image-gathering device that is optimized for the highest realizable information rate corresponds closely, under appropriate conditions, to the design of the human eye that evolution has

optimized for viewing the real world. This convergence of information theory and evolution toward the same design clearly supports the extension of information theory to the characterization of imaging systems.

Numerous other information-theoretic characterizations of imaging are now emerging. However, none of these characterizations accounts for the critical limiting factors that constrain image gathering and restoration, as outlined here. Hence, it is not possible for them to address the efficiency and accuracy with which images can be conveyed. Moreover, none of these characterizations combines observation with communication, or sensing with processing, in a general formalization. Instead, each of them still deals with a particular measurement or imaging problem. Hence, the role of information theory in imaging is far from mature. Many challenges remain: to develop general formalizations, to account for physical phenomena, and to correlate predictions with measurements.

Further Reading

- Fales, C.L., Huck, F.O., Alter-Gartenberg, R., Rahman, Z., 1996. Image gathering and digital restoration. *Philosophical Transactions of the Royal Society of London A354*, 2249–2287.
- Huck, F.O., Fales, C.L., 2001. *Encyclopedia of Imaging Science and Technology*, Characterization of image systems., New York: John Wiley & Sons.
- Huck, F.O., Fales, C.L., Rahman, Z., 1996. An information theory of visual communication. *Philosophical Transactions of the Royal Society of London A354*, 2193–2248.
- Huck, F.O., Fales, C.L., Rahman, Z., 1997. *Visual Communication: An Information Theory Approach*. Boston: Kluwer Academic Publishers.
- Lesurf, J.C.G., 1995. *Information and Measurement*. Bristol, UK: Institute of Physics Pub.
- O'Sullivan, J.A., Blahut, R.E., Snyder, D.L., 1998. Information-theoretic image formation. *IEEE Transactions Information Theory* 44, 2094–2123.
- Shannon, C.E., 1948. A mathematical theory of communication. *Bell System Technical Journal* 27, 379–423. and 28: 623–656.
- Shannon, C.E., Weaver, W., 1964. *The Mathematical Theory of Communication*. Urbana, IL: U. Illinois Press.
- Wiener, N., 1949. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. New York: John Wiley & Sons.

Inverse Problems and Computational Imaging

M Bertero and P Boccacci, University of Genova, Genova, Italy

© 2005 Elsevier Ltd. All rights reserved.

Introduction

A digital image is an array (2D image) or a cube (3D image) of quantized numbers, providing a representation of a physical object or a map of some distributed property of the object (density of matter, energy, etc.). It can be obtained by digitizing the original image (for instance, a photograph) or can be directly formed by the outputs of the imaging system (digital camera, microscope, telescope, etc.). When visualized on a suitable image display device, a digital image must provide a representation of the physical object which can be easily interpreted and analyzed. In many cases, however, this condition is not satisfied. We give two typical examples.

The first is that of an image degraded by blurring which can be introduced by defects of the optical systems (such as aberrations) or by image motion, image defocusing, etc. The representation of the object does not have the quality which should be desirable for its interpretation.

The second case is even more important and arises when the output of the imaging system is not directly related to the physical quantity to be imaged. For instance, in X-ray tomography, the output of a detector is the attenuation of an X-ray pencil crossing the body while the quantity to be imaged is the density of the body.

In the situations described above, the desired image can be obtained by solving a number of mathematical problems. If we denote f as the object, namely the target of the imaging system, and g as the image of f provided by the system, then the problems to be solved can be summarized as follows:

- develop a physical model of the system to obtain a mathematical relationship between the object f and the image g ; the computation of g from a given f is usually called the solution of the direct problem;
- solve the problem of obtaining the physical quantity f from given outputs g of the system; this second step is usually called the solution of the inverse problem.

The digital image of f , obtained by solving an inverse problem, is a computed one, generated by the computer where the algorithm for the solution of the inverse problem has been implemented.

It may be convenient to formulate both the direct and the inverse problem in terms of functions rather than arrays, cubes, etc., because such a formulation allows the use of powerful tools of functional analysis for their investigation. Next, the results obtained in the continuous case can be used for understanding the features of the corresponding discrete problem. In most important applications, for instance, the relationship between f and g is linear so that, in the discrete version of the direct problem, one obtains g by applying a suitable matrix A to f while, in the discrete version of the inverse problem, one obtains f from g by solving a linear algebraic system. However, the solution of the inverse problem is not so simple and the difficulties can be understood by looking at the corresponding continuous problem.

The basic reason relies on the fact that inverse problems are ill-posed in the sense of Hadamard: the solution may not exist; even if it exists it may not be unique; and, even if it exists and is unique, it may not depend continuously on the data. The last statement means that a small perturbation of the data, such as that caused by noise or any kind of experimental error, can completely modify the solution.

Inverse problems are ill-posed because imaging systems do not transmit complete information about the physical object. Therefore, the problem is to extract the useful information contained in the data or, in mathematical terms, to look for approximate solutions which are stable against noise. This is the purpose of a mathematical theory introduced in its general form by the Russian mathematician, Tikhonov and known as regularization theory.

We illustrate the general framework outlined above by means of a specific example – image deconvolution, also known as image deblurring, image restoration, etc. The object f is the image which should be recorded in the absence of degradation, while g is the image corrupted by aberrations or other causes of blurring. Both f and g are functions of 2D (or 3D) space variables $\mathbf{x}$ which are the coordinates of a point in the image domain. If the imaging system is isoplanatic, then it is described by a space-invariant point spread function (PSF) $K(\mathbf{x})$, which is the image of a point source located in the center of the image domain. The PSF provides the response of the system to any point source wherever it is located. On the other hand, its Fourier transform, the transfer function $\hat{K}(\omega)$ (the hat denotes Fourier transform and ω , the spatial frequencies associated to the space variables $\mathbf{x}$), tells us how a signal of a fixed frequency is propagated through the linear system, so that the blurring can also be viewed as a sort of frequency filtering.

Under these assumptions the image g is the convolution product of object f and PSF K :

$$g(\mathbf{x}) = \int K(\mathbf{x} - \mathbf{x}')f(\mathbf{x}')d\mathbf{x}' \quad (1)$$

and the solution of the direct problem is just the computation of a convolution product.

The image g , as given by Eq. (1), is the so-called noise-free image. The actually recorded image g_r is affected by a noise term and, in general, it can be written in the following form:

$$g_r(\mathbf{x}) = g(\mathbf{x}) + n(\mathbf{x}) \quad (2)$$

where $n(\mathbf{x})$ is a function describing noise contamination. It is a random function and, therefore, it is unknown even if one understands its statistical properties. We point out that Eq. (2) does not imply that the noise is additive or signal independent. For instance, in microscopy and astronomy, the contamination of the image is due both to photon noise, which satisfies Poisson statistics, and read-out noise, which is basically Gaussian and white. Therefore the noise term in Eq. (2) must only be intended as the difference between the noisy and the noise-free image.

In a first attempt at approaching the inverse problem, it seems quite natural to ignore the noise term $n(\mathbf{x})$ and to solve Eq. (1) for $f(\mathbf{x})$ with $g(\mathbf{x})$ replaced by $g_r(\mathbf{x})$. Then, by taking the Fourier transform of both sides of this equation and using the well-known convolution theorem, one finds the following relationship:

$$\hat{g}_r(\omega) = \hat{K}(\omega)\hat{f}(\omega) \quad (3)$$

which relates the Fourier transform of f and g to the transfer function of the imaging system.

The solution of this equation looks elementary. However, a first problem is due to the fact that an optical system is generally band-limited; its band Ω is the bounded domain of spatial frequencies where the transfer function is different from zero. Since the noise is not band-limited or, at least, has a band much broader than that of the optical system, outside Ω the right-hand side of Eq. (3) is zero while the left-hand side is not; in other words, the solution of the deconvolution problem, as formulated above, does not exist because no object f satisfies Eq. (3) everywhere. In this particular case, one can circumvent the problem by applying a suitable filter to $g_r(\mathbf{x})$ in order to suppress the out-of-band noise.

The second problem is the nonuniqueness of the solution of the problem with the filtered data. This is due to the so-called invisible objects, namely objects whose Fourier transform is zero on Ω , so that the corresponding images are exactly zero. If we find a solution of the deconvolution problem, by adding an arbitrary invisible object to this solution, we find another solution of the same problem. A well-defined solution can be obtained by requiring its Fourier transform to be zero outside Ω . In other words, we set to zero what is not transmitted by the imaging system.

The standard way for approaching the previous questions is to look for solutions in the least-squares sense, namely for objects which minimize the functional:

$$\varepsilon^2(f) = \|K * f - g_r\|^2 \quad (4)$$

where $*$ denotes the convolution product, as defined in Eq. (1), and the right-hand side is the square of the L^2 -norm (which can also be interpreted as energy) of the discrepancy between the recorded image g_r and the computed image $K * f$. If one looks for a least-squares solution with minimal energy then one automatically re-obtains the solution discussed above, namely an object whose Fourier transform in Ω is given by

$$\hat{f}_{\text{inv}}(\omega) = \frac{\hat{g}_r(\omega)}{\hat{K}(\omega)} \quad (5)$$

and is zero outside Ω . Such a solution is that provided by the so-called inverse filter.

However, there is an additional difficulty in that the transfer function usually tends to zero at the boundary of the band while the Fourier transform of the noise, hence that of g_r , does not, or tends to zero in a different way. It follows that $\hat{f}_{\text{inv}}(\omega)$ is divergent or very large at the boundary of the band. We conclude that its inverse Fourier transform does not exist or, if it exists, is affected by large artifacts due to noise propagation. The latter is the typical situation in the case of digital images. In Fig. 1 we give the result of a numerical simulation showing the typical effect produced by the inverse filter. The object is a beautiful picture of a galaxy recorded by the Hubble Space Telescope (HST), while the PSF is a computed one describing the blurring of HST before installation of the optics correction. The blurred image is obtained by convolving the object with the PSF and by adding white noise. The restored image produced by the inverse filter is completely corrupted by noise.

The inverse filter provides a solution which fits the data in the best possible way and therefore also fits the noise in the frequency domains where the signal is weakly transmitted by the imaging system. This can be achieved at the cost of an energy (L^2 -norm) of the solution which is too large. Therefore, such a conclusion suggests that one must search for a reasonable compromise between data fitting and energy of the solution. This is the basic idea of regularization theory, at least in its most simple form. The mathematical formulation is obtained by looking for objects which minimize the functional:

$$\Phi_\mu(f) = \|K * f - g_r\|^2 + \mu \|f\|^2 \quad (6)$$

where μ , the so-called regularization parameter, is a parameter controlling the trade-off between data fitting and energy of the solution: when μ is small, the data fitting is good and the energy is large while, when μ is large, the data fitting is poor and the energy is small.

For a given μ , the function f_μ which minimizes the functional of Eq. (6) is called the regularized solution of the problem. It is easy to show that its Fourier transform is given by

$$\hat{f}_\mu(\omega) = \frac{\hat{K}^*(\omega)}{|\hat{K}(\omega)|^2 + \mu} \hat{g}_r(\omega) \quad (7)$$

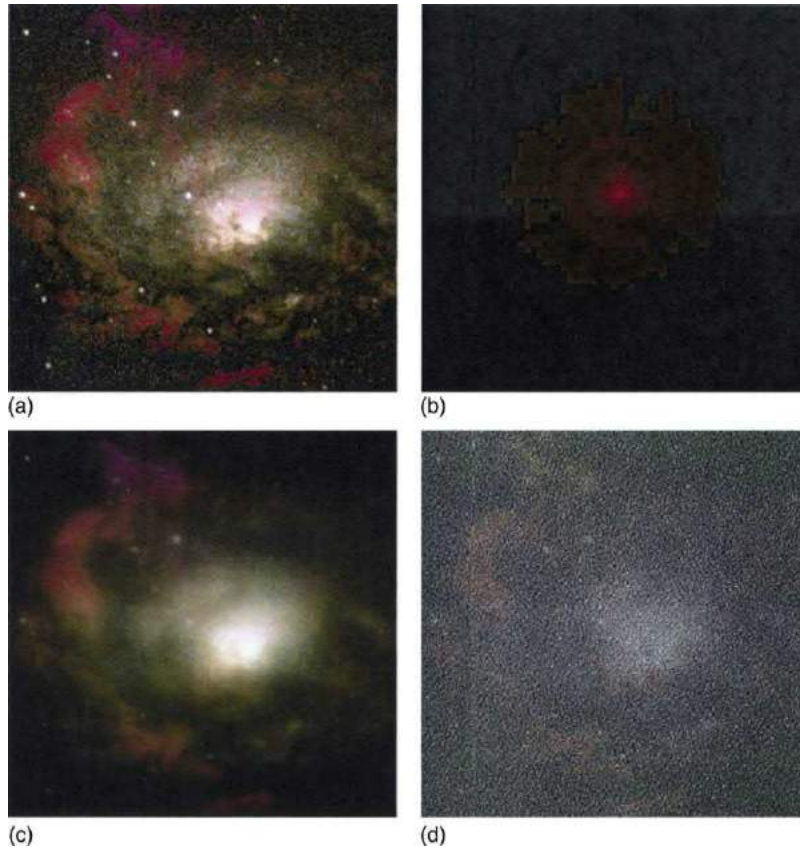


Fig. 1 Illustrating the effect of the inverse filter: (a) Image of the Circinus galaxy taken on 10 April 1999, with the Wide Field Planetary Camera 2 of the Hubble Space Telescope; (b) the PSF used for blurring the RGB components of the image (here shown in red); (c) the blurred image obtained by convolving the components of the image in (a) with the corresponding PSFs and adding noise; (d) the restoration obtained by applying the inverse filter to the three components of (c). Part (a) courtesy of NASA/Space Telescope Science Institute.

The regularized solutions form a one-parameter family of functions. When μ is small these functions are strongly contaminated by noise whose effect is gradually reduced when μ increases. For too large values of μ , one re-obtains blurred versions of the original object.

Such behavior is illustrated by the sequence of regularized solutions given in Fig. 2 and corresponding to the example of Fig. 1. The regularization parameter increases starting from left to right and from top to bottom. Its values are $\mu=0$ (inverse filter), 0.001, 0.01, 0.05, 0.1, and 0.5 (the noise is of the order of a few percent and the PSF is normalized in such a way that the sum of all its value is 1). The sequence makes evident the well-established theoretical result of the existence of an optimal value of the regularization parameter. Such an optimal value can be determined in the case of numerical simulations as that represented in Fig. 2. However, its estimation in the case of real images is more difficult. Several methods for the selection of μ have been proposed. We only mention the so-called discrepancy principle. It consists of determining the value of μ , such that the value of the discrepancy functional at f_μ , $\varepsilon^2(f_\mu)$ coincides with an estimate of the energy of the noise. In other words, the data are fitted with an accuracy comparable with their uncertainty.

As shown by Eq. (7), the regularized solution has the same band Ω of the imaging system, because $\hat{K}^*(\omega)$ is zero outside Ω . As is known, a band-limited function can be represented in terms of its samples taken at a rate which is roughly proportional to the size of the band. This is the content of the famous Shannon sampling theorem (more precisely its 2D extension). On the other hand, the sampling distance can be taken as a measure of the resolution provided by the restored image so that one concludes that the restoration method discussed above does not produce an improvement of the resolution of the imaging system. It certainly produces an improvement of the quality of the image which makes possible a visual verification of that resolution.

In many circumstances, the term super-resolution is used for describing methods which allow an improvement of resolution beyond the limit provided by the sampling theorem. Therefore, a super-resolving method must produce a restored image with a band broader than that of the imaging system. Therefore, super-resolution is related to the problem of out-of-band extrapolation.

Such an approach was already proposed in the 1960s, when it was observed that the Fourier transform of an object with a finite spatial extent (all objects have this property) is analytic; then the theorem of unique analytic continuation implies that the Fourier transform of the object can be determined everywhere from its values on the band of the imaging system.

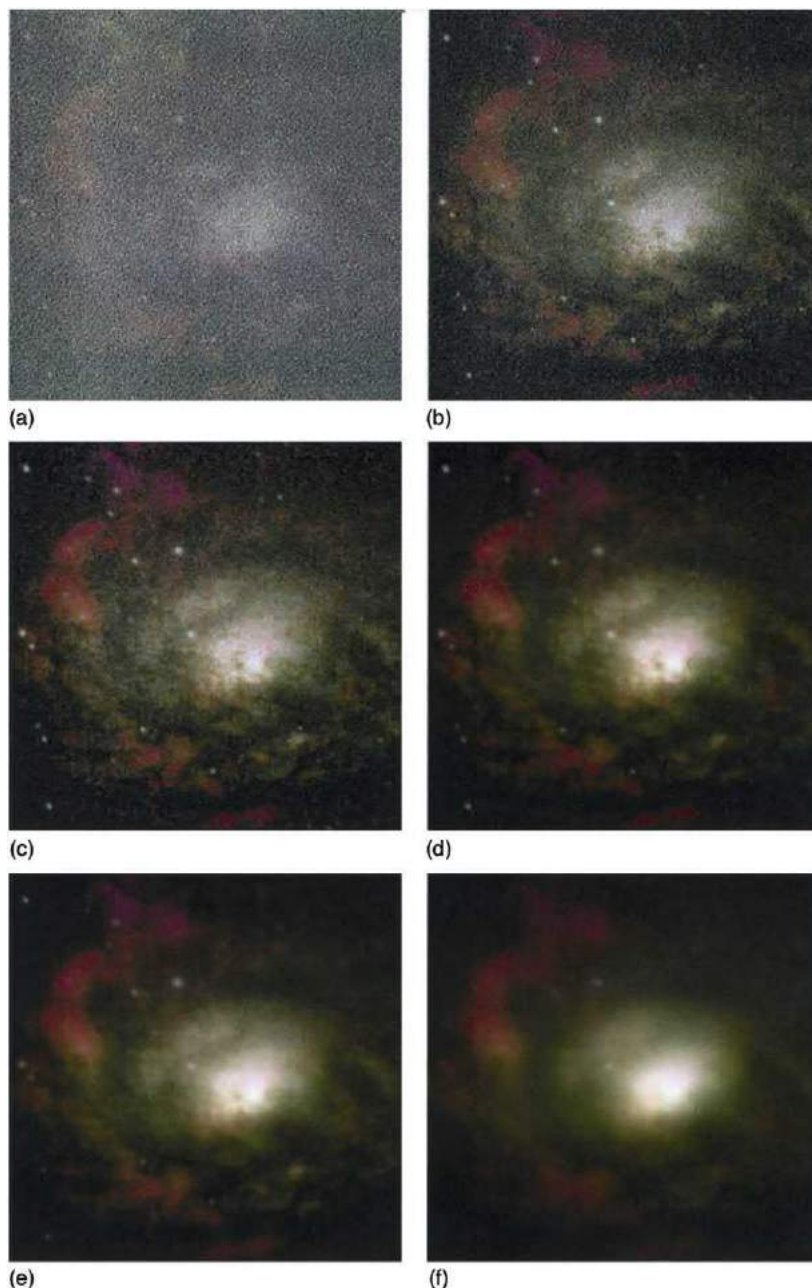


Fig. 2 Illustrating the effect of the regularization parameter. The blurred image is that shown in Fig. 1(c). The sequence is obtained by increasing the value of the regularization parameter: (a) $\mu=0$ (inverse filter); (b) $\mu=0.001$; (c) $\mu=0.01$; (d) $\mu=0.05$; (e) $\mu=0.1$; (f) $\mu=0.5$. It is evident that the best restoration is that shown in (c).

Unfortunately this problem is ill-posed, so that out-of-band extrapolation is not feasible in practice. Only recently it has been recognized that a considerable amount of super-resolution is possible when the spatial extent of the object is not much greater than the resolution distance of the instrument. For example, super-resolving methods could be used for detecting unresolved binary star in astronomical images.

The important point is that a super-resolving method must implement explicitly the finite extent of the object: the domain of the object must be estimated and the method must search for a solution which is zero outside this domain. This introduces an important question in object restoration and, more generally, in the theory of inverse problems, namely the design of regularized solutions satisfying additional conditions (constraints). This question is also important from the theoretical point of view: the problem is ill-posed because of insufficient information on the object as transmitted by the imaging system; the use of additional constraints reduces the class of the solutions which are compatible with the data and therefore can improve the restoration. This is the use of a priori information in the solution of inverse problems.

A constraint which has been widely investigated and used is the positivity of the solution, i.e., all the values of the restored image must be positive or zero. The physical meaning of the constraint is obvious. Its beneficial effect is to reduce ringing artifacts which affect the regularized solutions previously discussed, in cases where the object contains bright spots over a black background or sharp intensity variations from zero to some positive value. The ringing is essentially related to the well-known Gibbs effect in the truncation of the Fourier series.

The requirement of positivity is the requirement of a particular lower bound on the values of the solution. Therefore, one can consider more general constraints in requiring a given lower and/or upper bound on the values of the solution, eventually combined with a constraint on the domain to produce, for instance, a super-resolved positive solution. The most general form of these constraints requires that the solution belongs to a given closed and convex set C in some functional space. A general method producing regularized solutions in a convex set is an iterative method which is essentially a gradient method for the minimization of the least-squares functional, with a projection on the set C at each iteration. In the case of object deconvolution, the least-squares functional is that given in Eq. (4) and the method has the following form: if f_k is the result of the k -th iteration, then f_{k+1} is given by

$$f_{k+1} = P_C[f_k + \tau K^T * (g_r - K * f_k)] \quad (8)$$

where P_C is the convex projection onto the set C , τ is a relaxation parameter and $K^T(x) = K(-x)$. In general, the algorithm, known as the projected Landweber method, is initialized with $f_0 = 0$. An important property is that it has a regularization effect, in the sense that iterations must not be pushed to convergence but stopped after a suitable number of iterations to avoid strong noise contamination. In Fig. 3 we compare, in a specific example, the result obtained by means of the linear regularization method and that obtained by means of the projected iterative method with a lower and upper bound on the solution. The reduction of ringing is evident.

Among the methods producing positive solutions we must mention the most popular one in astronomy, the so-called Richardson–Lucy method, which is an iterative method for Maximum-Likelihood estimation. If the PSF has been normalized in such a way that the sum of all its values is one, it takes the following form:

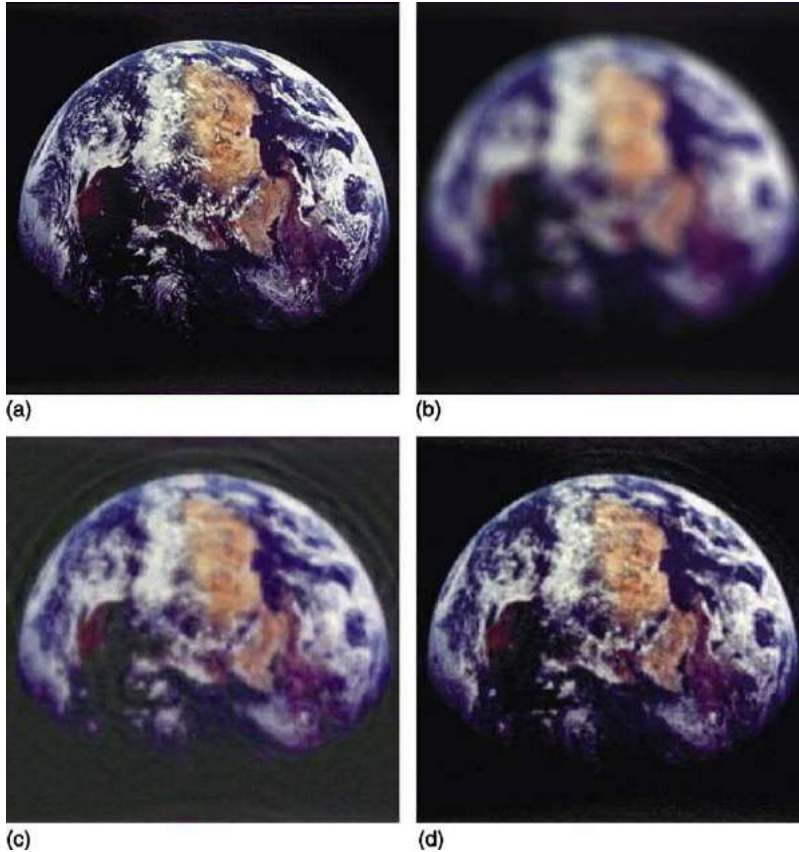


Fig. 3 Illustrating the effect of constraints on image restoration: (a) Image of the Earth taken during the Apollo 11 mission; (b) blurred image of (a) obtained by assuming out-of-focus blur and adding noise; (c) optimal restoration provided by the regularization method; the ringing around the boundary of the Earth is evident; (d) optimal restoration obtained by means of the projected Landweber method with a lower and an upper bound on the solution at each iteration. Part (a) courtesy of NASA.

$$f_{k+1} = f_k \left(K^T * \frac{g_r}{K * f_k} \right) \quad (9)$$

The method is normally initialized with a uniform image and it must not be pushed to convergence to avoid noise amplification. Another algorithm suggested by statistical arguments and producing positive solution is the so-called Maximum Entropy Method (MEM).

The image restoration methods described above have wide applications both in microscopy and astronomy. For instance, the minimization of the functional Eq. (6), with the additional constraint of positivity, has been used for deconvolving images in microscopy to reduce the effect of the missing cone. On the other hand, the Richardson–Lucy method was widely used for deconvolving the images of HST before installation of corrective optics.

Nowadays, image deconvolution is becoming more important for the ground-based telescopes equipped with adaptive optics. Indeed, even if adaptive optics is able to provide a considerable compensation of the atmospheric blur, a further improvement can be obtained by deconvolving the detected images. The PSF is provided by the image of a suitable guide star.

Finally, it is worth mentioning the Large Binocular Telescope, in construction on the Mount Graham in Arizona, because this instrument requires image restoration methods for a full exploitation of its imaging properties. The telescope (Fig. 4) consists of two 8 m mirrors on a common mount: both mirrors are equipped with adaptive optics and are combined interferometrically to reach the resolution of a 22.8 m mirror in the direction of the baseline. The telescope can be rotated to record images of the same astronomical target with different orientations of the baseline, i.e., the line joining the centers of the two mirrors. Finally the images must be processed by means of multiple images deconvolution methods to obtain a unique high-resolution image, equivalent to that produced by a 22.8 m mirror.

The applications of inverse problems with major social impact are in the area of medical imaging. The spectacular success of X-ray computerized tomography (CT), invented by GH Hounsfield at the beginning of the 1970s, has stimulated the development of other imaging modalities, based essentially on the same principle, such as positron emission tomography (PET) and single photon emission computerized tomography (SPECT). Computed images are also provided by magnetic resonance imaging (MRI). In such a case, however, the computational problem is rather simple since it consists essentially in Fourier transform inversion.

The basic principle of X-ray CT is described as a finely collimated source S (see Fig. 5(a)) emitting a pencil of X-rays which propagates through the body along a straight line L up to a well collimated receiver R . If we know both the intensity I_0 of the source and the intensity I detected by the receiver, the logarithm of the ratio I_0/I is the integral of the linear attenuation function (roughly proportional to the density function of the body) along the line L . If the source and the receiver are moved along two parallel lines, having direction θ and defining the plane Π (in practice a planar slice of the body under consideration), one obtains the integrals of the linear attenuation function $f(x)$ along all parallel lines orthogonal to θ . The scanning variables in the plane Π are defined in Fig. 5(b): φ is the angle formed by the unit vector θ with the x -axis and s is the signed distance between the origin and the integration line L . The integrals along all parallel lines orthogonal to θ define a function of s which is called the projection of $f(x)$ in the direction θ :

$$P_\theta(s) = \int_{-\infty}^{+\infty} f(s \cos \varphi - u \sin \varphi, s \sin \varphi + u \cos \varphi) du \quad (10)$$

The set of the projections of $f(x)$ for all possible θ is called the Radon transform of $f(x)$, in honor of the mathematician Johann

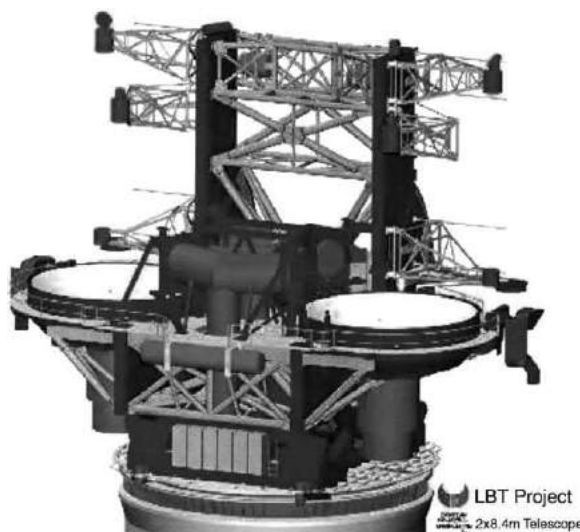


Fig. 4 Picture of the Large Binocular Telescope (LBT) in construction on the Mount Graham in Arizona. It consists of two mirrors on the same mount. The images of the two mirrors are combined interferometrically to produce a high-resolution image in the direction of the line joining the centers of the mirrors (baseline).

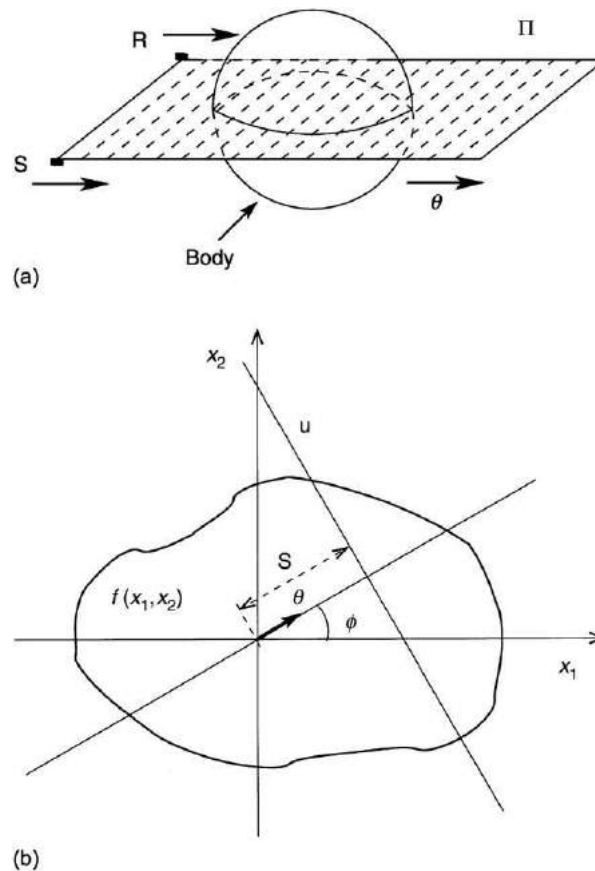


Fig. 5 (a) Scheme of the scanning procedure in X-ray tomography: the source(S)–receiver(R) pair is moved along a direction θ orthogonal to the line S–R, defining the plane Π to be imaged; when the scanning has been completed the system is rotated by a certain angle, the scanning procedure is repeated, and so on. (Reproduced with permission from Bertero M and Boccacci P (1998) *Introduction to Inverse Problems in Imaging*. Bristol, UK: IOP Publishing.) (b) Definition of the variables used in the representation of the Radon transform; φ is the angle formed by the S–R line with the x_1 -axis; s is the signed distance between the origin and the integration line (orthogonal to θ); and u is the coordinate on the integration line.

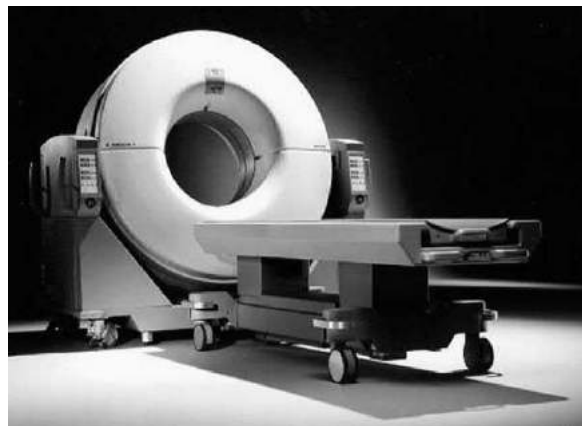


Fig. 6 Picture of a scanner for X-ray tomography. The annular part contains the acquisition and scanning systems.

Radon, who first investigated the problem of recovering a function of two variables from its line integrals. Today, such a problem is known as Radon transform inversion or object restoration from projections. Sampled values of the Radon transform, i.e., sampled values of the projections for a number of possible directions between 0 and π , are the outputs of the acquisition system of the medical equipment known as CT-scanners (Fig. 6). Their data-processing system contains the implementation of an algorithm for Radon transform inversion and the final result is a set of computed images providing maps of slices of the human body.

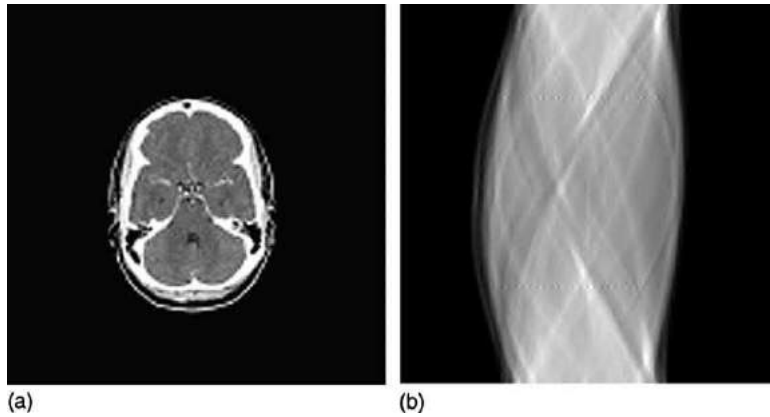


Fig. 7 (a) Picture of a brain slice; (b) the corresponding sinogram. As explained in the text, the sinogram is a gray-level representation of the Radon transform in the s, φ plane. Each horizontal line in (b) corresponds to a projection of (a).

The plot of the Radon transform of f in the (s, φ) -plane, obtained by representing its values as gray levels, is called the sinogram of f . In Fig. 7, we give the image of the sinogram of a brain slice. The horizontal rows provide representations of the projections of the phantom. In the image, the angle φ defining the direction θ takes values between 0 and 2π while the variable s takes values between $-a$ and a , if a is the radius of a disc containing the phantom. As is evident, the sinogram is symmetric with respect to the point $(0, \pi)$. Each point in the rectangle corresponds to a straight line crossing the phantom and the straight lines through a fixed point of the phantom describe a sinusoidal curve in the (s, φ) -plane. This property has given rise to the name sinogram.

The sinogram is, in a sense, the image of the body slice under investigation, as provided directly by a CT-scanner. However, an inspection of this image does not easily provide information about the body. Therefore, it must be processed to obtain a more significant image. Radon transform inversion is an example of inverse and ill-posed problem and its solution must be treated with extra care.

The crucial point in Radon transform inversion is the Fourier slice theorem, whose content is the following: the Fourier transform of the projection in the direction θ of the function f , is the Fourier transform of f on the straight line passing through the origin and having direction θ :

$$\hat{P}_\theta(\omega) = \hat{f}(\omega\theta) \quad (11)$$

This theorem clarifies the information content of CT data: each projection provides the Fourier transform of the function f along a well-defined straight line in the plane of the spatial frequencies. It is also the starting point for deriving the basic algorithm of data inversion in tomography, the filtered back-projection (FBP). It is obtained from Fourier transform inversion in polar coordinates and is a two-step algorithm: the first step is a filtering of the projections by means of a ramp-filter; the second is the back-projection of the filtered sinogram. More precisely the two steps are as follows:

- **Filtering:** for each θ the projection $P_\theta(s)$ is replaced by a filtered projection given by

$$Q_\theta(s) = \int_{-\infty}^{+\infty} |\omega| \hat{P}_\theta(\omega) e^{i\omega s} d\omega \quad (12)$$

where the multiplication by the ramp filter $|\omega|$ derives from the Jacobian of polar coordinates;

- **Back-projection:** this operation is the dual of the projection operation: indeed the projection assigns to a straight line L with coordinates (s, φ) , hence to a point of the domain of definition of the Radon transform, the integral of f along L ; on the other hand, the back-projection assigns the value of the Radon transform at (s, φ) to all points of the straight line L with the same coordinates (a pictorial representation of projection and back-projection is given in Fig. 8); as a consequence the back-projection operator $R^\#$ assigns to each point $\mathbf{x}$ the sum (integral) of all values of the Radon transform corresponding to straight lines passing through $\mathbf{x}$:

$$R^\# P_\theta(\mathbf{x}) = \int_0^{2\pi} P_\theta(x \cos \varphi + y \sin \varphi) d\varphi \quad (13)$$

The back-projection operator, when applied directly to the projections as in Eq. (13), provides a blurred image of f ; in order to get a satisfactory restoration it must be applied to the filtered projections $Q_\theta(s)$. This point is illustrated in Fig. 9.

We point out that the ramp filter introduced in Eq. (12) amplifies the high frequency components of the projections hence the noise corrupting these components. This remark clarifies that Radon transform inversion is an ill-posed problem and requires some kind of regularization. This is obtained by introducing an additional low-pass filter which reduces the effect of the high frequencies. The most frequently used combination of ramp and low-pass filter is the so-called Shepp–Logan filter.

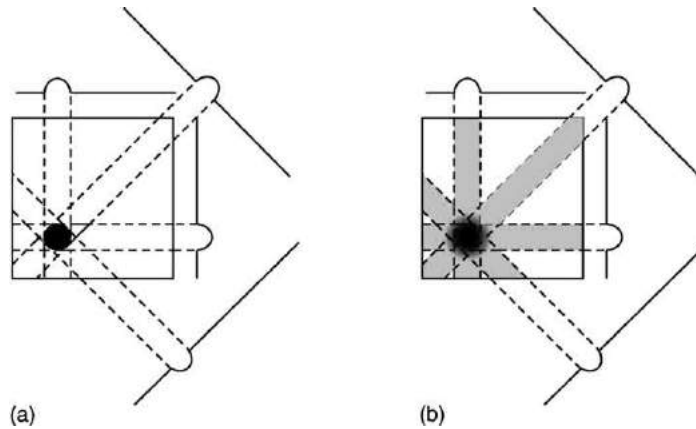


Fig. 8 (a) The effect of the projection operation is shown by drawing the projections of a circular phantom corresponding to three directions. (b) shows the result obtained by applying the back-projection operation to the three projections given in (a). It is evident that the picture provided by this operation has a maximum on the domain of the circular phantom.

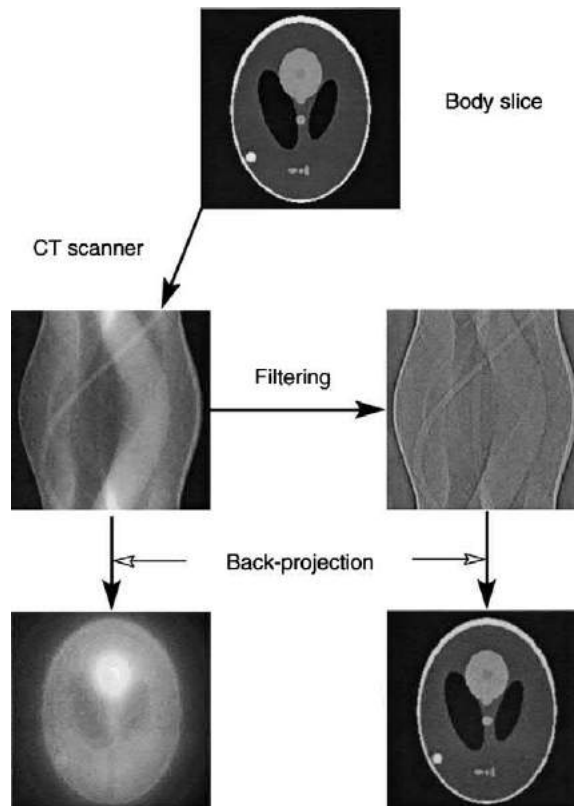


Fig. 9 Illustrating the effect of the filtered back-projection. Shown here is: a slice of the so-called Shepp–Logan phantom; the corresponding sinogram (indicated by the arrow CT scanner); the filtered sinogram (indicated by the arrow filtering); the restorations obtained by applying the back-projection operation both to the original sinogram and to the filtered sinogram.

X-ray tomography is also called transmission computerized tomography (TCT) because the image is obtained by detecting the X-rays transmitted by the body. It provides information about anatomical details of human organs because the map of the linear attenuation function is essentially the map of the density of the tissues.

A different type of information is obtained by the so-called emission computerized tomography (ECT) which is based on the administration of radio-nuclide-labeled agents known as radio-pharmaceuticals. Their distribution in the body of the patient depends on factors such as blood flow, metabolic processes, etc. Then a map of this distribution is obtained by detecting the γ -rays produced by the decay of the radio-nuclides. Therefore, ECT yields functional information, in the sense that the images produced by ECT show the function of the tissues of the organs.

Two different modalities of ECT are usually considered:

- single photon emission computerized tomography (SPECT), which makes use of radio-isotopes where a single γ -ray is emitted per nuclear disintegration;
- positron emission tomography (PET), which makes use of β^+ -emitters, where the final result of a nuclear disintegration is a pair of γ -rays, propagating in opposite directions, produced by the annihilation of the emitted positron in the tissues.

In both cases it is necessary to detect the γ -rays coming from well-defined regions of the body and, to this purpose, different methods are used in each case.

In SPECT, the discrimination of the γ -rays is obtained by means of a collimator, consisting of holes in a large lead slab covering the crystal detector face. In PET the collimation is obtained by means of pairs of detectors in coincidence. This technique, which is also called electronic collimation, is more accurate than the physical collimation used in SPECT and allows the design of systems with a great efficiency.

The basic reconstruction method in TCT, namely FBP, is often used in the reconstruction of SPECT and PET data. However, several corrections are necessary in practice. The principal ones are due to the collimator blur and to the scattering of photons in the body. If one develops a more accurate model of data acquisition by taking into account these effects, then the restoration problem implies the inversion of a very large matrix which is sparse and ill-conditioned. To this purpose iterative regularization methods are used such that the basic operation required at each step is matrix-vector multiplication. Other imaging modalities, which have been proposed for medical applications, are electrical impedance and microwave tomography. The underlying inverse problems are basically nonlinear so that the computational cost of the methods for their solution is, in general, too high for clinical applications. However the research in these areas is very active and, therefore, the situation could be improved in the near future.

Further Reading

- Bertero, M., 1989. Linear inverse and ill-posed problems. In: Hawkes, P.W. (Ed.), *Advances in Electronics and Electron Physics*, vol. 75. New York: Academic Press.
- Bertero, M., Boccacci, P., Boccacci 1998. *Introduction to Inverse Problems in Imaging*. Bristol, UK: IOP Publishing.
- Bertero, M., De Mol, C., 1996. Super-resolution by data inversion. In: Wolf, E (Ed.), *Progress in Optics*, vol. XXXVI. Amsterdam: Elsevier, pp. 129–178.
- Engl, H.W., Hanke, M., Neubauer, A., 1996. *Regularization of Inverse Problems*. Dordrecht, Germany: Kluwer.
- Groetsch, C.W., 1993. *Inverse Problems in Mathematical Sciences*. Braunsschweig, Germany: Vieweg.
- Herman, G.T. (Ed.), 1979. *Image Reconstruction from Projections*. Berlin: Springer.
- Herman, G.T., 1980. *Image Reconstruction from Projections. The Fundamentals of Computerized Tomography*. New York: Academic Press.
- Huang, T.S. (Ed.), 1975. *Picture Processing and Digital Filtering*. Berlin: Springer.
- Kak, A.C., Slaney, M., 1988. *Principles of Computerized Tomographic Imaging*. New York: IEEE Press.
- Krestel E (ed.) (1990) *Imaging Systems for Medical Diagnostics*. Berlin: Siemens Aktiengesellschaft.
- Natterer, F., 1986. *The Mathematics of Computerized Tomography*. New York: Wiley.
- Tikhonov, A.N., Arsenin, V.Y., 1977. *Solutions of Ill-Posed Problems*. Washington, DC: Winston/Wiley.

Introduction

The image quality of a perfect optical system is ideally limited by its diffraction figure. In the real world there are many effects which degrade that physical limit, resulting in poorer image quality at the focal plane. These effects may be static (e.g., time-independent, as, for example, error in the alignment of the optical train) or dynamic (e.g., variable with time, such as optical misalignment introduced by thermo-mechanical distortion). Active and adaptive optics, the topics of this article, are technologies able to correct for the optical aberrations induced by dynamical effects. There are a variety of causes liable for dynamical image degradation, among these time-variable optical misalignments, pointing and tracking errors (for motorized instruments), and atmospheric perturbations. In practice, both atmospheric turbulence and misaligned instrument introduce random spatial and temporal perturbations on the optical beam, distorting the spherical (ideal) wavefront incoming on to the focal plane. This leads to loss of both the diffraction limited condition and image resolution. A typical situation for astronomical application is shown in Fig. 1. A target (a star in this case) emits light whose wavefront, being far from the observing system, can be considered plane. When this wavefront passes through the atmosphere it is distorted with a temporal frequency of the order of hundreds of Hertz. This dynamically distorted wavefront fills the aperture of the optical system and will once more be distorted by the dynamic misalignment of the optical system itself. The final wavefront will be deformed with respect to the ideal (spherical) wavefront.

Dynamic optical misalignment may depend on gravitational and thermal deformation, while pointing error may appear as very slow (drift), slow (wander), or fast (jitter) tilt of the image on the focal plane. Atmospheric perturbations are produced by air turbulence and may be natural (as in the astronomical or satellite tracking cases) or induced by the experiment itself

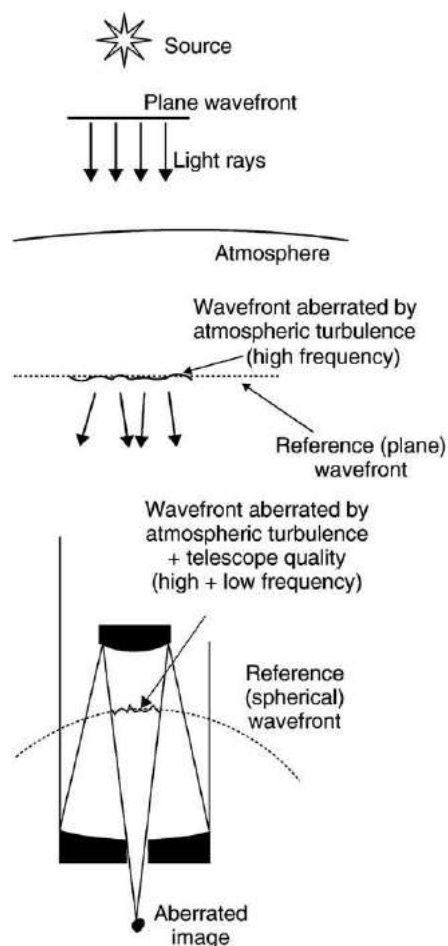


Fig. 1 Dynamical wavefront aberration.

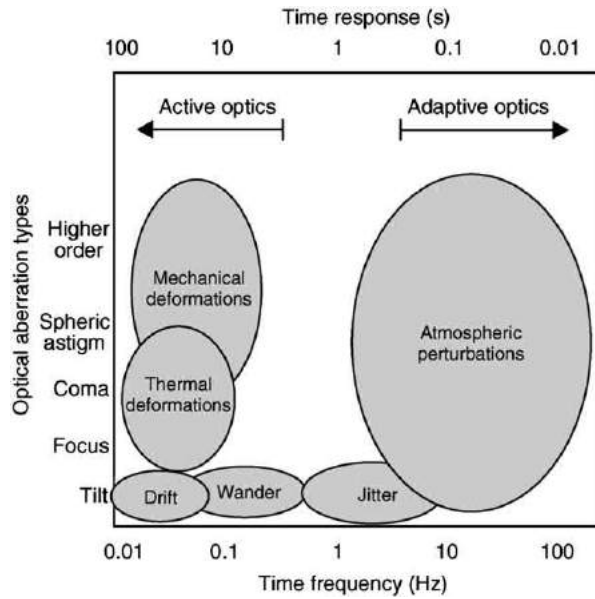


Fig. 2 Wavefront aberration time-scale and associated aberrations.

(thermal blooming in high power laser applications). All these effects appear on very different time-scales, as shown in [Fig. 2](#). As noticeable, atmospheric turbulence introduces aberrations of all orders, at a frequency above about 1 Hz. Pointing and tracking errors introduce an image tilt with different time-scale. Gravitational effects induce mechanical deformation of the optics, mainly in astronomical applications because of the massive components (the telescope primary mirror in particular). Thermal deformation produces optical misalignment among optical components mainly resulting in defocus. [Fig. 2](#) is purely illustrative, because some effects may produce wavefront aberrations at different frequencies than indicated there. For example, drift may be faster than indicated in the figure, and thermal blooming (a thermal effect) introduces aberrations at a higher frequency, and so on.

The correction of the wavefront errors is usually made by means of devices able to change the phase wave somewhere in the light path. Though the concept for correcting aberrated wavefront at slow or high frequency is the same, for example, the use of one or more deformable elements in the optical path, the technologies developed for each area are quite different: variations at slow frequency (approximately below 0.1 Hz in [Fig. 2](#)) are indicated as quasi-static errors and the techniques developed to overcome them are properly called active optics, while adaptive optics indicates the techniques to overcome the unwanted effects at high frequency (above about 1 Hz in frequency). The frequency value dividing the active from adaptive optics is not well defined and depends on the field of application and its scientific engineering communities. The important point is that a correct use of a combination of the two techniques may ultimately lead to bringing the imaging system close to its diffraction limit.

Slow Corrections: Active Optics

The analysis of the wavefront is required before its correction, and one or more deformable/movable mirrors and their control system, to operate the corrections. A typical system for low frequency correction is shown in [Fig. 3](#). A telescope contains three mirrors (M1, M2, and M3) and an instrument to detect its scientific target, usually placed on its optical axis. The dynamic deformations present in this kind of instrument are mainly third-order spherical (thermal deformation of M1), third-order astigmatism (a saddle-horse deformation of M1 due to gravity when the telescope is pointing away from the zenith), and third-order coma (due to gravitational decentering between M1 and M2). Since a telescope has a tracking system some pointing error may be present, causing motion of the image in the focal plane. The observation of an off-axis target (typically a star for astronomical applications) is made to measure the deformations of the incoming wavefront by means of a wavefront sensor (WS) system. This step is completed using some of the wavefront sensing technique described elsewhere in this encyclopedia. The WS must measure the information on the wavefront distortion with adequate spatial resolution to compensate for aberrations across the full aperture of the system. Once the wavefront distortion has been measured by the WS, the needed corrections are sent to the control system which apply the appropriate mirror deformation/movements. In the example of [Fig. 3](#) only the primary mirror (entrance pupil) is deformable and it is used for correcting spherical and astigmatism aberrations. M2 is used to compensate the decentering coma by means of lateral displacement. M3 may be tilted to correct for low-frequency pointing errors. In astronomical applications active optics is essential for telescopes with primary-mirror diameters larger than approximately four meters which cannot be manufactured rigidly at reasonable cost. Active optics requires a thin deformable or a segmented primary mirror maintained on the correct shape at every elevation. Using this method, low-frequency correction has been achieved to date in telescopes with primary mirrors of the order of ten meters.

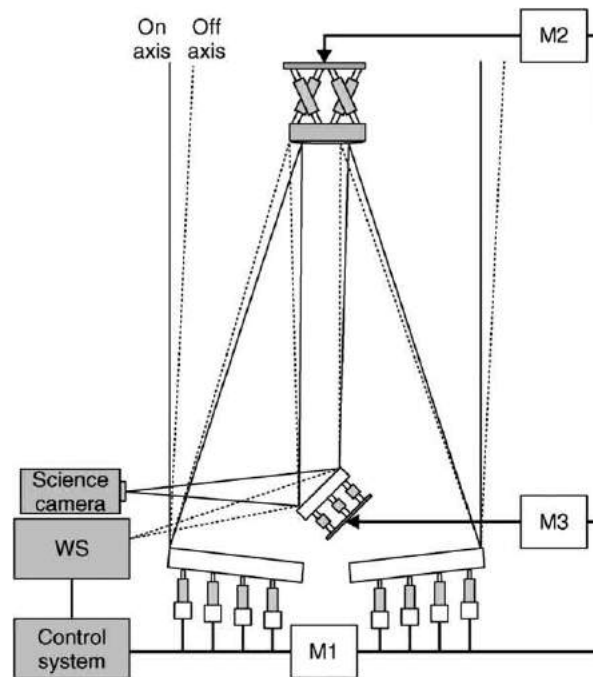


Fig. 3 A typical system for correcting low-frequency dynamical aberrations (active optics).

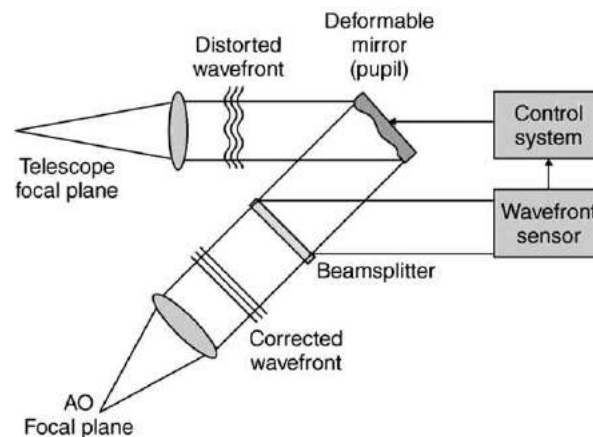


Fig. 4 A typical setup for correcting high-frequency dynamical aberrations (adaptive optics).

Fast Corrections: Adaptive Optics

It is clear that corrections at high frequency are impossible using the approach discussed above, because of the large dimensions and weight of the mirror and the low response time required. For an adaptive correction it is necessary to image a smaller system pupil in the optical train and make the active correction of the wavefront. A typical configuration is shown in Fig. 4. The light from the telescope focal plane (corrected for low frequency errors by means of active optics) is collimated and a pupil is located in this beam. Behind the active mirror a beamsplitter is used to send part of the light into the WS and measure wavefront distortions. The analysis results are sent to the control system which performs the active mirror deformation for the wavefront compensation. If the WS and the control system are sufficiently fast (typically faster than the time-variations of the wavefront distortion), this system is able to produce a plane wavefront which can be focused at the diffraction limit of the system. For this reason adaptive optics are often referred to as real-time system.

The seminal paper on the adaptive optics concept was written by H.W. Babcock of the Mount Wilson and Palomar observatories in 1953. Since then a great improvement has been made, especially in military and astronomical fields. In recent years, however, many applications of adaptive optics have developed in fields different from traditional ones, such as precision manufacturing by means of high-power lasers, telecommunications, aero-optics, medical physics, and ophthalmology.

We describe adaptive optics with particular reference to astronomical applications. We provide a brief description of other applications at the end of this article.

Imaging Through the Atmosphere

The optical effects of the atmospheric turbulence are due to local temperature variations that induce fluctuations in the refractive index of air. The refractive index changes depend on the air density as well as the range of the temperature variations. Air density is highest at sea level and decreases exponentially with altitude. Optical effects of turbulence decrease with the altitude of the site, the reason for locating astronomical telescopes on mountains. The effects of turbulence on a point source image are depicted in Fig. 5. Hereafter we consider an ideal telescope, for example, one without need for active correction at low frequencies. We consider the source 1 (continuous line). The radiation emitted by the source is a spherical wavefront that can be treated as a plane, because of the very long distance to the Earth's atmosphere.

In absence of wavefront distortion, the angular diameter of the image (Airy disk), is approximately $2.44\lambda/D$, with λ the wavelength of the observation and D the telescope aperture (the mirror diameter). In this case the image is said to be diffraction limited, and it is the best possible performance for the observation. This is the performance of a space-based ideal telescope.

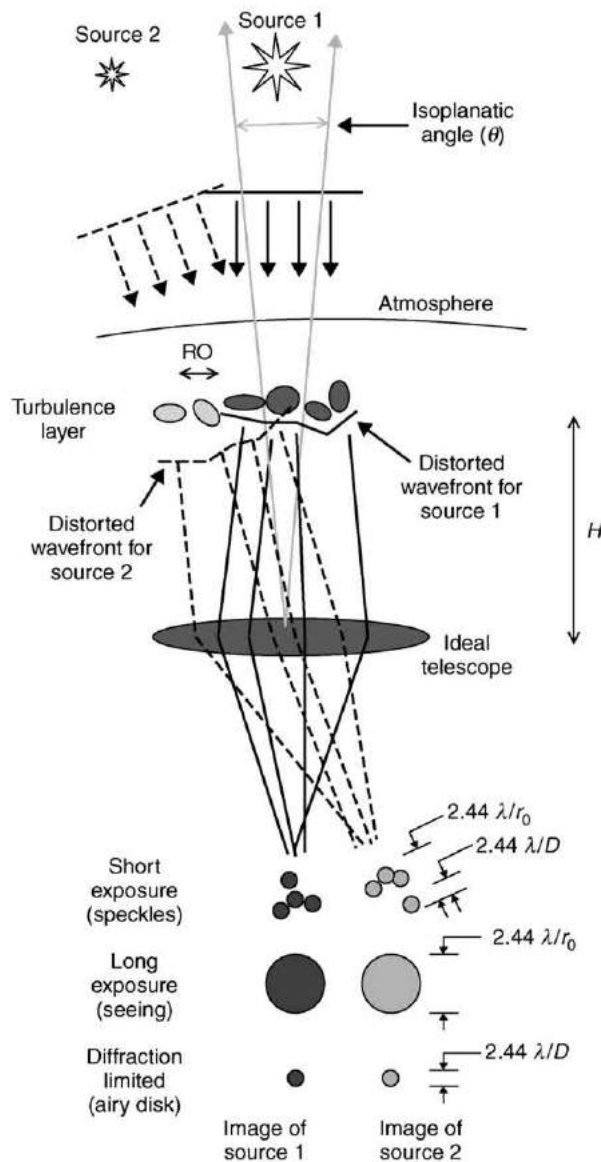


Fig. 5 Imaging through the atmosphere.

In the presence of the atmosphere the incoming wavefront is distorted by the turbulence and when it reaches the telescope aperture it is no longer a plane (see in Fig. 5). The so-called Fried's parameter r_0 (also called atmospheric coherence length) is the most common parameter used to quantify the distortion effects of atmospheric turbulence. It is defined as the distance over which the wavefront phase is highly correlated (in this way it provides a measure of the turbulence scale size). For an undistorted plane wave, r_0 is infinity, while for atmospheric distortion it typically varies between 5 cm (bad sites) and 20 cm (excellent sites). We depict Fried's parameter schematically in Fig. 5, as bubbles of dimension r_0 in the turbulent layer (referred to as Fried's cells). r_0 is a statistical parameter, which depends on time and wavelength ($r_0 = r_0(t, \lambda^{6/5})$). Its variability with time may be considerable, sometimes changing by a factor of 2 within a few seconds. A small telescope, with a diameter less than r_0 , will always operate at this diffraction limit. When the r_0 is smaller than the telescope aperture, the angular size of the image is different if we consider long- or short-time exposures. If the exposure time is very short (less than hundredths of a second) it is possible to image a structure in the focal plane (see short exposure image in Fig. 5). This structure, which varies with time, is due to high-order wavefront distortions that occur within the aperture. The short exposure image is broken into a large number of speckles of diameter $2.44\lambda/D$, which are randomly distributed over a circular region of approximately $2.44\lambda/r_0$ in angular diameter. A lower-order wavefront distortion is correlated with the size of the largest disturbances produced by atmospheric turbulence (the so-called outer scale) which varies considerably (from 1 m to over 1 km). Such disturbances produce a time-dependent tilt of the overall wavefront that causes fluctuations in the angle of arrival of the light from a star and randomly displace the image. For D of the order of the Fried parameter r_0 , image motion dominates, and a simple tip-tilt mirror (not just deformable) provides useful correction. For D greater than r_0 , speckles dominate and a deformable mirror is necessary to account for the wavefront distortion. For long exposures the dimension of the image is determined by the ratio λ/r_0 rather than $2.44\lambda/D$. The angular resolution is diminished due to the atmospheric turbulence (see Fig. 5). For a 4 m telescope, atmospheric distortion degrades the spatial resolution by more than one order of magnitude compared to the diffraction limit, and the intensity at the center of the star image is lowered by a factor of 100 or more. Even at the best astronomical sites, ground-based telescopes observing at visible wavelengths cannot, because of atmospheric turbulence alone, achieve an angular resolution better than telescopes of about 20 cm diameter.

Most of the distortions imposed by the atmosphere are in the wave-phase and an initially plane wavefront traveling 20 km through the turbulent atmosphere accumulates, across the diameter of a large telescope, phase errors of a few micrometers.

Turbulence is distributed along the propagation path through the atmosphere, and then the wavefront errors become uncorrelated as the field angle increases. This effect is known as angular isoplanatism and is one of the major limitations to the astronomical adaptive optics. The structure becomes completely uncorrelated out of the isoplanatic angle, approximately equal to r_0/H , where H is the altitude of the turbulence layer. This decorrelation is sketched in Fig. 5, where the short exposure images for objects 1 and 2 have different speckles structure.

In Earth's atmosphere significant turbulence occurs at altitudes up to 20 km, often with large peaks in the vicinity of the tropopause, at around 10 km there exist, in practice, more than one parameter H in Fig. 5, one for each turbulence layer. This three-dimensional distribution of the turbulence considerably restricts the angle over which adaptive compensation is effective and leads to the design of multiconjugate adaptive optics, as we next explain.

Adaptive System Design

As stated above, an active system is composed of at least one wavefront sensor and one deformable mirror. Typically, the wavefront is corrected by making independent tip, tilt, and piston movements of the deformable mirror surface, as shown in Fig. 6. As a rule of thumb, a deformable mirror used for atmospheric correction should have one actuator for each Fried's cell, such that the total number of actuators should be equal to the number of Fried's cells filling the telescope aperture, i.e., D^2/r_0^2 . For instance, a near-perfect correction for an observation in visible light ($0.6 \mu\text{m}$) with an 8 m telescope placed in a good astronomical site ($r_0=20 \text{ cm}$), would require a deformable mirror with approximately 6400 actuators in its adaptive optics system. Recalling that r_0 is strongly dependent on the wavelength (i.e., proportional to $\lambda^{6/5}$), similar performance in the infrared at $2 \mu\text{m}$ requires only 250 actuators. This is one reason why adaptive correction is easier in the infrared. A large number of Fried's cells requires a similarly large number of subapertures in the wavefront sensor. Commonly used WSs are based on modifications of the classic Hartmann screen test or of curvature type. In the former case the adaptive device is made with individual piezoelectric actuators (see Fig. 6), while in the latter the correction is achieved with a bimorph adaptive mirror, made of two bonded piezoelectric plates. With both

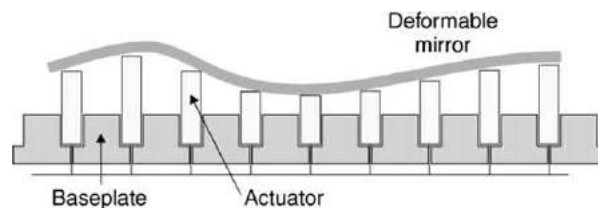


Fig. 6 A sketch of a deformable mirror.

methods, wavefront sensing is done on a reference off-axis object (Fig. 3), or even on the observed object itself (Fig. 4) if it is bright enough. In the former case the angular distance of the reference object must fall inside the isoplanatic angle and must be sufficiently luminous. In typical astronomical sites the isoplanatic angle at a wavelength of $0.6\text{ }\mu\text{m}$ is ~ 5 arcsec, while it can be ~ 20 arcsec at $2\text{ }\mu\text{m}$. This is an additional reason why the adaptive optics system is most efficient in the infrared range. Correction in the visible requires a reference star approximately 25 times brighter than in the infrared. Most current astronomical systems are designed to provide diffraction-limited images in the near-infrared (1 to $2\text{ }\mu\text{m}$) with the capability for partial correction in the visible. However, some military systems for satellite observations do provide full correction in the visible on at least 1 m class telescopes.

The control system is generally a specialized computer that calculates, from the WS measurements, the commands to send the actuators of the deformable mirror. The calculation must be performed quickly (within 0.5 to 1 ms), otherwise the state of the atmosphere may have changed making the wavefront correction inaccurate. The required computing power can exceed several hundred million operations for each command set sent to a 250 actuator deformable mirror. As in active optics systems, zonal or modal control methods are used. In zonal control, each zone or segment of the mirror is controlled independently by wavefront signals that are measured for the subaperture of that zone. In modal control, the wavefront is expressed as the best-fit linear combination of polynomial modes of the wavefront distorted by the atmospheric perturbations.

Future Development

As explained above, a convenient method for measuring the wavefront distortion is to look for an off-axis reference star (also known as natural guide star, NGS) within the isoplanatic angle. However, it is generally impossible to find a bright reference star close to an arbitrary astronomical target. Conditions are far better in the infrared than in the visible because atmospheric turbulence (and in particular its high spatial frequencies) is, for a given AO correction, less pronounced at longer wavelengths. The spatial and temporal sampling of the disturbed wavefront can be reduced, which in turn can permit the use of fainter reference stars. Coupled with the larger isoplanatic angle in the infrared, this gives a much better outlook for AO correction than in the visible.

Nevertheless, even for observations at $2.2\text{ }\mu\text{m}$, the sky coverage achievable by this technique (equal to the probability of finding a suitable reference star in the isoplanatic patch around the chosen target) is of the order of 0.5 to 1% . It is therefore unsurprising that most scientific applications of AO have to date been limited to the study of objects which provide their own reference object. The most promising way to overcome the isoplanatic angle limitation is by the use of artificial reference stars, also referred to as laser guide stars (LGS). These are patches of light created by the back-scattering of pulsed laser light of sodium atoms in the high mesosphere or by molecules and particles located in the low stratosphere. The laser beam is focused at an altitude of about 90 km in the first case (sodium resonance) and 10 to 20 km in the second case (Rayleigh diffusion). Such an artificial reference star can be created as close to the astronomical target as desired, and a wavefront sensor measuring the scattered laser light is used to correct the wavefront aberrations on the target object.

Military laboratories have reported the successful operation of adaptive optics devices at visible wavelengths using laser guide star on both a 60 cm telescope and a 1.5 meter telescope, achieving an image resolution of 0.15 arcsec . Nevertheless, there are still physical limitations of a LGS. The first problem, focus anisoplanatism, or cone-effect, was realized long ago. Since the artificial star is created at a relatively low altitude, back-scattered light collected by the telescope forms a conical beam, which does not cross the same turbulence-layer areas as light from the astronomical target. This leads to a phase estimation error, which in principle can be solved by the simultaneous use of several laser guide stars surrounding the targeted object. A more significant problem is image motion or tilt determination. Since the paths of the outgoing and returning light rays are the same, the centroid of the artificial light appears to be stationary in the sky, while the apparent position of an astronomical source suffers lateral motions (also known as tip/tilt). The simplest solution is to supplement the AO system and LGS with a tip/tilt corrector set on a (generally) faint close NGS. Performance is then limited by poor photon statistics in the tip/tilt correction.

Adaptive optics with a multicolor laser probe is another concept that has been studied to solve the tilt determination problem of laser beacon based AO. Only applicable to sodium resonant scattering at 90 km , it excites different states of the sodium atoms and makes use of the slight variation in the refraction index of air with wavelength. Its main drawback is the limited returned flux, due to the saturation of mesospheric sodium layer.

The quality of a site for adaptive optics is determined not only by the median seeing, but also by the height and thickness of the turbulent layers (often only a few thin layers dominate). Some adaptive-optics systems can be adapted to correct for turbulence originating at a particular height by using several reference stars. This technique is known as multiconjugate adaptive optics (MCAO) and is still in its infancy. This concept deals with a 3-D view of the atmosphere, and for this reason it is also called turbulence tomography. Several wavefront sensors are needed for tomography. Turbulence tomography is also useful to correct for the cone effect when using LGS. When several deformable mirrors are placed on the appropriate conjugate planes, as shown in Fig. 7, anisoplanatism reduces and the size of the corrected field of view increases. The situation is roughly depicted in Fig. 7, where each turbulent layer (indicated as TL) has a proper conjugated plane (C-TL).

Another challenging optical solution is one of utilizing large secondary adaptive mirrors. This application seems especially interesting at thermal wavelengths, where any additional mirror raises the significant thermal background detected by instruments.

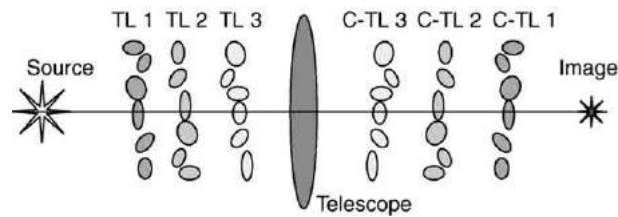


Fig. 7 Conjugated planes of turbulent layers.

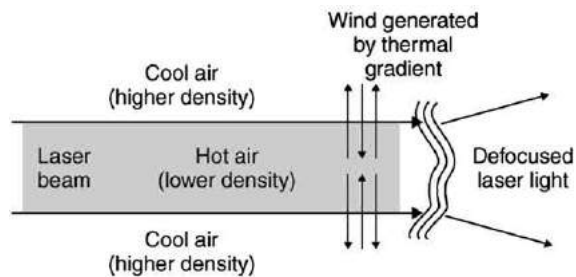


Fig. 8 Thermal blooming effect in high-power laser applications.

Other Adaptive Optics Applications

In recent years the adaptive optics technique has been developed in many applications beside astronomy and the military, for example, as in high-power laser welding or cutting, in ophthalmology and in aero-optics.

In high-power laser applications, thermal blooming takes place when laser light is absorbed by the gases and particles that make up the atmosphere. Since energy is conserved, the energy lost by the laser (its intensity diminishes) is gained by the atmosphere that heats up. This inhomogeneity at the boundary of the laser beam generates wavefront distortions by turbulence, as shown in Fig. 8. This means that laser light that passes through the air at a later time interacts with a different atmosphere than the light that passed by earlier. The net effect is that, in the case of a Gaussian laser beam, light following a wavefront which passed at an earlier time, 'sees' a concave lens that deflects the successive rays out of the beam (away from the axis) causing the beam to appear to grow in radius or 'bloom'. Adaptive optics techniques for real-time compensation of blooming in laser beams were developed in the late 1960s. Given the high coherence and monochromaticity of a laser, the technique is more straightforward than for more common white-light applications of the astronomical applications.

Density fluctuations in a compressible shear layer are sufficiently high to cause a time varying index of refraction of the layer. This situation is common in a turbulent flowfield, and the study of the wavefront aberrations induced by this flowfield, when it is of relatively short propagation length in the near-field of the optical system, is referred to as aero-optics. Turbulent flowfield induces various types of degradations in the performances of an optical system. For example the jitter and defocusing of laser-based weapons cause a reduction in the amount of energy delivered to the target. Tracking systems experience bore-site errors and imaging systems experience blurring. Recent work suggests that adaptive optics compensation may successfully overcome these problems.

In ophthalmology the adaptive technique finds application in studying the retina of the eye. In this case the eye itself produces wavefront aberration, and the retina is seen as blurred. Techniques in ophthalmic laser surgery are being developed to use adaptive optics to correct for detected distortions.

In the field of telecommunications, studies are in progress for using adaptive techniques in data transmission between ground-based stations and satellites.

See also: Nonlinear Optical Phase Conjugation. Phase Conjugation and Image Correction

Further Reading

- Beckers, J.M., 1993. Adaptive optics for astronomy: principles, performance and application. *ARAA* 31, 13–62.
 Fried, D.L., 1965. Statistics of a geometric representation of wavefront distortion. *JOSA* 55, 1427–1435.
 Hardy, J.W., 1998. *Adaptive Optics for Astronomical Telescopes*. New York: Oxford University Press.
 Jumper, E.J., Fitzgerald, E.J., 2001. Recent advances in aero-optics. *Progress in Aerospace Sciences* 37, 299–339.
 Liang, J., Williams, D.R., Miller, D.T., 1997. Supernormal vision and high-resolution retinal imaging through adaptive optics. *JOSA A* 14 (11), 2884–2892.

- Lukin, V.P., 1995. *Atmospheric Adaptive Optics*. Bellingham: SPIE Optical Engineering Press.
- Ragazzoni, R., Marchetti, E., Valente, G., 2000. Adaptive-optics corrections available for the whole sky. *Nature* 403 (6765), 54–56.
- Roddier, F., 1999. *Adaptive Optics in Astronomy*. Cambridge: Cambridge University Press.
- Roggeman, M., Welsh, B., 1996. *Imaging Through Turbulence*. Boca Raton, FL: CRC Press.
- Tyson, R.K., 2000. *Introduction to Adaptive Optics*. Bellingham: SPIE Optical Engineering Press.
- Welford, W., 1974. *Aberrations of the Symmetrical Optical System*. New York: Academic Press.
- Wilson, R.W., Jenkins, C.R., 1996. Adaptive optics for astronomy: theoretical performance and limitations. *MNRAS* 268, 39.

Phase Conjugation and Image Correction

EN Leith, University of Michigan, Ann Arbor, MI, USA

© 2005 Elsevier Ltd. All rights reserved.

Introduction

Phase conjugation of a wave is a process that has long been recognized. However, the technique for producing high-quality phase conjugation is relatively new. Also, the principal use of phase conjugation, the correction of phase errors in waves that have propagated through phase-distorting media, is relatively new, having originated in holography in the 1960s. Today, phase conjugation is an important area of modern optics. This article describes the conjugation of a wavefield by the process of holography and the holographic phase conjugation method of removing phase distortions from a wavefield, thereby permitting high-quality images to be produced from phase-distorted waves.

The basic meaning of a conjugated wave is explained with the aid of Fig. 1. Fig. 1(a) shows a wavefield propagating to the right. For simplicity, we show only one wavefront, i.e., one contour of constant phase. It is understood that a wavefield consists of a sequence of such wavefronts, with each wavefront differing in phase by 2π from the adjacent one, and separated by the wavelength λ . Of the two points, A and B, on the wavefront, we say that the wavefront at A lags that at B, since when the wavefront crosses a plane normal to the direction of travel, point A of the wavefront crosses that plane behind point B. The wavefield produces, at the plane P, a field $ae^{i\phi}$, where a is the amplitude and ϕ is the phase of the wavefront at the plane P.

We place a phase conjugator in the path of the wave. The wave passes through the conjugator, emerging as a phase conjugated wave, in which the point B, which formerly led point A in phase, now lags behind it by a corresponding amount. This illustrates the basic meaning of the term 'phase conjugation'. This device could be considered as a forward phase conjugator, since the phase conjugated wave travels in the same direction as the original wave.

Next (Fig. 1(b)) we describe a backward phase conjugator. The conjugate wave travels in a direction opposite to that of the incident wave. The conjugate wave emitted by the phase conjugator seems to be identical to the incident wave; we show the two waves to be in coincidence, or more precisely, for purposes of clarity, almost in coincidence. The points A on the two wavefronts coincide, as do the points B. As before, on the incident wave, point A lags behind point B. On the exiting wave, since the wave is traveling in the opposite direction, point B is now lagging point A, i.e., is lagging in phase. In either case, forward or backward conjugation, the wave emerging from the conjugator has the complex amplitude $a(x,y)e^{-i\phi(x,y)}$.

A phase conjugator (backward variety) is like a mirror, in that it obeys the basic mirror law, that the angle of reflection equals the angle of incidence, but in a rather different way, as Fig. 2 shows. In fact, the reflected ray travels back on itself. A simple way to generate a conjugate wave, and one that has been known for a very long time, is with the use of an array of retroreflectors, or corner cubes. As shown in Fig. 3, each ray undergoes two reflections and emerges traveling in exactly the opposite direction, although with some displacement. If we considered the third dimension, there could be three reflections. The amount of displacement depends on where the ray strikes; if it strikes near the apex of the corner cube, the ray displacement will be slight; if at the edge, the displacement is greatest. Thus, the image formed in reflection is degraded. This geometrical degradation is greater for larger corner

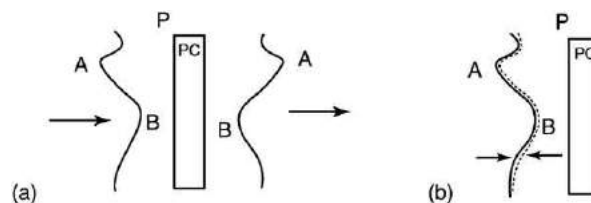


Fig. 1 A phase conjugator. (a) shows a forward phase conjugator; (b) shows a backward phase conjugator. For clarity, in (b) the conjugated wave is shown with dotted lines. PC is phase conjugator, and P is the entrance plane of the conjugator.

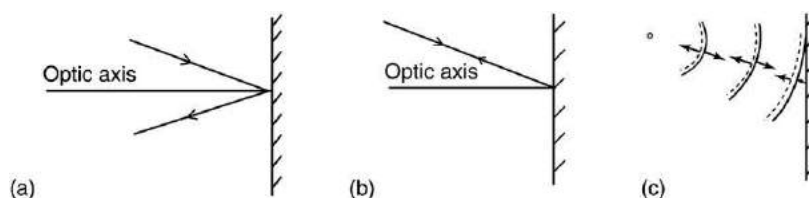


Fig. 2 Light reflected from mirrors. (a) Light falling on a conventional mirror and being reflected; (b) light reflected from a conjugating mirror; (c) as in (b), except that illustration is with wavefronts instead of rays.

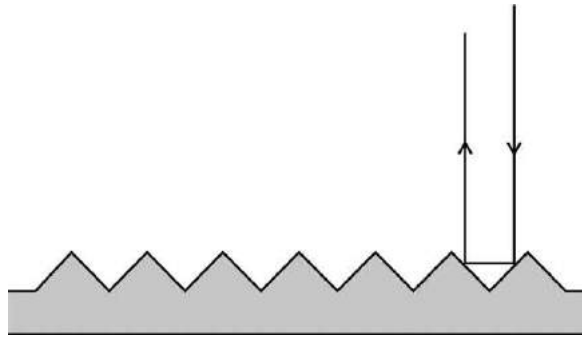


Fig. 3 A ray of light reflected from a retroreflector (or corner cube).

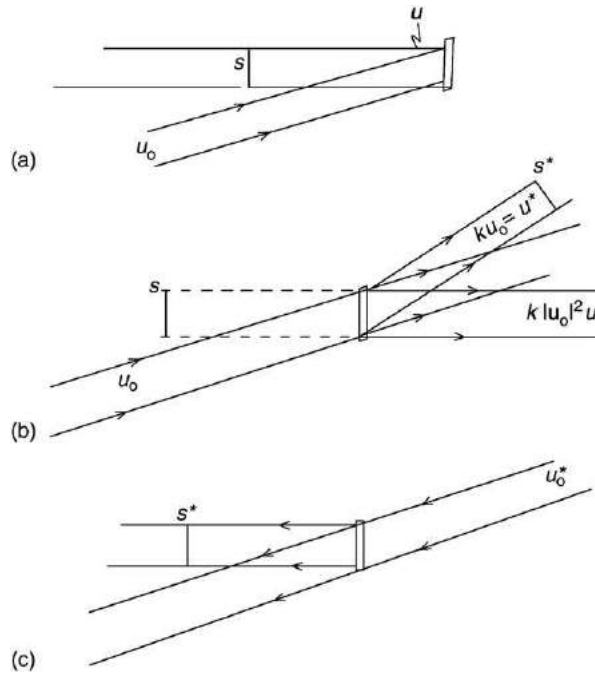


Fig. 4 The holographic process. (a) Making the hologram; (b) reading out the hologram, using a readout beam that duplicates reference beam; (c) reading out with reversal of reference beam.

cubes. Thus, why not make the cubes very small? Then, however, the image is degraded by diffraction effects, since each corner cube constitutes a small aperture. Thus, there is an optimum size for the cubes; if too large, geometrical factors degrade the image; if too small, diffraction effects degrade the image.

Phase Conjugation by Holography

With holography, however, came a method of conjugating a wavefront that yields high-resolution imagery. The holographic process inherently produces a phase conjugate image, as was noted by Gabor in his early papers on holography. The holographic process is illustrated in Fig. 4. For simplicity, we assume the object to be a planar transparency $s(x, y)$. A coherent light wave, generally a plane or spherical wave, passes through the transparency and a diffracted field u is produced at the recording plane P, as in Fig. 4(a). The field u is typically the Fresnel diffraction pattern of the object s , but if the propagation distance is sufficiently great, u could be the Fraunhofer pattern of s . Of course, lenses can be placed between the object and recording planes, thus modifying the diffraction pattern.

The field u is combined with a reference wave u_0 , which is generally an off-axis planar or spherical wave. The two beams produce an interference pattern with intensity

$$I = |u_0 + u|^2 = |u_0|^2 + |u|^2 + 2\text{Re } u_0^* u \quad (1)$$

where the final term can be written as $u_o^*u + u_o u^*$; the $*$ denotes complex conjugate. The process of forming the interference of the two beams, i.e., forming the intensity of their sum, has generated the conjugate term u^* . When the intensity is recorded, as on photographic film, for example, both terms are stored on the film. Suppose for simplicity that u_o is a plane wave $a_o e^{i2\pi f_o x}$, where a_o is a constant, and $f_o = (\sin \theta)/\lambda$, where θ is the angle that the reference wave makes with the optic axis and λ is the wavelength of the light. Also, u can be written as a product of an amplitude term a and a phase term $e^{i\phi}$, or $u = a e^{i\phi}$. Let the photographic plate, after exposure and development, have a transmittance t proportional to the intensity I , or $t = kI$ with k a constant; this is an oversimplification of the recording process, but preserves the essence of the holographic process. Now let the photographic plate, which after development becomes a hologram, be illuminated with a readout wave that, for simplicity, we let be u_o , a duplicate of the reference wave. Then, the field emerging from the hologram is

$$u_e = u_o k [|u_o|^2 + |u|^2 + u_o^* u + u_o u^*] \quad (2)$$

Only the last two terms are of interest. These can be written as $k|u_o|^2 u$ and $ku_o^* u^*$, respectively, or alternatively, as $ka_o^2 a e^{i\phi}$ and $ka_o^2 a e^{i(4\pi f_o x - \phi)}$. The term $k|u_o|^2 u$ represents a wave that is, to within a constant, identical to the object wave u ; the other term represents a wave that is, to within a constant and a linear phase term, a conjugate of the original wave. The factor $e^{i4\pi f_o x}$ has the physical meaning that the wave is propagating in a direction 2θ given by $\sin 2\theta/\lambda = 2f_o$. The readout waves are shown in Fig. 4(b).

The so-called true wave u and the conjugate wave u^* thus propagate angularly away from each other and from the direct transmitted beam, which carries the uninteresting components $|u_o|^2$ and $|u|^2$.

However, a more suitable way to form the conjugate wave is to illuminate the hologram with a readout wave that is, again, a duplicate of the original reference wave, but propagating in the opposite direction, as in Fig. 4(c). The conjugate wave is now propagating along the z axis, along the path of the original object wave. It is readily shown that this counterpropagating wave is identical in shape to the original object wave recorded by the hologram, except that it travels in the opposite direction, exactly the situation portrayed in Fig. 2. This wave propagates so as to form an image that is, to within a constant, identical to the original object, except for the counterpropagation, which produces a conjugate image $s^*(x, y)$. The image intensity, however, is $|s|^2$, which is identical to the original image intensity.

This conjugate wave was, in the early days of holography, a detriment, since without the use of the off-axis method (reference wave introduced at an angle to the object wave) the two images were inseparable, and the true image could be observed only against the background of the defocused conjugate image. Thus, an out-of-focus image overlaid the true image, making it noisy.

In 1962, the same year that the off-axis reference wave for separating the two images was introduced, a possible use for the conjugate image was suggested. If the original propagation path of the object wave were aberrated, as would occur if the wave propagated through an irregular medium, then the conjugate wave propagating through the medium would undergo a correction and would arrive at the original object position perfectly corrected, resulting in a sharp image, a process sometimes called wavefront healing. The idea lay dormant for 3 years, until in 1965, experimental demonstration of this idea was reported. In one case, the object was a transparency containing opaque lettering against a transparent background and the inhomogeneity was a piece of frosted glass, which resulted in extremely severe aberrations, so that the object wave, having traversed the diffuser, was no longer capable of forming an image. However, when the distorted wave was recorded holographically, conjugated, and sent back through the same diffuser, the result was a highly legible image.

The principle of wavefront healing is illustrated in Fig. 5, when the inhomogeneity is an extended, or three-dimensional, structure. A point object O sends a spherical wave into the inhomogeneity. As the wavefront propagates, it becomes distorted, and loses its sphericity. When it emerges from the inhomogeneity, the wavefront is highly distorted. The wave is conjugated at the conjugator C and emerges as a back-propagating wave that retraces the original wave path through the inhomogeneity, gradually losing its aberrations and looking more like a spherical wave. When it emerges from the inhomogeneity, it is a perfect spherical wave that converges to a point image that coincides with the original object. At each plane throughout the inhomogeneity, the back-propagating conjugate wave (dashed lines) has the same shape as did the original wave when it crossed that plane (solid line).

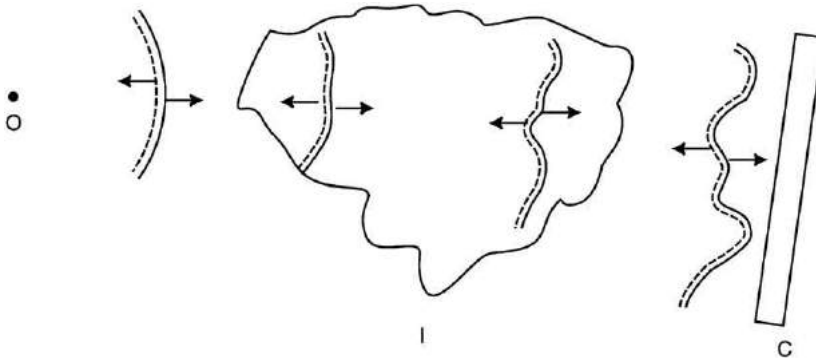


Fig. 5 The wavefront healing process. C is the conjugator, I is the inhomogeneity, and O is a point object.

This interesting wavefront healing property of holographic phase conjugation was the beginning of a new branch of optics, called phase conjugation, giving rise to a number of applications. However, the holographic method had the problem that it did not occur instantaneously, but had to await the processing of the photographic film on which the hologram was recorded. This delay limited the range of potential applications. Later, in the 1970s, nonlinear crystals became available that did the equivalent of holography, but in real time, or near real time – a process known as four wave mixing. There are yet other types of real-time phase conjugators with nonlinear devices.

Pseudoscopic Image

As long as the object is only two-dimensional, a function of x and y only, with no extent along the axial dimension, z , the intensity distribution of the true and conjugate images is identical. The conjugateness, being a pure phase phenomenon, disappears when the observable is the image intensity $|s|^2$, and not the complex amplitude s . Since the intensity is always the observable, whether the image is viewed by eye, or with a camera, the conjugateness generally is of no significance.

However, when the object for holography is three-dimensional, the situation is different, and the conjugate image exhibits some strange properties. These arise from conflicting cues pertaining to the three-dimensionality of the image. To explain, we review some basics about how a three-dimensional world is perceived by a viewer. First, there is the stereo effect, resulting from the viewer having two eyes, with each eye seeing an object from slightly different positions, a phenomenon called binocular disparity. The mind uses this stereo effect to perceive three-dimensionality. But there are other cues that also contribute to perception of three-dimensionality. Another is parallax; as the viewer moves his head, objects farther from the observer change their angular position more slowly than objects close by. The observer can sense the three-dimensionality of an object, even with only one eye, by moving the head, thereby inducing parallax. Also, there is interposition, whereby near object portions block the viewing of object portions lying behind them, again giving a sense of depth. There are still other cues, more subjective, such as perspective, parallel lines converging toward infinity, shadowing, etc. With these observations in mind, we consider the hologram under illumination and the two images that it forms (Fig. 6). As depicted in Fig. 6, these two images lie in mirror symmetrical positions relative to the hologram. Both images have the same side facing the hologram. It must be this way, because when the hologram was recorded, it was side a that faced the recording plane. And both images must therefore have side a closer to the hologram.

When a viewer observes the true image (a virtual image), he views it looking through the hologram, and he sees the image at the position where the original object was when the hologram was recorded, and it is, in its visual properties, indistinguishable from the actual object; it has all of the three-dimensional and the parallax properties that the original object would exhibit. However, when the viewer observes the conjugate image, he must view it from the side of the image opposite from the hologram, i.e., he views the image from the back side. When a viewer observes the conjugate image, he sees it from the back side (side b). All of the above cited cues are in agreement about the orientation of the image, viz., that side b is closer to the observer than is side a . That is, all cues but one. The conjugate image is formed as the hologram recording plate saw it, with side a being closer. Thus, all interpositions are formed with the same relation that the object had with respect to the recording plate. The side a is the part of the object that was closer to the hologram, so that when interpositions occur, side a blocks side b . Since the observer thus sees side a partially obscuring side b , he must conclude that side a is closer than is side b . However, this observation is in direct conflict with the other cues, especially the stereo and the parallax cues, which are themselves strong cues, as is the interposition cue. The conflict of these cues, with two of them interpreting the image as being formed with one depth orientation, the other with the opposite, gives a number of strange effects. For example, as the viewer moves his head so as to induce parallax, the object appears to rotate. If the image is that of a human head, for example, the observed effect is that as the viewer changes his position, the head appears to

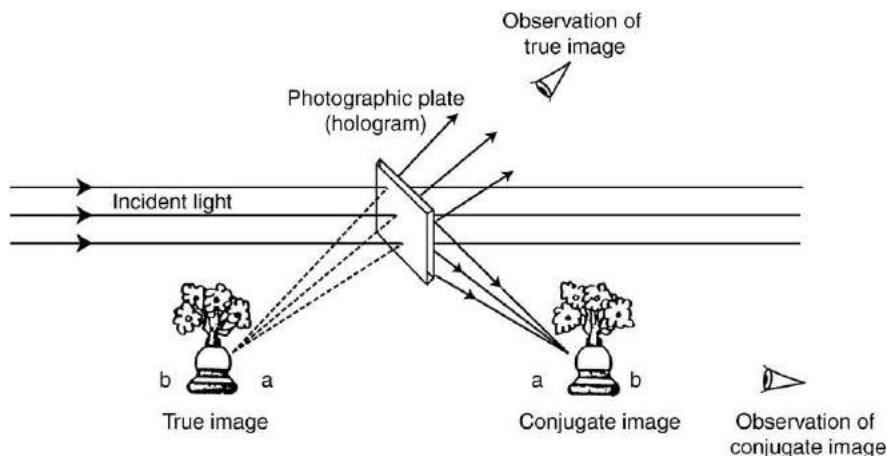


Fig. 6 Formation of the true and conjugate images from a hologram.

rotate in synchronism with the observer, as if the image had its gaze fixed on the viewer, an effect that is often perceived as being eerie.

Similarly, if the object were two flat plates, (Fig. 7) in the situation depicted, the planes have no overlap, i.e., no interposition, as seen by the observer. And the observer unequivocally perceives plate A as being closer than plate B. However, if the viewer moves his head to a position where interposition occurs, the conflict of cues arises, A becomes partially blocked by B, which is quite correct as observed at the position of the hologram, but is quite incorrect from the viewer's vantage point. The result of this conflict is that the viewer, who previously perceived the three-dimensionality, now sees the three-dimensionality perception suddenly collapse, and the sense of three-dimensionality is lost. Other strange perception anomalies can also be produced. Such images, with conflicting depth cues, are called pseudoscopic, as opposed to orthoscopic images, where the cues do not conflict.

One Way Phase Conjugation

A significant limitation of phase conjugation aberration correction is that the image and object are on the same side of the inhomogeneity. One would usually want them to be on opposite sides. In a modified form of the phase conjugation process, shown in Fig. 8, this result is achieved. The first step of the process is carried out by passing a plane wave, instead of the object wave, through the inhomogeneity. This plane wave originates from a point source located at the focal plane of a collimating lens.

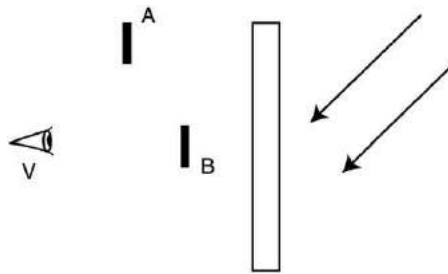


Fig. 7 Pseudoscopic conjugate image formed from an object consisting of two plates. The viewer V perceives plate A as being closer than B and there are no conflicting cues, since there is no interposition. However, if the observer moves his head upward, there is partial interposition, thus conflicting cues, and the sense of three-dimensionality is lost.

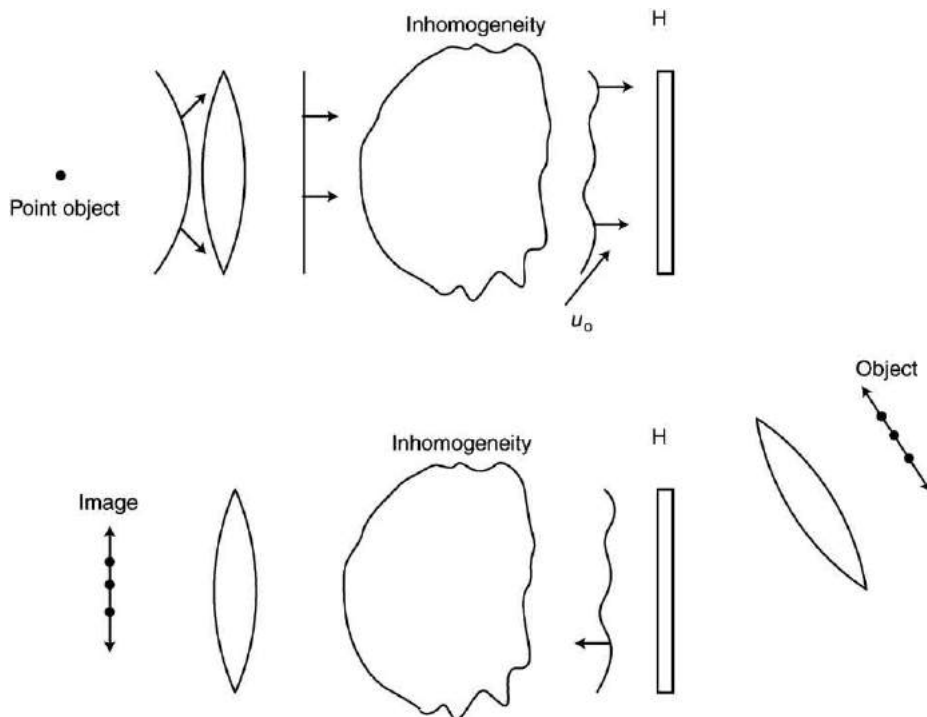


Fig. 8 One-way phase conjugation. Top: Making the hologram. Bottom: The viewing process. We show three points on the object. The center point forms a well-corrected image, but the other two points produce imperfectly corrected images.

The plane wave thus carries information about the inhomogeneity to the recording plate, where it is combined with a reference wave and a hologram of the distorted plane wave is recorded. Thus, information about the phase defect is stored on the hologram.

Next, the hologram is read out, not with a counterpropagating replica of the reference wave as before, but instead with an object-bearing wave that passes through the hologram. This wave emerges from the hologram and now has on it the recorded aberration information. As it counterpropagates along the path of the original object wave, it passes through the inhomogeneity, whereupon the aberration is cancelled and a corrected image forms at the focal plane of the collimating lens, where previously the point source had originated. The goal is achieved; the object and the corrected image are now on opposite sides of the inhomogeneity.

This correction is only approximate. To explain, we think of the object as being a collection of many points. The radiation from each point is converted into a plane wave by the lens, with different object points producing plane waves propagating at different angles. These plane waves then impinge on the hologram and emerge with each one bearing the aberration stored on the hologram. They continue on their paths and enter the inhomogeneity. The plane wave component, whose path direction matches that of the reference wave used in recording the hologram, will emerge from the inhomogeneity as a perfectly corrected wave, thus forming a well-corrected point image at the position of the original point radiator. However, another plane wave component, traveling in an incorrect direction, will be laterally displaced from the proper position for entering the inhomogeneity, and the aberration correction will be imperfect, the degree of imperfection being in direct relation to this displacement error. The displacement error is the product of the angular error and the propagation distance between hologram and inhomogeneity. If the inhomogeneity is slowly varying as a function of x and y , and if the object does not have a large extent, so that the range of angles is small, the image degradation will be small.

If the last two conditions do not apply, two options are available. First, the distance between inhomogeneity and hologram can be reduced, either by bringing the hologram closer to the inhomogeneity, or by imaging the hologram onto the inhomogeneity. The correction is now valid over a wider extent of the image, especially if the inhomogeneity can be treated as planar rather than extending over a volume. Alternatively, a scanning process can be carried out on the object beam falling on the hologram, so that in sequence, each plane wave component assumes the proper angle to counterpropagate along the path for perfect correction, and will form a well-corrected image point. This output spot then falls on a pinhole positioned to separate it from other, imperfectly corrected points. Then, over time, all points of the object will form well-corrected image points, and the entire object will thus be well-imaged. The one way phase conjugation method was first used to correct the spherical aberration of a lens, and the result was quite successful.

Phase conjugation has become an important branch of modern optics. The process is often carried out using photographic materials. More often nowadays, it is carried out using nonlinear crystals. Some of the crystal methods are essentially holography in real time, but others, such as phase conjugation by Brillouin scattering, go beyond holography.

See also: Adaptive Optics. Nonlinear Optical Phase Conjugation

Further Reading

Caulfield, H.J. (Ed.), 1979. *Handbook of Optical Holography*. New York: Academic Press.

Goodman, J.W., 1996. *Introduction to Fourier Optics*, 2nd edn. New York: McGraw-Hill, chap. 9. This chapter is devoted to holography, with section 9.12.3 dealing with phase conjugation aberration correction.

Kogelnik, H., 1965. Holographic image projection through inhomogeneous media. *Bell Systems Technology Journal* 48, 2451–2452.

Leith, E.N., Upatnieks, J., 1965. Holograms: their properties and uses. *SPIE Journal* 4, 3–6. This is the earliest publication demonstrating the principle of wavefront healing by holographic phase conjugation.

Smith, H.M., 1975. *Principles of Holography*, 2nd edn. New York: Wiley.

Hyperspectral Imaging

ML Huebschman, RA Schultz, and HR Garner, University of Texas, Dallas, TX, USA

© 2005 Elsevier Ltd. All rights reserved.

Introduction

Hyperspectral imaging techniques on a microscopic level have paralleled the hyperspectral remote astronomical sensing which has developed over the years. It is especially apparent in remote sensing of our own planet's land masses and atmosphere, such as in the LANDSAT satellite program. In biology, biochemistry, and medicine, hyperspectral imaging and multispectral imaging are often interchanged, although there is a distinct difference in the two techniques.

A multispectral imager produces images of each of a number of discrete but relatively wide spectral bands. Multispectral imagers employ several wide bandpass filters to collect the data for the images. A hyperspectral imager produces images over a contiguous range of spectral bands and, thus, offers the potential for spectroscopic analysis of data. Hyperspectral imagers may use gratings or interferometers to produce the contiguous range of spectral bands to record.

Fluorescent dyes (also referred to as fluorophores or fluorochromes) and recently quantum dots, as markers, are a very powerful tool for analyzing organic tissues for researchers in cytology (study of cell function), cytogenetics (study of the human chromosomes), histology (study of structure and function of tissues), microarray scanning (analysis of concentrations of multiple varied probes), and biological material science. Without the correlation of the spectral and spatial data in biological images, there would be little useful information for the researcher. Therefore, hyperspectral imaging in the biological sciences must combine imaging, spectroscopy, and fluorescent probe techniques. Such imaging requires high spatial resolution to present a true physical picture, and the spectroscopy must also have high spectral resolution in order to determine the quantities of biochemical probes present. The correlation must provide the color-position relation for identification of meaningful information.

This description will begin with an initial review of multispectral imaging to lead into a discussion of hyperspectral imaging. The discussion will examine aspects of hyperspectral imaging based on an interferometric method and on a grating dispersion method similar to the methods in multispectral imaging.

Image Parameters

The analysis of fluorophores involves the excitation of the dye with a shorter wavelength and thus higher energy than the emission spectrum. The use of the appropriate excitation wavelength will yield a maximum emission intensity at a characteristic wavelength. The quality of a fluorescent image generally combines contrast and brightness. Contrast depends on the ratio of emission intensity to that of background light. Consequently, the use of filters to block the excitation wavelength and prevent it from overwhelming the field is critical in providing high contrast for the detection of fluorophore emission. However, the use of narrow filters to restrict the background can severely limit the brightness of the fluorescent image. This is due to the barrier filters allowing only a portion of the actual emission spectrum to pass. Brightness is an intrinsic property of each individual fluorophore, but is also dependent on the ability to illuminate with the peak excitation wavelength for that fluorophore.

The balance between contrast and brightness may be further complicated by applications designed to facilitate the evaluation of multiple fluorophores. The selection of appropriate excitation and emission filters within a single set to allow for the simultaneous viewing of multiple fluorophores can be very inhibiting. Although quad-band beamsplitters are currently available, commercial advances to date have yielded only triple-pass combinations for simultaneous detection, such as one of which permits the simultaneous detection of the widely used fluorophore combination fluorescein-5-isothiocyanate (FITC), Texas Red and 4',6'-diamidino-2-phenylindole (DAPI) (e.g., Chroma Technology Corp.).

Although the capture of digital images has propelled this field, efforts to collect multicolored images as digital files also suffer from technical limitations. Color CCD cameras often utilize filters or liquid crystals to segregate the red, green, and blue wavelengths onto the imaging chip. The employment of these cameras is restricted to resolving colors as discriminated by the filters used. For scientific applications, then, the greatest sensitivity and selectivity has been achieved through black and white CCD cameras with optimized filters for a specific application. While this approach is widely used, it suffers from the inherent limits of the spectra of data that can be analyzed since it is defined by the characteristics of the optical filters employed. Thus, the system is restricted to the use of fluorophores that can be effectively resolved from each other by filters, and the filters selected limit the data collected to a portion of the total emission spectrum for any one fluorophore.

Multispectral Imaging

Multispectral imaging systems in biomedical research are composed of a microscope, filters and/or interferometer for spectral selection, a charged coupled device (CCD) camera to record the data, and computer control for acquisition, analysis, and display

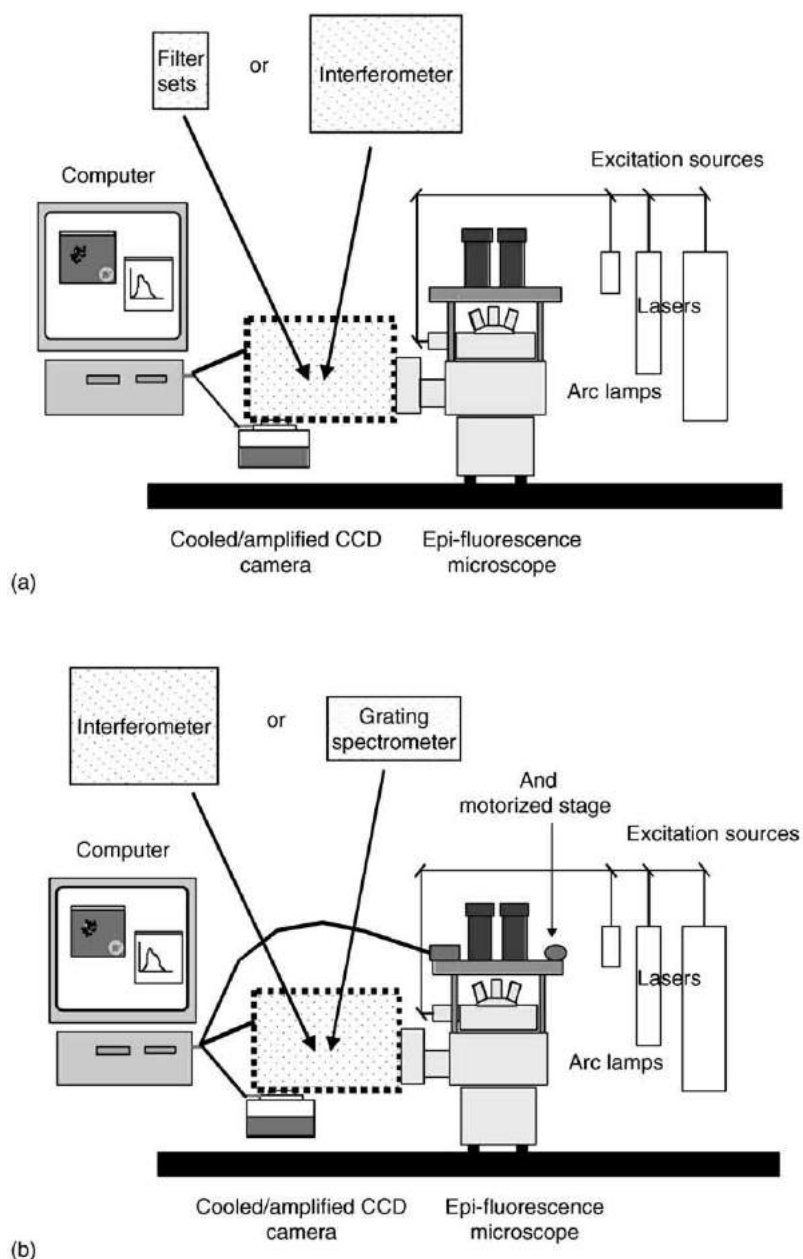


Fig. 1 The instruments for multispectral and hyperspectral imaging are similar. (a) Multispectral imagers use interferometers or bandpass filters to collect noncontiguous emission wavelengths. (b) Hyperspectral imagers employ interferometers or gratings to collect a contiguous wavelength band. Computers, microscopes, CCD cameras, and excitation light sources are common to all the techniques. The systems which use filter sets or interferometers use motorized filter placement systems or automated interferometers. The gratings systems employ a motorized stage for the sample movement. Adapted with permission from Schultz RA, Nielsen T, Zavaleta JR, Ruch R, Wyatt R and Garner HR (2001) Hyperspectral imaging: a novel approach for microscopic analysis. *Cytometry* 43: 239–247.

(Fig. 1(a)). The introduction of CCD cameras has offered a means to document results and capture images as digital computer files. This has facilitated the analysis of results with more and more complex probe sets and has allowed the use of more sophisticated computer algorithms for data analysis. The other major component of the multispectral imaging system is a light source to excite the sample's fluorescent dyes whose emissions are selected by filters or an interferometer for recording on the CCD camera.

An example of the use of multispectral imaging is in the identification of the 24 human chromosomes. A scheme was presented for the analysis of human chromosomes by differentially labeling each chromosome with a unique combination of 5 fluorophore-labeled nucleotides (building blocks of DNA and RNA, consisting of a nitrogenous base, a five-carbon sugar, and a phosphate group). These complex probes are simultaneously hybridized to metaphase chromosomes and analyzed to determine which of the

fluorophores are present. Combinations of one, two, or three of the five fluorophores are enough to distinguish a particular chromosome from the others.

An interferometric multispectral system used to perform this type of analysis is the Spectral Karyotyping (chromosome typing) or SKY System (Fig. 1(a)). This technology does not represent hyperspectral imaging, but is presented for comparison. Illumination light passes through a triple band-pass dichroic filter which allows for the simultaneous excitation of all the fluorochromes in, for instance, Speicher's scheme. Modern multi-band pass filters add little to image distortion while effectively blocking unwanted autofluorescence or excitation illumination. The fluorescence emissions from the sample are split into two separate beams by the interferometer. The beams are then recombined with an optical path difference (OPD) between them such that the two beams are no longer in register and interfere with each other. By adjusting the OPD, a particular wavelength band can be selected for, which results in constructive interference and thus can be collected by the CCD camera pixels. After adjusting the OPD to record the 5 dye wavelengths, each pixel's 5 records are then compared to the known probe-labeling scheme to identify the corresponding chromosome material. The image of the metaphase chromosome spread can be karyotyped using specific pseudo-colors for maximal visual discrimination of the chromosomes.

Another nonhyperspectral method is the Multi-Fluorescence *In Situ* Hybridization or M-FISH System. It uses five single-band-pass excitation/emission filter sets where each filter set is specific for one of the five fluorochromes used in the combinatorial labeling process (Fig. 1(a)). Again, five separate exposures are taken (one exposure for each filter set) and combined to create a multicolor image. Most M-FISH systems now have motorized filter set selectors which minimize any image registration shifts that may occur between exposures. For each pixel, the presence/absence of signal from each of the five fluorochromes is assessed and then matched to the probe-coding scheme to determine which chromosome material is present. As with the SKY System, an image of the metaphase spread is converted into a karyotype and specific pseudo-colors are used to maximize visual discrimination of the chromosomes.

Hyperspectral Imaging

As has been stated, the main difference between multi- and hyper-spectral imaging is the collection of a broad contiguous band of wavelengths. The apparatus and methods for hyperspectral imaging are similar to multispectral imaging (Fig. 1). Instead of collecting sections of the emission spectrum using filters or interferometers, hyperspectral imaging collects a continuous emission spectrum by dispersing the wavelength of light using gratings and measuring the wavelength intensity, or by dispersing Fourier components of light using interferometers and measuring interference intensity (Fig. 1(b)). Multispectral images are limited to measuring total intensity within the selected wavelength band. Determining the abundance of more than one dye per pixel is less accurate when only a small segment of each dyes' spectrum is passed by the emission filter. This has not been considered a limitation in the past because most research depended on just identifying the presence and not the abundance of several spectrally broad and overlapping dyes which also occupied relatively large spatial areas. With the advent of smaller and smaller featured microarrays, the desire to use more dyes in identification schemes, and the use of quantum-dot markers which can be used to indicate both genetic makeup and quantity, the limitations of multispectral imaging has led to a move to hyperspectral imaging.

In contrast to the fluorescence multispectral imaging approaches, a hyperspectral image is fully three-dimensional. Every XY spatial pixel of the image has an arbitrarily large number of wavelengths allied with it. An intensity cube of data is collected. Two of the axes of the cube are spatial and one axis is wavelength (see Fig. 2). At each of the 3D points, there is an associated intensity of emitted energy. Therefore, it is a four-dimensional information space. The individual intensity signature of differing emitting bodies can then be separated to give a complete view. Spectral decomposition is a mathematical/algorithmical process by which individual emission spectra of the dyes can be uniquely separated from a composite (full spectral) image by knowing the signature of each dye emission spectra component measured separately. Spectral decomposition may be performed using standard algorithms with codes which are readily available.

Interferometric Hyperspectral Imaging

The system architecture of the interferometric hyperspectral imager resembles that of the SKY System (Fig. 1). But, instead of just sequencing the interferometer to select a set of specific wavelengths to record as in the SKY System, the interferometer in the hyperspectral imager is sequenced through discrete optical path differences to record the intensity at each pixel for each OPD. Fourier spectroscopy techniques are used to transform the OPD intensities at each pixel into the wavelength intensities at each pixel. Each pixel may be thought of as a Fourier spectrometer which is not dependent on any other pixel. Then, spectral decomposition algorithms can be run to determine which dyes are present (see the next section).

Like the multifilter methods, the interferometer approach provides a real two-dimensional image that can be observed, focused, and optimized prior to measurement. But like the filtering methods, the excitation source must be filtered for sufficient contrast to view the fluorescent emissions. Interferometers have high optical throughput, relative to other methods, and can have high and variable spectral resolution. The spectral range can cover from the ultraviolet to the infrared. The high and variable spectral resolution allows the user to trade off sensitivity, spectral resolution, and acquisition time in a way that is not possible with any other technique. And interferometers do not cause polarization of light as a filter may.

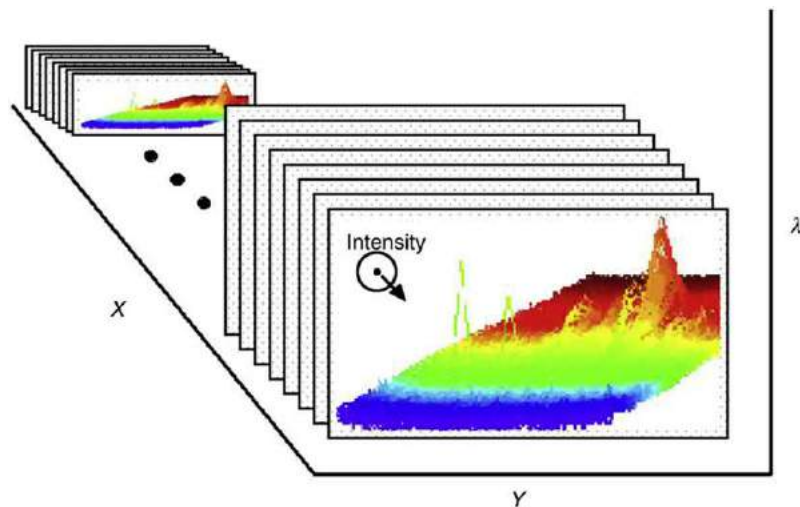


Fig. 2 Intensity data cube. In this data cube, each CCD exposure captures a measurement in intensity (number of counts) by wavelength for a small spatial area in Y and ΔX . Many displacements in X with subsequent CCD acquisitions yield a 'cube' of data in X , Y , and λ . The intensity counts may be thought of as another dimension, $I(x,y,\lambda)$. Another form of the data cube is with each CCD acquisition comprising intensity counts for a spatial X and Y area for a small $\Delta\lambda$ and multiple displacements in λ for each acquisition.

However, interferometric instruments are susceptible to image registration and lateral coherence difficulties. They are limited by the extreme precision required of the optical components which results in the continuous need for monitoring and adjustment, since aberrations are not tolerated in interferometric instruments. The throughput is reduced by loss due to the beamsplitter design and from the mirrors. Additionally, the Fourier transform instrument deconvolutes based on a two-step process; the correlation of the light passed by an interferometer followed by an inverse Fourier transform in software.

To mitigate the extreme precision and sensitivity of the most interferometric systems, Sagnac interferometers have been employed because of their common path intrinsic stability. The common path compensates any misalignment, vibration, or temperature affect. The OPD is provided by the Sagnac affect. When the Sagnac interferometer is rotated, the path traveled in one direction will be different from that of the light on the common path in the other direction. Although, this interferometer method is trading one difficulty for another, it is somewhat easier to manage the rotation of a physical system than to maintain the exacting precision of mirror alignments in a standard interferometer.

Spatial Resolution

The spatial resolution of a Sagnac Hyperspectral Imager is set by the size of the pixels of the CCD camera and the microscope system magnification, since the interferometric hyperspectral viewer sees the entire field of view at once. Let the x - and y -dimensions of the CCD pixel be δx and δy , respectively, and the magnification be M , then the spatial resolutions in the x and y directions are $2\delta x/M$ and $2\delta y/M$, respectively. For CCD pixel sizes of 10 microns square, binning together two pixels in the x - and two pixels in the y -direction, and an overall magnification of 60, the x and y resolutions are both 0.66 microns.

Field of View

For N pixel rows and C pixel columns, the field of view will be $N\delta y/M$ and $C\delta x/M$ for the length of the area along the y - and x -axis, respectively. For a CCD array of 1536 columns by 1024 rows, and the parameters above, the field of view is 171 microns by 256 microns, x and y , respectively.

The sensitivity, dynamic range, and spectral response are set by the CCD camera optical components.

Spectral Resolution

Spectral resolution will depend on the ability to vary the OPD and the Fourier transform. While the interferometer is used basically as a filter in the multispectral imager to select a wavelength, in the hyperspectral imager it is used to vary the OPD for all the wavelengths whose energy is integrated at the CCD camera pixels.

The electric field, E , for emissions at one wavelength, λ , or wave number, $k=1/\lambda$, radial frequency, ω , and amplitude, A , can be written as

$$E(k) = A(k)e^{-i(2\pi kx - \omega t)} \quad (1)$$

and the intensity which is measured is the magnitude squared of the electric field:

$$I = E^* E \quad (2)$$

Emissions from fluorescent sources are usually noncoherent and composed of a range of wavelengths, even with only one dye present. The intensity measured at a pixel is the superposition of many wave numbers. After traveling a certain OPD, L , the intensity integrated over all wave numbers may be written as

$$I_{\text{pxl}}(L) = \frac{1}{2} \left[\int I(k) dk + \int I(k) \cos(2\pi k L) dk \right] \quad (3)$$

or

$$I_{\text{pxl}}(L) = C_0 + C_1 \cdot \text{Re}\{\text{FT}[I(k)]\} \quad (4)$$

where FT stands for Fourier transform.

The inverse transform is

$$\begin{aligned} \int I_{\text{pxl}}(L) e^{-i2\pi k' L} dL &= C_0 \int e^{-i2\pi k' L} dL + C_1 \int \int I(k) e^{-i2\pi k L'} dk \cdot e^{i2\pi k' L} dL \\ \int I_{\text{pxl}}(L) e^{-i2\pi k' L} dL &= C_3 + C_4 I(k) \end{aligned} \quad (5)$$

The integration constant, C_3 , can be ignored for large wave numbers, as in the case of the visible range, since C_3 will be proportional to the inverse of the wave number. The intensity as a function of wave number is then

$$I(k) = C_5 \int I_{\text{pxl}}(L) e^{-i2\pi k L} dL \quad (6)$$

For absolute quantitative measurements, the constant C_5 can be determined by calibration with a known emission. For discrete measurements made by changing the optical path length, then the working formula is

$$I_{\text{pxl}}(k) = \sum_{L_i} I_{\text{pxl}}(L_i) e^{-i2\pi k L_i} \Delta L \quad (7)$$

This is the discrete Fourier transform which must be performed for each pixel to determine the intensity as a function of wavelength. Per the Nyquist sampling tenets:

$$\Delta L = \frac{1}{(N/2)\Delta k} \quad \text{or} \quad \Delta k = \frac{2}{N\Delta L} \quad (8)$$

where Δk is the discrete wave number interval and N the number of different OPDs. Sagnac interferometers can make rotations which give at least a change of $\Delta L = \lambda/4$, so substituting in Eq. (8):

$$\begin{aligned} \Delta k = k_1 - k_0 &= \frac{1}{\lambda_1} - \frac{1}{\lambda_0} = \frac{\Delta \lambda}{\lambda^2 + \lambda \Delta \lambda} = \frac{8}{N\lambda} \\ \frac{\Delta \lambda}{\lambda^2} &\approx \frac{8}{N\lambda} \quad \text{or} \quad \frac{\Delta \lambda}{\lambda} \approx \frac{8}{N} \end{aligned} \quad (9)$$

A 5% resolution in wavelength (or 20 nm at 400 nm) requires $N=160$ and, of course, a 2.5% resolution requires $N=360$. Therefore, about 400 OPD changes and acquisitions of the intensity fringes will give calculated resolutions of less than 10 nm in wavelength.

Analysis of the spectrum based on this resolution can be conducted to determine which dyes are present. Software techniques for displaying data will be described in the next section.

Grating-Dispersion Hyperspectral Imaging

A hyperspectral imaging microscope collects the complete visible emission spectrum from a microscope slide by using a grating spectrometer to provide the wavelength dispersion. This microscope correlates spatial and spectral information with minimal use of optical filters. A hyperspectral microscope system is composed of a standard epi-fluorescence microscope (Olympus IX70) equipped with objectives lens (including $1.25\times$ and $10\times$ Plan Fluorite, $40\times$ Plan Apochromat Dry, and $60\times$, and $100\times$ Plan Apochromat Oil), as well as a set of standard filter cubes (Chroma Technology) to allow for traditional visualization of dyes through the eyepiece. (Prototype typical components are listed in parentheses.) Multiple excitation sources can be utilized including standard 100 W mercury and xenon arc lamps, tungsten brightfield illumination, a 532 nm solid state laser (Brimrose), a helium neon laser (Uniphase), an argon ion laser (Uniphase), and a pulsed-doubled nitrogen dye laser (Laser Science, Inc.) (Fig. 1(b)).

One side port of microscope is optically coupled to a spectrograph (Acton Research SpectraPro[®] 556i). A narrow entrance slit into the spectrograph allows for only one line (Δx) of the image to be captured at a time. A 50-micron slit permits 80% of maximum throughput for the spectrometer. The spectrograph separates the polychromatic light illuminated from this single line into its light components. Normal operation uses a grating with 50 grooves/mm and a 600 nm-blaze wavelength, which allows

wavelengths of approximately 400–780 nm to be measured. Different spectrograph gratings can be used for special applications from near UV to near IR. The center wavelength and grating can be shifted under computer control. The output of the spectrograph is then captured with an air/liquid-cooled CCD camera (Photometrics-Quantix) and an optional image intensifier (Video Scope International VS4-1845). The camera has a resolution of 1536×1024 pixels with a depth of 12 bits per pixel. The 1536 pixels are aligned in the spatial direction to allow the greatest spatial resolution. During normal operation the wavelength resolution is 'binned' to 128 pixels, sufficient to resolve ~ 8 different spectra in a single excitation scan. Higher resolution can be employed for special applications.

A line of excitation light can be provided by a line generator when laser excitation is used or by a combination of slits and cylindrical lenses when using traditional lamps or burners (mercury or xenon). This line of light illuminates the sample in the same position that maps the resulting emission light through the imaging spectrograph and onto the camera. Some advantages of illumination with a line of light include decreased sample bleaching, reduced background due to scattering, and efficient use of available excitation power. A traditional bandpass, a low-pass dichroic filter, or a linear variable filter (Raynard Corp.) may be employed to select or restrict the excitation wavelength during specific applications.

Linear Variable Filter (LVF)

The linear variable filter (LVF) (Raynard Corporation) can be used to replace the standard excitation filter cube to provide a capability for continuous excitation wavelength band selection. The LVF is $1'' \times 2.25''$ and is coated such that it continuously filters light from ~ 400 nm to ~ 800 nm along its length. Custom filters that span other wavelengths are available. The coating and therefore wavelength is uniform along a short distance, making it an ideal method for selecting an excitation wavelength band to paint across the slide, excite the slide's contents, and map the contents' emission into the slit of the imaging spectrograph. The bandwidth passed is about 21 nm wide. The LVF introduces $\sim 50\%$ attenuation of the intensity.

Spatial Resolution

The spatial resolution of a grating spectrograph Hyperspectral Imager is set by the size of the pixels of the CCD camera in the y -direction and the microscope system magnification. However, differing from the interferometric system, the spatial resolution in the x -direction depends on the spectrometer slit width and the microscope system magnification. As before, let the x - and y -dimensions of the CCD pixel be δx and δy , respectively, and the magnification be M . Let w_s be the slit width of the spectrometer, and note that the slit width is always going to be larger than a few δx . Then the spatial resolution in the y -direction is again $2\delta y/M$, but the spatial resolution in the x -direction is now $2w_s/M$. Again, for CCD pixel sizes of 10 microns square, binning together two pixels in the y -direction, and an overall magnification of 60, the y -resolution is 0.66 microns, as it was for the interferometer method. For a slit width of 50 microns, the x -resolution is 1.67 microns, independent of the number of pixels binned in the x -direction. The control software moves the stage the right number of steps to minimize the overlap or skipped areas. The slit width can be reduced, but at a loss of throughput or increased exposure time. The control software can compensate for different spatial resolution to display the image as 1:1.

Field of View

For N pixel rows (y -dimension) and C pixel columns (x -dimension), the field of view will be $N\delta y/M$ for the length of the area along the y -axis. If the spectrometer slit length is less than the height of the CCD, then the length is just h_s/M , where h_s is the slit length. We orient our CCD array so that the 1536-pixel dimension is along the y -axis. The y -view is thus 256 microns as in the interferometric example. To achieve the same field of view as the Sagnac method imager, we would need to take 103 stage movements and acquisitions. This highlights the difference in the two hyperspectral methods. The grating method collects the complete spectrum on each acquisition with no spectrometer adjustment but needs to step down the slide to produce the full spatial image. The interferometer method collects the complete spatial image on each acquisition with no position adjustment but needs to step the interferometer through a range of optical path differences to acquire multispectral content.

The sensitivity, dynamic range, and spectral response are set by the CCD camera optical components.

Spectral Resolution

Spectral resolution depends on spectrometer instrument function which includes the slit width, the dispersion of the grating, the CCD pixel size, and number of pixels binned together. The dispersion for normal operation was determined to be ~ 40 nm/mm or 0.4 nm/pixel in the 10 micron CCD pixel example. For example, binning of 8 pixels yields a spectral resolution of 3.2 nm at all wavelengths. To achieve this spectral resolution, the interferometer method would require 1000 OPD increments and acquisitions.

Analysis of the spectrum based on this resolution can be conducted to determine which dyes are present. Although the spatial resolutions differ considerably in the two methods, as long as the dye peaks are separated by the spectral resolution spacing, the spectral decomposition algorithms will find all dyes within the spatial resolution area. We have demonstrated that the analysis algorithms can resolve 5 dyes with peaks separated by more than 6 nm in various combinations and known quantity ratios against the dye standards. The program successfully identifies the dyes present with a 91% accuracy in the quantities.

Software

Software (*HyperScope*, a 'C' language program) controls all the hardware components for acquisition. It includes features for analyzing and displaying the resultant $X - Y$ images and spectral information for the grating spectrometer Hyperspectral Imager. During a scan, individual pictures are taken in the $Y - \lambda$ plane, while the stage is incremented in the X -direction. A collection of such $Y - \lambda$ scan files, which can total from 50–1000, is merged to build an image cube (Fig. 2). Another advantage of the single-line acquisition mode is that time-dependent features can be monitored continuously. The software was designed to permit repeated collection of a single line at variable time points. The utility of this feature has been demonstrated by capturing images of fluorescent objects moving through UV-transparent capillaries.

The software was designed to display the resulting $X - Y$ plane extracted in various ways from the image cube. A graph of the emission spectrum for any pixel within that image may be viewed by clicking on any pixel of interest in the $X - Y$ image (Fig. 3(c)). A standard linear curve-fitting algorithm is used to determine the contribution of individual dyes to the measured emission spectrum. The software also features a windowing technique (integration of the emission spectra over a fixed window) which emulates standard filtering. An overlay feature within the software allows the curve fitting results to be displayed in false color and compared by viewing the composite for multiple fluorophores in a single image. Additionally, it permits the contribution of a fluorophore to be emphasized for visualization purposes by either displaying the contribution of a single fluorescent signature in the absence of the others that were present in the image or by scaling individual display intensity values up or down. Overlapping pseudo-colors assigned to represent each fluorophore (either by curve fitting or by windowing) may be displayed as added, averaged, maximum or minimum values. For each new fluorophore–excitation source combination, the characteristic emission spectra is acquired and added to a library of available spectra that can be used as components in the spectral decomposition. This permits the user to introduce new fluorophores at any time. Moreover, background spectra (defined as a region of interest by the user) can be acquired and used as one element of the decomposition. This background component can be turned off in the false color view (as any of the spectral components may be) resulting in enhanced noise reduction and visual contrast improvement.

Description of Acquisition

In order to acquire a full image, a slide is placed on the microstage mover (Ludl Electronics) that allows for precise movement of the sample across the image acquisition plane. The user can position the stage via a joystick or via software to move to pre-determined coordinates. The image is then scanned, with the number of scans being selected at the time of the image acquisition.

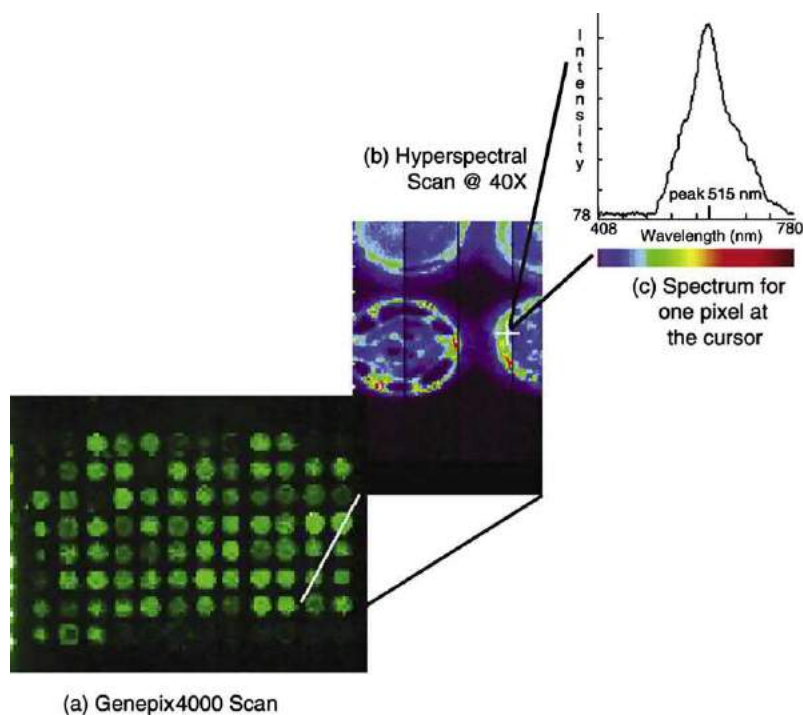


Fig. 3 The benefit of spectral imaging is demonstrated in a spatial high-resolution scan of an expression microarray. (a) A portion of the scan of an expression microarray on a microscope slide using the Genepix4000 scanner. It records the intensity emitted by the Cy3 dye when excited with a 532 nm laser. (b) A portion of the scan of the same microarray with the Hyperspectral Imaging Microscope using a 40X objective and Hg excitation lamp. False color is displayed to indicate the intensity variations. (c) A *HyperScope* computer program display of the spectrum recorded when the cursor is placed over one pixel in a scan.

Furthermore, the sample can be scanned at one magnification or resolution and then rescanned at another. Each $Y - \lambda$ image is saved as a single $Y - \lambda$ file for each increment along the sample. Thus, each image file corresponds to the excitation line of light and the emission collection region that maps through the imaging spectrograph to the CCD camera. Saving individual files permits unlimited $Y - \lambda$ image acquisition at the same Y coordinate, valuable for time-dependent studies. This feature also provides some robustness by enabling a partial $X - Y$ image to be displayed should acquisition of one or more Y scans fail for any reason. After acquiring each image, the stage is moved automatically, one or more steps in the X direction, to acquire emissions at the next line and generate the next $Y - \lambda$ file. Therefore, only the fluorescent probes in a line at a single X -location are examined at each acquisition time. Step sizes can be adjusted by the software to give: (i) the best resolution; (ii) the best 1:1 image; (iii) the minimum overlap; or (iv) the fast acquisition time.

After a variable number of rows of associated emission have been captured (typically 50–1000 scans), an image cube is loaded into memory comprising all the $Y - \lambda$ data files. A graph of the spectrum from any pixel (or bin) in the $X - Y$ plane may be displayed in a separate window by clicking on the cube (**Fig. 3(c)**). A typical acquisition scan of 250 $Y - \lambda$ files, which are used to construct a $250 \times 768 \times 128$ ($\times 12$ bit) image cube, consists of approximately 100 Mb of information.

MicroArray Scanning and Customized Excitation

Another use of this technology is to analyze spotted microarrays. The Hyperspectral Imager has been compared to commercial two-color scanners with comparable sensitivity at approximately 0.5 fluors per square micron (**Fig. 3**). Increased sensitivity (500–10000 gain) can be provided by using a microchannel plate amplifier (Model VS4-1845, Videoscope International, Ltd.) installed between the CCD camera and the imaging spectrograph.

Since the hyperspectral imaging microscope collects images that reflect only the $Y - \lambda$ plane, it contains the potential to use a single line of light as a source for image illumination. Therefore, lights systems being developed, which are capable of generating a line of light composed of variable or complex wavelengths, could be employed to precisely control excitation. The incorporation of such an excitation source in hyperspectral imaging microscopy could be useful in many clinical and basic research applications.

Further Reading

- Anguiano, A., 2000. Fluorescence *in situ* hybridization (FISH): Overview and medical applications. *Journal of Clinical Ligand Assay* 23, 33–42.
- Berland, K., Jacobson, K., French, T., 1998. Electronic cameras for low-light microscopy. *Methods in Cell Biology* 56, 19–44.
- Bernheim, A., Vagner-Capodano, A.M., Couturier, J., 2000. From cytogenetics to cytogenomics of tumors. *S-Medical Science* 16, 528–539.
- Bezrookove, V., Hansson, K., van der Burg, M., *et al.*, 2000. Individuals with abnormal phenotype and normal G-banding karyotype: improvement and limitations in the diagnosis by the use of 24-colour FISH. *Human Genetics* 106, 392–398.
- Boland, M.V., Murphy, R.F., 1999. Automated analysis of patterns in fluorescence-microscope images. *Trends in Cell Biology* 9, 201–202.
- Cardullo, R.A., Alm, E.J., 1998. Introduction to image processing. *Methods in Cell Biology* 56, 91–115.
- Garini, Y., Katzir, N., Cabib, D., Buckwald, R.A., Soenksen, D., Malik, Z., 1996. Spectral bio-imaging. In: Wang, X.F., Herman, B. (Eds.), *Fluorescence Imaging Spectroscopy and Microscopy*. New York: John Wiley and Sons, pp. 87–122.
- Garini, Y., Macville, S., du Manoir, *et al.*, 1996. Spectral Karyotyping. *Bioimaging* 4, 65–72.
- Gothot, A., Grosdent, J., Palulus, J., 1996. A strategy for multiple immunophenotyping by image cytometry: Model studies using latex microbeads labeled with seven Streptavidin-bound fluorochromes. *Cytometry* 24, 214–225.
- Gribble, S.M., Roberts, I., Grace, C., Andrews, K.M., Green, A.R., Nacheva, E.P., 2000. Cytogenetics of the chronic myeloid leukemia-derived cell line K562: Karyotype clarification by multicolor fluorescence *in situ* hybridization, comparative genomic hybridization, and locus-specific fluorescence *in situ* hybridization. *Cancer Genet Cytogenetics* 118, 1–8.
- Han, H., Gao, X., Su, J., Nie, S., 2001. Quantum-dot-tagged microbeads for multiplexed optical coding of biomolecules. *Nat Biotech.* 19, 631–635.
- Huebschman, M.L., Schultz, R.A., Garner, H.R., 2002. Characteristics and capabilities of the hyperspectral imaging microscope. *Eng in Med and Bio.* 21, 104–117.
- Inoue, T., Gliksmann, N., 1998. Techniques for optimizing microscopy and analysis through digital image processing. *Methods in Cell Biology* 56, 63–90.
- Jain, A.K., 1997. In: *Fundamental of Digital Image Processing*. New Jersey: Prentice-Hall, pp. 80–128.
- Jalal, S.M., Law, M.E., 1999. Utility of multicolor fluorescent *in situ* hybridization in clinical cytogenetics. *Genet Med.* 1, 181–186.
- Johannes, C., Chudoba, I., Obe, G., 1999. Analysis of X-ray-induced aberrations in human chromosome 5 using high-resolution multicolour banding FISH (mBAND). *Chromosome Research* 7, 625–633.
- Kruse, F., 1993. The Spectral Image Processing System (SIPS) – Interactive visualization and analysis of imaging spectrometer data. *Remote Sensing Environment* 44, 145–163.
- Lu, Y.J., Morris, J.S., Edwards, P.A.W., Shipley, J., 2000. Evaluation of 24-color multifluor-fluorescence *in-situ* hybridization (M-FISH) karyotyping by comparison with reverse chromosome painting of the human breast cancer cell line T-47. *Chromosome Research* 8, 127–132.
- Malik, Z., Cabib, D., Buckwald, R.A., Talmi, A., Garini, Y., Lipson, S.G., 1996. Fourier transform multipixel spectroscopy for quantitative cytology. *Journal of Microscopy* 182, 133–140.
- Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T., 1986. *Numerical Recipes, the Art of Scientific Computing*. New York: Cambridge University Press.
- Raap, A.K., 1998. Advances in fluorescence *in situ* hybridization. *Mutat. Res. Fundam. Mol. Mech. Mutagen* 400, 287–298.
- Sakamoto, M., Pinkel, D., Mascio, L., *et al.*, 1995. Semiautomated DNA probe mapping using digital imaging microscope: II. system performance. *Cytometry* 19, 60–69.
- Schrock, E., duManoir, S., Veldman, T., *et al.*, 1996. Multicolor spectral karyotyping of human chromosomes. *Science* 273, 494–497.
- Schultz, R.A., Nielsen, T., Zavaleta, J.R., Ruch, R., Wyatt, R., Garner, H.R., 2001. Hyperspectral imaging: a novel approach for microscopic analysis. *Cytometry* 43, 239–247.
- Schwartz, S., 1999. Molecular cytogenetics: show me the colors. *Genet Med.* 1, 178–180.
- Speicher, M.R., Ballard, S.G., Ward, D.C., 1996. Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nature Genetics* 12, 368–375.
- Wang, Y.L., 1998. Digital deconvolution of fluorescence images for biologists. *Methods in Cell Biology* 56, 305–315.
- Zhao, L., Hayes, K., Glassman, A., 2000. Enhanced detection of chromosomal abnormalities with the use of RxFISH multicolor banding technique. *Cancer Genet Cytogenetics* 118, 108–111.

Imaging Through Scattering Media

AC Boccara, Ecole Supérieure de Physique et Chimie Industrielles and Centre National de la Recherche Scientifique, Paris, France

© 2005 Elsevier Ltd. All rights reserved.

Introduction

We are often faced with the problem of seeing through scattering media such as smoke, fog, clothes, blood, tissues, ceramics, walls, etc. Tremendous effort has been made to introduce techniques which will allow us to overcome the intrinsic difficulty of selecting 'good' photons which are able to provide valuable information on structures which appear fuzzy or, for most of the time, completely hidden.

Although a number of concepts and techniques that we will discuss in this article could be applied to a large number of scattering samples, we will restrict ourselves to biological tissues.

Studying light propagation through biological tissues is becoming more and more popular among physicists, biologists, engineers, and doctors. Indeed this research domain appears to be growing rapidly as seen by the number of published articles, conferences, and international meetings. Optical techniques are attractive and are starting to compete with other well-established techniques (magnetic resonance imaging, X-ray imaging, ultrasonic imaging, positron emission tomography, etc.), because they are nonionizing, nondestructive, they can reach high resolution and are usually much cheaper. Optical techniques reveal an optical contrast which is valuable in providing morphologic as well as functional information.

Most biological tissues are weakly absorbing in the deep red and near infrared region of the spectrum; the main difficulty in performing imaging arises from the high level of scattering. Fig. 1 shows the picture of a small flash light bulb placed in the mouth: it appears fuzzy and red.

In Fig. 2 we consider various photon trajectories between a short pulsed source S and a fast detector D.

Unscattered photons are called 'ballistic' photons, their number decreases exponentially, and they are completely damped after one or two millimeters.

Then there are scattered photons which propagate close to the ballistic trajectory reaching the detector; these are named 'snake-like' photons and can be used for imaging purposes.

Most of the energy which reaches D is related to the scattered photons which may exhibit very long trajectories (typically ~ 10 times the distance S-D).

The description of various experimental methods used to overcome the difficulty introduced by scattering will be associated with the various photon families that we have mentioned: ballistic, snake-like, and scattered photons. We will first give a few orders of magnitude of the parameters which will dictate the way light interacts with the tissues. Let us emphasize that these parameters also carry the information related to the contrast mechanism. The ultimate goal is to realize an 'optical biopsy', noninvasive but as close as possible to the regular biopsy followed by histopathology.



Fig. 1 Biological tissues are highly scattering but the absorption is low in the red (as seen in this digital visible photography) and the near infrared.

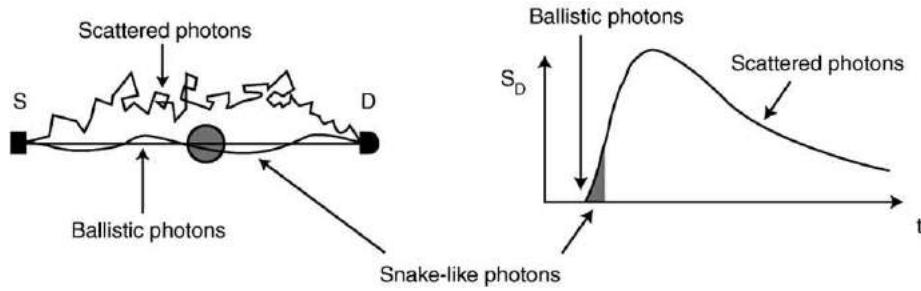


Fig. 2 Between an ultrafast light source S and a detector D can be distinguished various classes of photons which can be used for imaging.

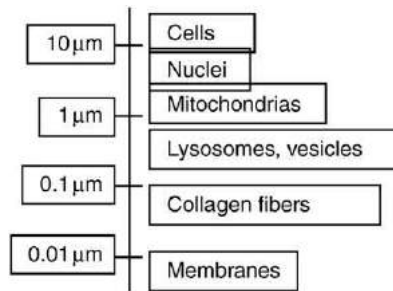


Fig. 3 Various sizes of scatterers may be found in cell structures.

Optical Parameters

Scattering Cross-Section

As shown in **Fig. 3**, biological tissues are constituted of structures which exhibit various characteristic scales: from a few nanometers in membranes to about ten micrometers for an entire cell.

As we know from scattering theories these various scales correspond to various scattering properties. For small particles ('small' means much smaller than the optical wavelength λ) we are in the Rayleigh regime which is based on the evaluation of a time-varying dipole induced by a uniform time-varying (at the light frequency scale) electric field. This dipole radiates in the far field and its scattering cross-section, which is the ratio of the scattered power to the incoming irradiance for a sphere of radius r , is given by

$$\sigma = \frac{\langle P \rangle}{I} = \left(\frac{n^2 - 1}{n^2 + 2} \right)^2 \frac{8\pi}{3} r^6 k^4 \quad (1)$$

n being the refractive index of the sphere divided by the index of the surrounding medium and $k = 2\pi/\lambda$. The scattering diagram for such particles is isotropic if we use unpolarized light. This cross-section is much smaller than the 'geometrical' section S of the particle; for instance, if $n = 1.5$, $a = 10$ nm, $\lambda = 0.6$ μm , $\sigma = 15 \cdot 10^{-8}$ μm^2 , which is much smaller than the actual section (3×10^{-4} μm^2).

For larger particles the sphere (or the ellipse) interacting with a plane polarized wave is rather complicated because we are not restricted to the dipolar approximation (uniform field), but induced multipolar moments and their radiating fields have to be taken into account. This theory was proposed by Mie and can be found in optics textbooks. As an example, **Fig. 4** shows the scattering cross-section of a micron-size particle as a function of the wavelength: a number of 'resonances' can be observed.

The scattering diagram is no longer isotropic, it exhibits lobes, and forward scattering is much higher than back-scattering. Scattering is characterized by a scattering coefficient, μ_s , which is the inverse mean free path l_s of photons between scattering events:

$$l_s = 1/\mu_s = \sigma \rho \quad (2)$$

where σ is the scattering cross-section and ρ the number of scatterers by unit volume.

The scattering coefficients μ_s of tissues are found within the range 100–1000 cm^{-1} . After a path length L , the number of ballistic photons will be damped by a factor $\exp(-L/l_s)$, so tissue imaging cannot rely only on ballistic photons as they survive only to a path length larger than about one mm. One needs to consider in more detail the scattering properties of the scatterers.

The dominant sizes of the scatterers in tissue (cells) demand consideration of the scattering in the Mie regime. Such a scattering event does not result in isotropic scattering angles. Instead, the scattering in tissue is biased in the forward direction.

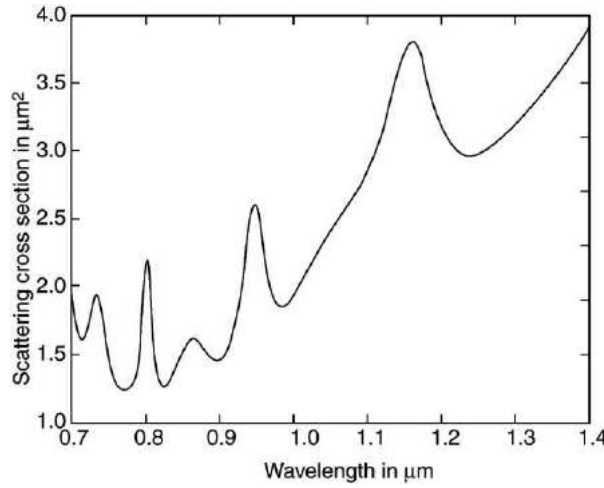


Fig. 4 In the Mie scattering regime resonances are observed when varying the wavelength (here a 1 μm sphere, refractive index: 2).

Anisotropic scattering is quantified in a coefficient, g , which is defined as the mean cosine of the scattering angle θ , $p(\theta)$ being the probability of a particular scattering angle:

$$g \equiv \langle \cos(\theta) \rangle \equiv \frac{\int_0^\pi p(\theta) \cos(\theta) \sin(\theta) d\theta}{\int_0^\pi p(\theta) \sin(\theta) d\theta} \quad (3)$$

For isotropic scattering, $g=0$. For complete forward scattering, $g=1$. In tissues, g varies from 0.7 to 0.95.

Photon migration, especially in the so-called diffusion regime, is often based on isotropic scattering such as for phonons (heat) or charge carriers scattering. For diffusion-like models, it has been shown that one may use an isotropic scattering model with a corrected scattering coefficient, μ'_s , and obtain equivalent results using:

$$\mu'_s = \mu_s(1 - g) \quad (4)$$

One can consider $l'_s = 1/\mu'_s$ as the length over which the incoming photon loses the memory of its initial direction, μ'_s is usually called the transport-corrected scattering coefficient.

Absorption Cross-Section

The absorption coefficient, μ_a (cm^{-1}) represents the inverse mean path length of photons before absorption. $1/\mu_a$ is also the distance in a medium at which intensity is attenuated by a factor of $1/e$ (Beer's Lambert law). Absorption in tissue is strongly wavelength dependent and is due to chromophores and water. Among the chromophores in tissue, the dominant component is the hemoglobin in blood. In the visible range, the blood absorption is relatively high, whereas it is much weaker in the near infrared. Water absorption is low in the visible and NIR region and increases above 1 μm , with tissues turning completely opaque above 2 μm . Thus, for greatest penetration of light in tissue, wavelengths in the 700–1300 nm spectrum are used. This region of the spectrum is called 'the therapeutic window'. Different spectra of chromophores allow one to separate the contribution of varying functional species in tissue such as quantification of oxy- and deoxy-hemoglobin to study tissue oxygenation.

Finally, when light is propagating in scattering media of thickness L , the damping of the incoming energy involves a combination of its absorption and scattering properties, and is given by

$$\exp(-L/a) \quad \text{with} \quad a^2 = 1/3\mu_a(\mu'_s + \mu_a) \quad (5)$$

This formula reflects the fact that the photon's pathlength is increased by scattering, leading to an increase of the damping, which is related to absorption.

These few examples provide a 'static' view of light behavior in scattering media. Studying in a more detailed way, the spatial and temporal distribution of light, is more difficult.

Photon Migration in Scattering Media

Photon migration models have been inspired by other fields, such as astrophysics, soft matter physics, and neutronics, where physical media are random in both space and time. To avoid difficulties linked to the microscopic complexity such investigations rely on a statistical basis.

The most widely used theory is the time-dependent diffusion approximation to the transport equation:

$$\vec{\nabla} \cdot (D\vec{\nabla}\Phi(\vec{r}, t)) - \mu_a\Phi(\vec{r}, t) = \frac{1}{c} \frac{\partial\Phi(\vec{r}, t)}{\partial t} - S(\vec{r}, t) \quad (6)$$

where $\vec{r}$ and t are spatial and temporal coordinates, c is the speed of light in tissue, and D is the diffusion coefficient related to the absorption and scattering coefficients by the formula:

$$D = \frac{1}{3[\mu_a + \mu'_s]} \quad (7)$$

The quantity $\Phi(\vec{r}, t)$ is called the fluence, defined as the power divided by the unit area. This equation does not reflect the scattering angular dependence; the use of μ'_s instead of μ_s ensures that isotropic scattering has been reached. Indeed, for the use of the diffusion theory which implies anisotropic scattering, the diffusion coefficient is expressed in terms of the transport-corrected scattering coefficient; $S(\vec{r}, t)$ is the source term.

The diffusion equation has been solved analytically for different types of measurements such as reflection and transmission modes, assuming that the optical properties remain invariant throughout the tissue. To incorporate the finite boundaries, the method of images has been used.

The diffusion approximation equation in the frequency-domain is the Fourier transform of the time-domain equation. This leads to the equation:

$$\vec{\nabla} \cdot (D\vec{\nabla}\Phi(\vec{r}, \omega)) - \left[\mu_a + \frac{i\omega}{c} \right] \Phi(\vec{r}, \omega) + S(\vec{r}, \omega) = 0 \quad (8)$$

Here the time variable is replaced by the frequency ω which is the modulation angular frequency of the source. In this model, the fluence can be seen as a complex number describing the amplitude and phase of the so-called 'photon density wave' in addition to a DC component:

$$\Phi(\vec{r}, \omega) = \Phi_{AC}(\vec{r}, \omega) + \Phi_{DC}(\vec{r}, 0) = I_{AC} \exp(i\theta) + \Phi_{DC}(\vec{r}, 0) \quad (9)$$

where θ is the phase shift of the diffusing wave whose wavelength is:

$$2\pi \sqrt{\frac{2c}{3\mu'_s\omega}} \quad (10)$$

Although analytical techniques provide a rapid understanding of most phenomena involved in static or dynamic situations, real biological samples are much more complex than homogeneous isotropic media. In these cases, direct and inverse algorithms map the spatially varying optical properties.

Monte Carlo simulations are very popular nowadays; a number of useful programs are available on the websites of various groups. Results may include time or polarization dependence of the photons along their paths.

The other method used in tissue optics is the random walk theory (RWT) on a lattice. RWT models the diffusion-like motion of photons in turbid media in a probabilistic manner. Using RWT, an expression may be derived for the probability of a photon arriving at any point and time given a specific starting point and time.

Imaging in the Ballistic Regime: Taking Advantage of Various Coherence Properties

Shallow Tissue Imaging Through Selection of Ballistic Photons

When imaging thin (less than 2 mm) tissue samples, it is possible to use photons which have not been scattered, called the ballistic photons which can be used to form high-resolution images in the same manner as if no scattering had taken place.

Their number, however, decreases exponentially with propagation distance and this ballistic signal is usually hidden by multiply scattered photons that obscure the image. For relatively shallow tissue depths it is possible to use various spatial filtering techniques to block the multiply scattered photons and there is much current research that aims to extend optical imaging and biopsy to depths of a few mm.

Confocal microscopy rejects much of the scattered light through spatial filtering and has been shown to form 3-D images at tissue depths of up to 200–300 μm , both *in vitro* and *in vivo*, which corresponds to about 50 dB of damping for the ballistic photons.

Two-photon fluorescence microscopy is particularly interesting in this regime since it ensures that all the detected fluorescence photons originate from the desired image plane. This technique also ensures a better penetration (using IR excitation) and a better resolution (nonlinear response) than single photon confocal microscopes.

Imaging to much greater tissue depths does not appear to be practical using optical microscopy and spatial filtering alone and more sophisticated techniques must also be employed to discriminate in favor of the ballistic photons. These techniques exploit either the fact that the ballistic signal photons retain spatial coherence of incident light or the fact that the ballistic photons keep the optical phase memory that multiply diffused photons have lost.

One of the most successful ballistic light-imaging techniques based on the latter filtering approach is optical coherence tomography (OCT) which uses coherence gating using low temporal coherence light and an interferometric detection. Coherent

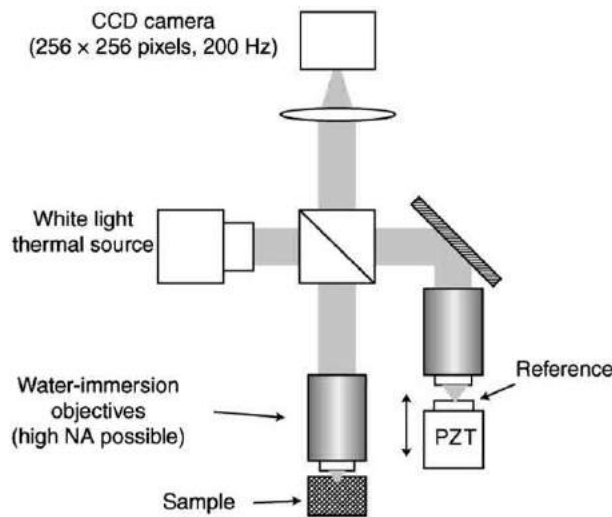


Fig. 5 The full-field OCT setup based on a Linnik interferometer. The PZT actuator is used to modulate the path difference between the two arms.

detection is used to discriminate against the multiply scattered photons that scatter back into the trajectories of ballistic photons. The ballistic light signal is measured by interference with a reference beam and the resulting pattern is detected with high sensitivity using homodyne or heterodyne detection. By using low coherence length sources and matching the time-of-flight of the reference beam signal to that of the desired ballistic light, OCT can acquire depth-resolved images in scattering media such as biological tissues. For powers in the mW range for tissue irradiation in the infrared, the unscattered ballistic (or more precisely single back scattered) component of the light will be limited by the shot noise detection limit after propagating through approximately 25 scattering mean free paths (MFP) of a scattering medium (or more than 100 dB). This corresponds to a 'typical' tissue thickness of about 1 mm tissue depth for the reflection geometry which is potentially useful for clinical applications such as screening for skin cancer, or for retinal examination.

OCT has proved clinically useful as a means of acquiring depth-resolved images in the eye and is being developed for application in strongly scattering biological tissue. Although the image acquisition requires pixel-by-pixel scanning, the use of new high-power superluminescent diodes in the 100 mW range or of ultrafast lasers to provide high average power, low coherence length radiation (sub-two cycle lasers) has resulted in an *in vivo* OCT system providing real-time imaging.

There are a number of other approaches to coherent imaging, including whole-field 2-D acquisition techniques which remove the need to scan transversely pixel by pixel.

Whole-field imaging using a temporally and spatially incoherent source and a well-balanced interferometric microscope intrinsically provides higher image acquisition rates and can take advantage of inexpensive spatially incoherent, broadband sources such as high-power LEDs, and white light thermal sources.

As an example, Fig. 5 shows Linnik-type white light interferometer with optical path modulation and Fig. 6 an image of an *ex-vivo* rat retina where one can recognize the main feature expected from histology.

Holography, which works in the Fourier space rather than in the object or image space, is also an interferometric technique able to discriminate between ballistic and scattered photons. Since the development of electronic holography, which takes advantage of available arrays of detectors with multimillion pixels, this approach is rapidly growing.

Real-time 3-D imaging systems based on photorefractive holography in multi-quantum-wells devices, which can potentially acquire depth-resolved images at thousands of frames per second, have been demonstrated.

These ballistic light techniques can extend the depth of tissue imaging to the mm range when imaging in reflection. When it is necessary to image through cms, rather than mms of tissue, however, there will be no detectable ballistic light signal and one must extract useful information from the scattered photons.

Imaging in the Snake-Like Regime: Taking Advantage of Time (or Frequency) Gating

The first approach is to use time gating to select the earliest arriving scattered light, which will provide the most useful image information.

For moderate tissue depths one can exploit the fact that biological tissue is highly forward scattering, with most photons being only slightly deviated from their original direction upon each scattering event. After a few cm of tissue, there can be a significant number of photons that have followed reasonably well-defined 'snake-like' paths about the original direction through the tissue, arriving at the detector after the ballistic light but before the fully diffuse photons whose trajectories are effectively randomized.

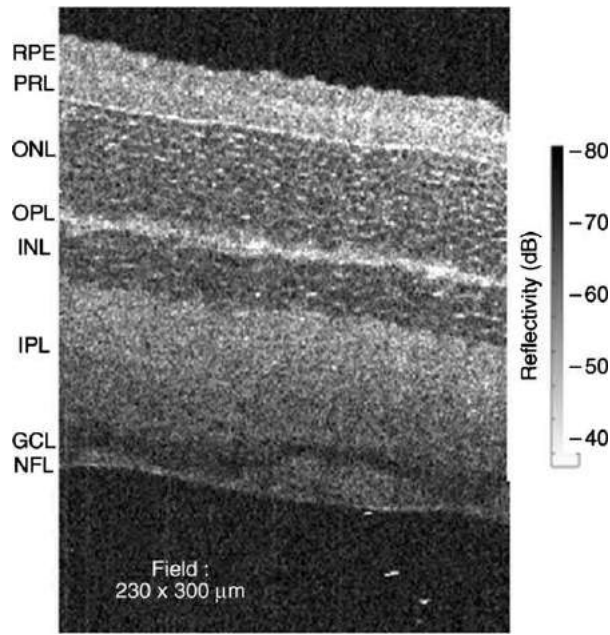


Fig. 6 OCT rat retinal image: retinal pigmented epithelium. PRL: photoreceptor layer. ONL: outer nuclear layer. OPL: outer plexiform layer. INL: inner nuclear layer. IPL: inner plexiform layer. GCL: ganglion cell layer. NFL: nerve fiber layer.

This early arriving light can be selectively gated using ultrafast lasers and a variety of detectors such as streak cameras, time-gated optical image intensifiers, and time-gated fast photon counting systems, which provide picoseconds time gates.

An alternative, sometimes cheaper, approach to temporal gating is the use of (high) frequency signal discrimination (which is the Fourier transform of the pulse experiment). The periodic solution of the diffusion equation with a periodic source term gives rise to the so-called 'photon density waves'. Such diffusive waves, like thermal waves, are used in the near-field regime (their amplitudes are damped by $\exp(-2\pi)$ after propagating over a single wavelength path) and they can reveal subsurface structures through their refractive and diffusive behavior. Coherence gating and polarization gating can be used when the early arriving light still retains some degree of coherence of the incident light. Indeed polarization discrimination has been used to select those photons which undergo few scattering events and therefore preserve a fraction of their initial polarization state, from the photons which experience multiple scattering resulting in a complete loss of memory of their initial polarization state.

As an example, **Fig. 7** shows the image of sub-mm light sources (LEDs) taken through 1 cm of polystyrene solution ($\mu'_s \sim \mu_s \sim 10 \text{ cm}^{-1}$) by selecting photons which kept the memory of their initial circular polarization (here a few less than 1%). One can see the sources quite clearly.

For many important biomedical applications, such as mammography or imaging brain function, it is necessary to penetrate through 5–10 cm of tissue, after which all the detected signal is diffuse. Detecting earlier arriving photons can provide more information but, if the time window is too narrow, the detected signal becomes too weak to use with acceptable data acquisition times.

This problem is, however, being addressed with some success using statistical models of photon transport, with different degrees of approximation ranging from full Monte Carlo simulation of photon propagation to the diffusion equation. The approach is to address the inverse problem, i.e. to measure the scattered light signal as comprehensively as is practical and to calculate what distribution of absorption and scattering properties would have produced the measured signal. This provides a means to quantify the optical properties of biological tissue, averaged over a volume corresponding to a particular distribution of photon paths, and to form relatively low-resolution tomographic images.

Multiplying the number of sources and detectors and correlating the various intensity distributions over the whole structure under examination gives, as expected, better precision in localizing the size and position of the local structure to be identified.

Despite the multiple input and multiple output approaches, real-life situations with the previously mentioned problem of the inhomogeneity of the optical properties of human tissues means that the results are still poor in term of resolution and sub-centimeter structures can hardly be observed in breast imaging, for instance.

One important exception to this limited success in finding precisely the position, the size, and the optical properties of hidden structures is the case of 'activation' studies. This corresponds to a 'differential' situation in which an unknown structure is locally changing with time. Here the goal is not to reach a full description of the unknown structure but only to characterize the local changes induced by the activation.

The main field of application of this approach is certainly in brain activation studies. A matrix of light emitters (usually at two wavelengths which differentiate oxy- and deoxy-hemoglobin, for instance 780 and 840 nm) is coupled to a matrix of light

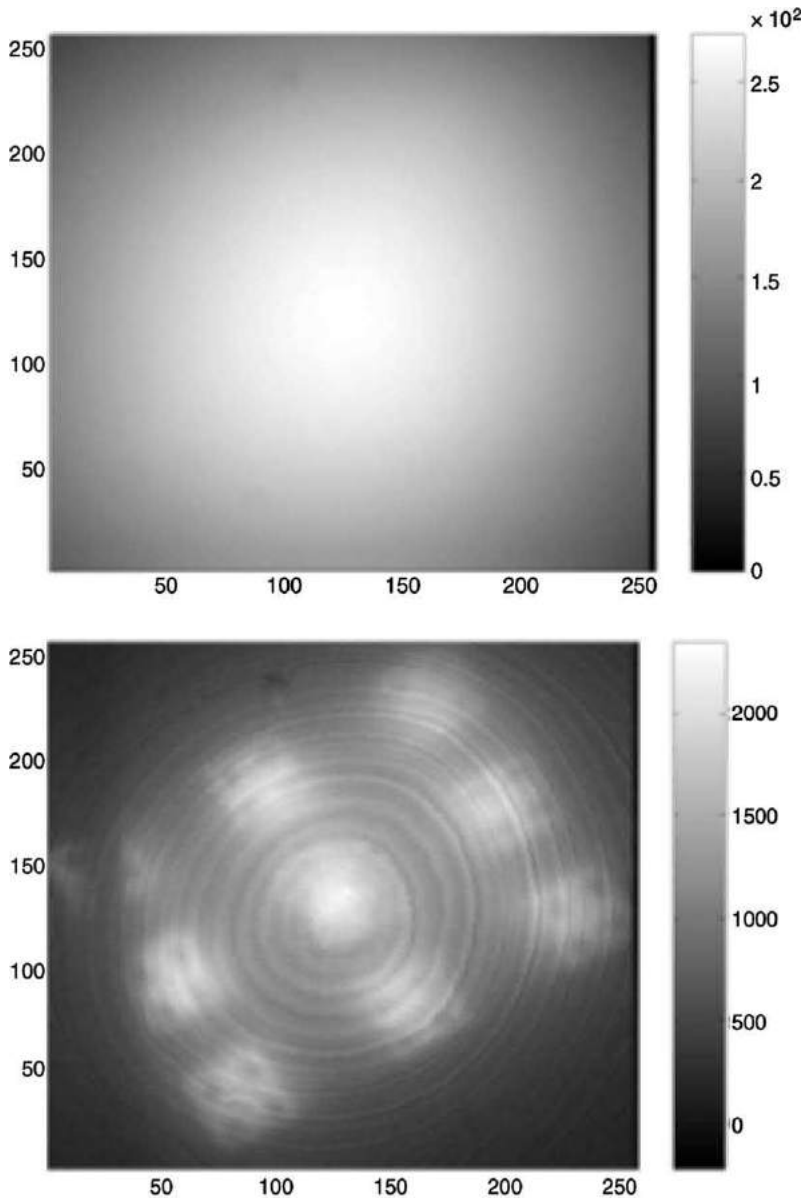


Fig. 7 Polarization imaging: here the degree of circular polarization is measured in the upper image. The lower image shows the field without polarization filtering.

detectors. The signals are usually balanced and any local change which breaks the symmetry of the scattered photon paths induces a differential signal, easy to detect, because there is no background signal before activation takes place.

Moreover it has been demonstrated that for weak perturbations (weak variations of the optical parameters), the inverse problem can be rigorously solved. This geometrical selection provides a cheap and simple setup with rather spectacular results. Sometimes a modulation is added and the phase shift of the photon density wave is enhanced by the differential approach.

Imaging in the Multiple Scattering Regime: Taking Advantage of all the Photons

Light Only: Transillumination and Fluorescence Imaging

The optical systems used for CW measurements are rather simple and require only the alignment of a number of light sources and detectors. Compared to time-gated techniques due to multiple scattering a much stronger path length distribution is present in the signal and the result is a loss of localization and resolution. In the so-called transillumination geometry, collimated detection is used to isolate multiply scattered photons close to the normal emergence angle.

Direct imaging with the required resolution using CW techniques in very thick tissues, such as breast tissue, has not been established even with sophisticated multi-sources/multi-detectors combinations (typically a few tens of sources and detectors at three different wavelengths).

Despite rapid progress in the modeling of light behavior in scattering media it has not yet been demonstrated that CW imaging can provide images with suitable resolution in thick tissue and time-dependent measurement techniques discussed earlier are dominantly used.

For fluorescence imaging, one has to account for the strong attenuation of light as it passes through tissue. Fluorescent dyes (porphyrin, indo-cyanine, and more recently quantum dots) have been developed that are excited and re-emit in the 'biological window' at near-infrared (NIR) wavelengths, where scattering and absorption coefficients are lower, which is favorable for deep fluorescence imaging. Fluorescent intensity in deep tissues depends on the location, size, concentration, and intrinsic characteristics (e.g., lifetime, quantum efficiency) of the fluorophores, and on the scattering and absorption coefficients of the tissue at both the excitation and emission wavelengths. In order to extract intrinsic characteristics of fluorophores within tissue, it is necessary to describe the statistics of photon path lengths which depend on all these different parameters. The amount of light emitted at the site of the fluorophore is directly proportional to the total amount of light impinging on the fluorophore.

Using fluorescence contrast has the unique feature of carrying two spectroscopic contrasts related to both excitation and fluorescent photons.

The next two techniques take advantage of another way to handle the optical information by coupling optics and acoustics.

Coupling Light and Sound: Acousto-Optic and Photo-Acoustic

Although the two approaches that we will now describe are different in their basic principles as well as in the experimental setups, their names are similar – opto-acoustics (sometimes called photo-acoustics) and acousto-optics.

Opto- (or Photo-)Acoustics

In this technique a pulsed (or modulated) electromagnetic beam irradiates the structure under examination. Despite their tortuous paths (about 10 to 100 cm) photons propagate through the volume in a few nanoseconds. Local absorbing centers then absorb the electromagnetic energy, they experience a fast thermal expansion, and become sources of acoustic waves.

The optical problem of tomographic reconstruction is then transformed into a simpler problem of acoustic reconstruction of source distribution because the speed of the sound is much slower than the speed of the light. More precisely we can compute the signal generated by a short light pulse impinging on a scattering medium which contains an absorber (thickness e) as can be seen on Fig. 8, when $e\mu_a \ll 1$.

The amount of power per unit surface which is deposited in the absorbing slice is $E\mu_a$ (E is the fluence in W/m^2). When the light pulse duration τ is short enough, despite the propagation in the scattering medium ($\tau \ll e/v$ where v is the speed of the sound), we find a variation of the mass by volume unit:

$$\Delta\rho \sim \rho\beta\Delta T \sim \rho \frac{\beta E\mu_a e}{C_v e} \quad (11)$$

The corresponding pressure is:

$$\Delta p = \frac{v^2}{\gamma} \Delta\rho \quad \text{or} \quad \Delta p \sim \rho \frac{v^2 \beta E}{C_p} \mu_a \quad (12)$$

Although this approach is new, compared to the purely optical transillumination approach, it has led to a large number of studies in which time gating brings new information to the accurate determination of the US emitter depth. It has been applied to breast tumor examination to complement standard imaging techniques such as echography and X-ray mammography. Recently this technique has demonstrated submillimeter resolution for *in vivo* blood distribution in rat brain imaging.

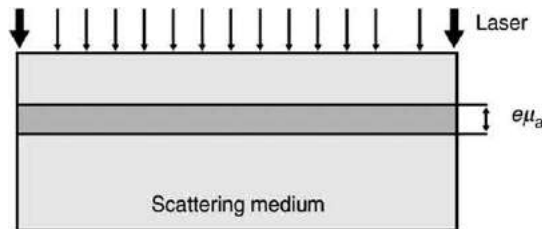


Fig. 8 Photo acoustic signal generation: 1-D model.

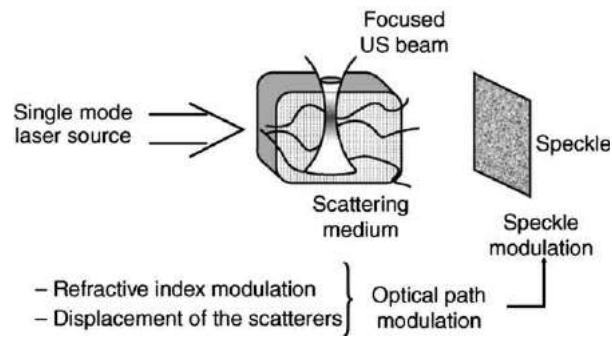


Fig. 9 Acousto-optic imaging principle: photons are tagged by the acoustic field.

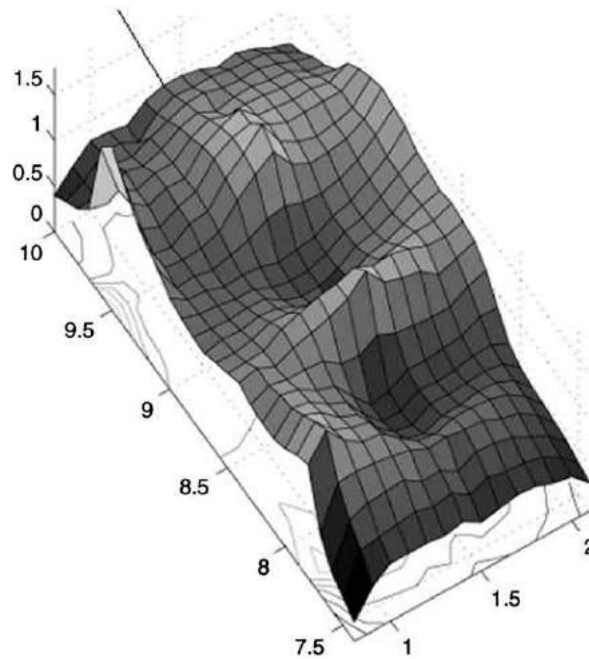
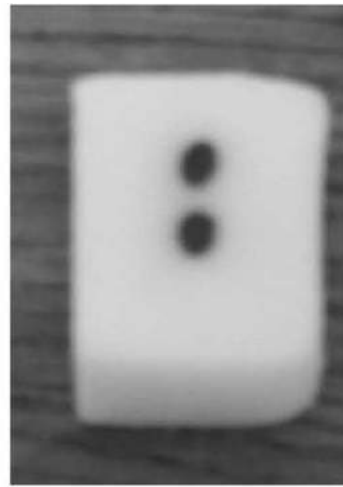


Fig. 10 Acousto-optic imaging (lower image) of two absorbing spheres embedded in a scattering medium (upper image shows a section of the gel sample).

Acousto-Optic

Here a DC single frequency laser is used to illuminate the sample. For a typical biological sample a few cm thick the trajectories are distributed between the sample thickness value and more than ten times this value. Due to the good temporal coherence of the source all the scattered photons are likely to interfere when they emerge from the sample volume. These interferences are seen as a speckle field which can be observed at a suitable distance from the sample surface.

As seen in [Fig. 9](#), adding an ultrasonic field will mainly induce a small (smaller than the optical wavelength) displacement of the scatterers (it also induces local variation of the refractive index) which will cause a speckle modulation at the ultrasonic frequency.

If the zone irradiated by the ultrasonic field is optically absorbing, less photons will emerge from this zone than from a nonabsorbing one and thus less modulation will be seen when scanning across the sample volume. This modulation can be detected by selecting a single speckle grain and averaging over a number of uncorrelated speckle distributions to obtain a suitable average value (for instance with latex particles). This is not always possible, in particular because the overall geometry is rather stable in a semisolid tissue. Using a single detector which receives many speckle grains gives less contrast. A multiple detector and a parallel processing of many speckle grains improve greatly the speed and the signal-to-noise ratio: a typical image of absorbing spheres, embedded in a scattering phantom, is shown in [Fig. 10](#).

Conclusion

The complexity associated with multiple scattering in turbid media, such as tissues of the human body, prevents an easy use of light as a routine tool for in-depth examination, whereas superficial examinations have always been sources of diagnostics.

Optical techniques will obviously progress to better characterization of tissues' optical properties, improvement of laser properties and detection techniques, as well as new mathematical approaches of the inverse problem.

We believe that this field of research is still open to new ideas, and new experimental schemes leading to new breakthroughs in the delicate problem of using light in highly scattering media.

Acknowledgements

I would like to warmly thank all the research staff (Arnaud Dubois, Benoit Forget, François Ramaz, Vincent Lorient) and PhD students (Michael Atlan, Kate Grieve, Gael Moneron, Juliette Selb) at ESPCI whose results have been used to illustrate this article. I also acknowledge discussions with my colleagues Pr Paul French (Imperial College London) and Dr Amir Gandjbakhche (NIH Bethesda).

See also: Wavefront Sensors and Control (Imaging Through Turbulence)

Further Reading

- Boccara, A.C., Fink, M. (Eds.), 2001. *Compte Rendu de l'Académie des Sciences, Physics and Astrophysics série IV tome 2, no. 8. Optical and Acoustical Imaging of Biological Media*. Paris: Elsevier.
- Born, M., Wolf, E., 1997. *Principles of Optics*. New York: Oxford University Press.
- Goodman, J.W., 2000. *Statistical Optics*. New York: Wiley-Interscience.
- Hecht, E., Zajac, A., 1979. *Optics*. Reading, MA: Addison-Wesley.
- Sebbah, P. (Ed.), 2001. *Waves and Imaging through Complex Media*. Dordrecht: Kluwer Academic Publishers.
- Van de Hulst, H.C., 1981. *Light Scattering by Small Particles*. New York: Dover.
- Van Tiggelen, B., Skipetrov, S. (Eds.), 2003. *Wave Scattering in Complex Media: From Theory to Applications*. Dordrecht: Kluwer Academic Publishers.

Infrared Imaging

K Krapels and RG Driggers, US Army Night Vision and Electronic Sensors Directorate, Fort Belvoir, VA, USA

Published by Elsevier Ltd.

Introduction

All objects above 0 Kelvin emit electromagnetic radiation associated with the thermal activity on the surface of the object. For terrestrial temperatures (around 300 Kelvin), objects emit a good portion of the electromagnetic flux in the infrared part of the electromagnetic spectrum. The visible band spans wavelengths from 0.4 to 0.7 micrometers (infrared engineers typically use micrometer/ μm for wavelength rather than the nanometer or wavenumber more commonly used in other fields). Infrared imaging devices convert energy in the infrared portion of the electromagnetic spectrum into images that can be created by the human eye. The unaided human eye cannot image infrared energy because the lens of the eye is opaque to infrared radiation.

The infrared spectrum begins at the red end of the visible spectrum where the eye can no longer sense energy. It spans from 0.7 μm to 100 μm . The infrared spectrum is, by common convention, broken into five different bands (this may vary according to the application). The bands are typically defined as: near infrared (NIR) from 0.7 to 1.0 μm ; short wavelength infrared (SWIR) from 1.0 to 3.0 μm ; mid-wavelength infrared (MWIR) from 3.0 to 5.0 μm ; long wavelength infrared (LWIR) from 8.0 to 14.0 μm ; and far infrared (FIR) from 14.0 to 100 μm . These bands are depicted graphically in Fig. 1, which shows the atmospheric transmission for a 1-kilometer horizontal ground path for a 'standard' day in the United States. These types of transmission graphs can be tailored for any conditions, using sophisticated atmospheric models such as MODTRAN (from www.ontar.com). Note that there are many atmospheric 'windows', such that an imager designed with such a band selection can see through the atmosphere.

The primary difference between a visible spectrum camera and an infrared imager is the physical phenomenology of the radiation from the scene being imaged. The energy used by a visible camera is predominantly reflected solar, or some other illuminating energy, in the visible spectrum. The energy imaged by infrared imagers (Forward Looking Infrared, commonly known as FLIRs) in the MWIR and LWIR bands, is primarily self-emitted radiation. From Fig. 1, the MWIR band has an atmospheric window in the 3 to 5 μm region and the LWIR band has an atmospheric window in the 8 to 12 μm region. The atmosphere is opaque in the 5 to 8 μm region, so it would be pointless to construct a camera that responds to this waveband.

Figs. 2 and 3 are images that show the difference in the source of the radiation sensed by the two types of cameras. The visible image (Fig. 2) is light that was provided by the Sun, propagated through Earth's atmosphere, reflected off the objects in the scene, traversed through a second atmospheric path to the sensor, and then imaged with a lens and a visible band detector. A key here is that the objects in the scene are represented by their reflectivity characteristics. The image characteristics can also change by any change in atmospheric path or source characteristic change. The atmospheric path characteristics from the Sun to the objects change frequently, because the Sun angle changes throughout the day and also weather and cloud conditions change. The visible imager characterization model is therefore a multipath problem that is extremely difficult.

The LWIR image given in Fig. 3 is obtained primarily by the emission of radiation by objects in the scene. The amount of electromagnetic flux depends on the temperature and emissivity of the objects. A higher temperature and a higher emissivity correspond to a higher flux. The image shown is 'white-hot' or the whiter a point in the image the higher the flux leaving the object. It is interesting to note that trees have a natural self-cooling process since a high temperature can damage foliage. Objects that have absorbed a large amount of solar energy are hot and are emitting large amounts of infrared radiation. This is sometimes called 'solar loading'.

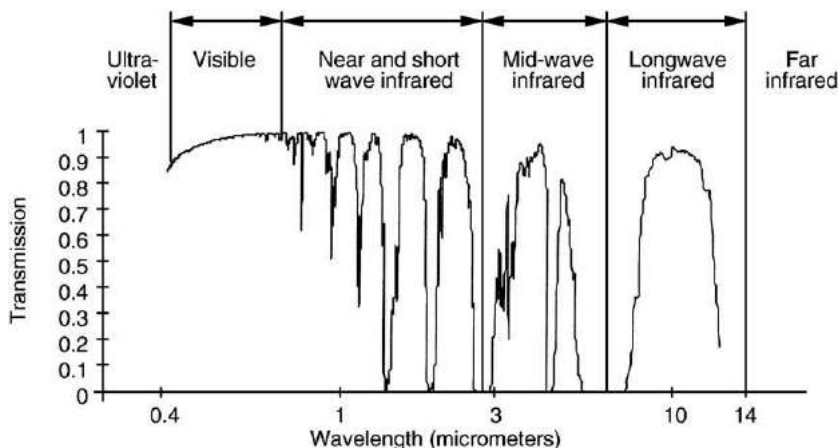


Fig. 1 Visible and infrared spectra.



Fig. 2 Visible image-reflected energy.

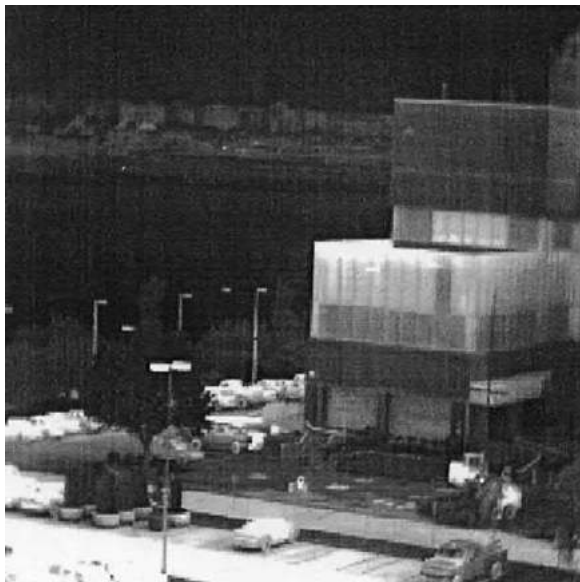


Fig. 3 Infrared image-self-emitted energy.

The characteristics of the infrared radiation emitted by an object are described by Planck's black-body law in terms of spectral radiant emittance

$$M_\lambda = \varepsilon(\lambda) \frac{c_1}{\lambda^5 (e^{c_2/\lambda T} - 1)} \text{ W/cm}^2 \mu\text{m} \quad (1)$$

where c_1 and c_2 are constants of $3.7418 \times 10^4 \text{ W } \mu\text{m}^4/\text{cm}^2$ and $1.4388 \times 10^4 \text{ } \mu\text{m K}$. The wavelength, λ , is provided in micrometers and $\varepsilon(\lambda)$ is the emissivity of the surface. A black-body source is defined as an object with an emissivity of 1.0, so it is a perfect emitter. Source emissions of black-bodies at typical terrestrial temperatures are shown in Fig. 4. Often, in modeling, the terrestrial background temperature is assumed to be 300 K. The source emittance curves are shown for other temperatures for comparison purposes. One curve corresponds to an object colder than the background and two curves correspond to temperatures hotter than the background.

Planck's equation describes the spectral shape of the source as a function of wavelength. It is readily apparent that the peak shifts to the left (shorter wavelengths) as the body temperature increases. If the temperature of a black-body is increased to that of

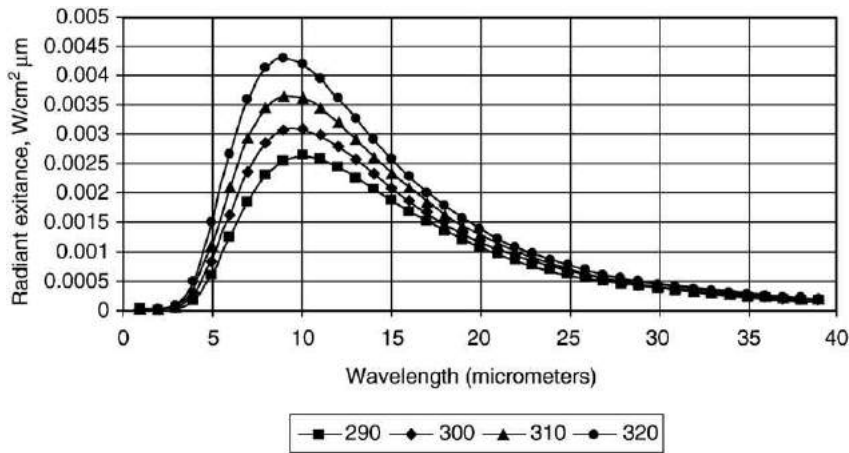


Fig. 4 Planck's blackbody radiation curves for four temperatures from 290 K to 320 K.

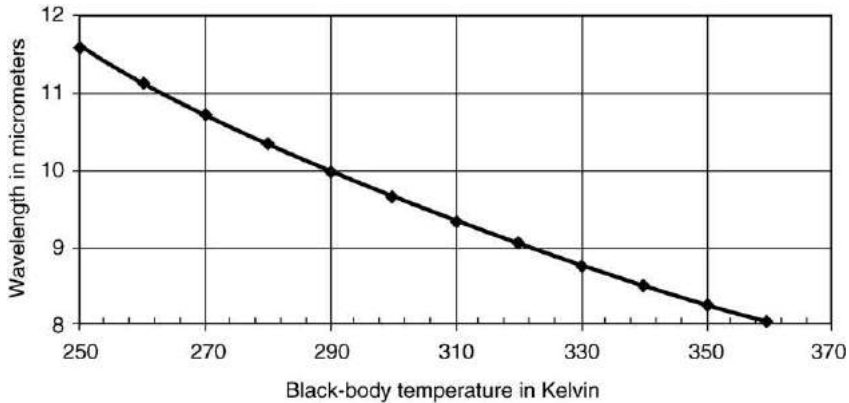


Fig. 5 Location of peak of black-body radiation, Wien's Law.

the Sun (5900 Kelvin), the peak of the spectral shape would decrease to $0.55\ \mu\text{m}$ or green light. This peak wavelength is described by Wien's displacement law

$$\lambda_{\text{max}} = 2898/T, \mu\text{m} \quad (2)$$

Fig. 5 shows the peak wavelength as a function of temperature through the longwave region. It is important to note that the difference between the black-body curves is the 'signal' in the infrared bands. For an infrared sensor, if the background is at 300 K and the target is at 302 K, the signal is the difference in flux between these curves. Signals in the infrared ride on very large amounts of background flux. However, in the visible band this is not the case. For example, consider the case of a white target on a black background. The black background is generating no signal, whereas the white target is generating a maximum signal (given that the sensor gain has been adjusted). Dynamic range may be fully utilized in a visible sensor. In the case of an IR sensor, a portion of the dynamic range is used by the large background flux radiated by everything in the scene. This flux never has a small value, hence sensitivity and dynamic range requirements are much more difficult to satisfy in IR sensors than in visible sensors. **Table 1** summarizes this characteristic numerically for the 300 K background and targets shown in **Fig. 4**.

A typical infrared imaging system scenario is depicted in **Fig. 6**. The scene consists of two major components, the target and the background. In an IR scene, the majority of the energy is remitted from the constituents of the scene. This emitted energy is transmitted through the atmosphere to the sensor. As it propagates through the atmosphere it is degraded by absorption and scattering. Obscuration by intervening objects and additional energy emitted by the path also affect the target energy. All of these contributors, which are not the target of interest, essentially reduce one's ability to discriminate the target. So, at the entrance aperture for the sensor, the target signature is reduced from the values observed at the target. Then, the signal is degraded by the optics and scanner (as applicable) of the sensor. The energy is then sampled by the detector array and converted to electrical signals. Various electronics amplify and condition this signal before it is presented to either a display for human interpretation or an algorithm like an automatic target recognizer for machine interpretation. A linear systems approach to modeling allows the components' transfer functions to be treated separately as contributors to the overall system performance. This approach allows for

Table 1 Signal/dynamic range limitation in IR bands

Blackbody temperature	Exitance in 8–12 μm band (W/cm^{-2})	Contrast
290 ($-10\Delta T$ K target)	$1.25\text{E}-02$	7.97%
300 (background)	$1.48\text{E}-02$	
310 ($+10\Delta T$ target)	$1.74\text{E}-02$	
320 ($+20\Delta T$ target)	$2.02\text{E}-02$	15.39%

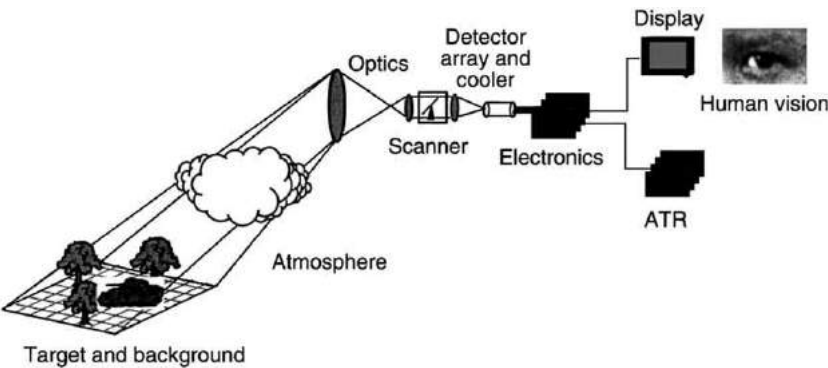


Fig. 6 Typical infrared imaging scenario.

straightforward modifications to a performance model for changes in the sensor or environment when performing trade-off analyses.

History of Infrared Imaging

Night vision systems began to be developed extensively during World War II. Satisfying the requirement to image at night was approached with two different methodologies. The first method was image intensification (I^2). This method involves amplification of any small light that was available and displaying it directly to the eye. Typically I^2 devices are sensitive to visible light and a small portion of the short-wave IR (SWIR, 0.7–3.0 μm) band. They are often classified (along with visible band imagers) as electro-optical (EO) imagers. With an I^2 imager, there must be some source of illumination for them to function well (as little as starlight is sufficient for approx 20/40 acuity with current systems). The second type of night vision device is the infrared or thermal imager (also known as FLIRs). FLIRs are sensitive to the radiation in either the mid-wave IR (MWIR, 3–5 μm) or long-wave IR (LWIR, 8–14 μm) bands.

The first infrared cameras used photographic film that was sensitive to infrared wavelengths. Intelligence and reconnaissance imaging from aircraft drove the infrared system development as a night augmentation to existing visible cameras. These imagers used either continuous moving film or stationary film, like an ordinary camera. A slit aperture was used in the continuous moving film cameras and aircraft motion provided the scan of the scene along the film to build a continuous image. The introduction of television, as electronic imaging in the visible spectrum, led to the development of electronic imagers in the infrared band. Electronic infrared imagers were not restricted to looking down and using aircraft motion for scan, hence these devices came to be known as FLIRs (Forward Looking InfraRed). Today, an electronic infrared imager may be called a FLIR or a thermal imager, with the trend being to use FLIR to describe military applications.

Early FLIRs were often accompanied by illumination devices. They were not very sensitive compared to modern FLIRs so active illumination was necessary. Infrared searchlights or illuminators were used to get higher signal-to-noise ratios. These types of FLIRs are called ‘active’ IR imagers, while FLIRs that do not use illuminators are ‘passive’ imagers. Generally, active imagers are used in civil applications and are declining in military use. Broadly, applications for thermal imagers fall into two categories; military and commercial. Table 2 lists some of the purposes and users for modern thermal imagers. The design and performance criteria vary widely for some of the applications.

Types of Infrared Imagers

Infrared imagers are classified by different characteristics: scan type, detector material/cooling requirements, and detector physics. The scan type refers to the method used to construct the electronic image. The camera may use a single detector which is raster scanned over the input scene to build an image. Alternatively, a parallel scan uses a linear array of detectors scanned across a scene

Table 2 Applications of infrared imaging

<i>Category</i>	<i>User</i>	<i>Application</i>
Military	Intelligence analyst	Intelligence surveillance and reconnaissance
	Pilot, driver	Navigation/pilotage
	Gunner, weapons officer	Targeting/fire control
	Air defense	Search and track
Commercial	<i>Civil government</i>	
	Police	Surveillance, fugitive tracking
	Fire	Rescue, hot-spot location
	<i>Environmental</i>	
	EPA	Emission tracking
	Interior department	Resource management
	NIMA/USGS	Mapping
	NOAA/MWS	Weather forecasting
	<i>Industrial</i>	
	Manufacturers	Machine vision
	Maintenance	Non-destructive testing
	<i>Medical</i>	
	Doctors	Diagnostic imaging

to build an image. The latest advances in materials have led to staring arrays of detectors. In a staring system, a detector is present in two dimensions to represent each image pixel and no mechanical motion of the focal plane is necessary to construct the image. There are some hybrid FLIR types that combine the different imaging techniques. Usually this combination leads to an improvement in signal-to-noise ratio or to reduce undersampling.

The second classification, by detector material/cooling requirements, usually describes FLIRs built using differing materials. For example, a typical LWIR FLIR material is mercury cadmium telluride (HgCdTe or MCT). In order to achieve high sensitivity, these devices are cooled to decrease dark current. Usually the cooling is based on liquid nitrogen or a cryogenic cooler, and the detectors operate at 77 K. Another common detector material is indium antimonide (InSb), which is used for MWIR FLIRs and is also cooled. A newer class of infrared cameras is not cooled, being referred to as 'uncooled' FLIRs. These uncooled FLIRs are microbolometers (resistive elements) or pyrometers (capacitive elements) and have less sensitivity than cooled imagers. Typically, the cooled imagers are comprised of photon detectors while the uncooled imagers are based on thermal detectors, the uncooled FLIRs being used in lower-performance applications.

Infrared Imager Performance

There are three general categories of infrared sensor performance characterizations. The first is sensitivity and the second is resolution. When end-to-end, or human-in-the-loop (HITL), performance is required, the third type of performance characterization describes the visual acuity of an observer through a sensor. The former two are both related to the hardware and software that comprises the system, while the latter includes both the sensor and the observer. Sensitivity is determined through radiometric analysis of the scene/environment and the quantum electronic properties of the detectors. Resolution is determined by analysis of the physical optical properties, the detector array geometry, and other degrading components of the system in much the same manner as complex electronic circuit/signals analysis.

Sensitivity describes how the sensor performs with respect to input signal level. It relates noise characteristics, responsivity of the detector, light gathering of the optics, and the dynamic range/quantization of the sensor. Radiometry describes how much light leaves the object and background and is collected by the detector. Optical design and detector characteristics are of considerable importance in sensor sensitivity analysis. In infrared systems, noise equivalent temperature difference (NETD) is often a first-order description of the system sensitivity. The 3D noise model describes more detailed representations of sensitivity parameters (see Further Reading).

The second type of measure is resolution. Resolution is the ability of the sensor to image small targets and to resolve fine detail in large targets. Modulation transfer function (MTF) is the most widely used resolution descriptor in infrared systems. Alternatively, it may be specified by a number of descriptive metrics such as the optical Rayleigh criterion or the instantaneous field-of-view of the detector. Where these metrics are component-level descriptions, the system MTF is an all-encompassing function that describes the system resolution.

Sensitivity and resolution can be competing system characteristics and they are the most important issues in initial studies for a design. For example, given a fixed sensor aperture diameter, an increase in focal length can provide an increase in resolution, but may decrease sensitivity. Typically, visible band systems have plenty of sensitivity and are resolution-limited, whereas infrared imagers have been more sensitivity-limited. With staring infrared sensors, the sensitivity has seen significant improvements.

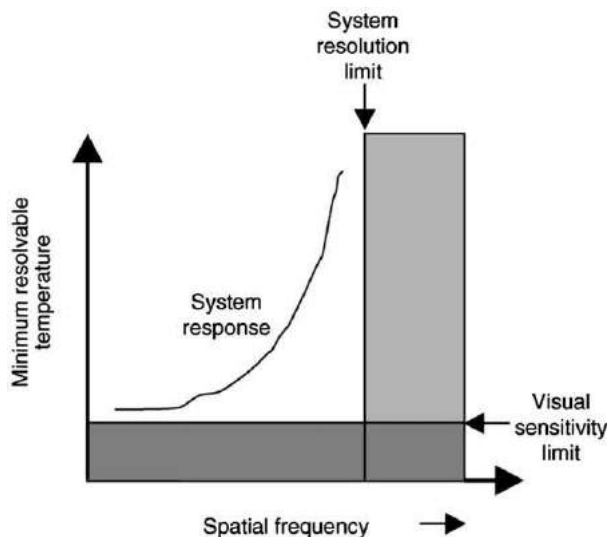


Fig. 7 IR imager performance.

Often metrics, such as NETD and MTF, are considered to be separable. However, in an actual sensor, sensitivity and resolution performance are not independent. As a result, minimum resolvable temperature difference (MRT or MRTD) has become the primary performance metric for infrared systems. MRT is a quantitative performance measure in terms of both sensitivity and resolution and a simple MRT curve is shown in Fig. 7. The performance is bounded by the sensor's limits and the observer's limits. The temperature difference (or thermal contrast) required to image smaller details in a scene increases with detail size. The inclusion of observer performance yields a single sensor performance characterization, MRT. It describes sensitivity as a function of resolution, and includes the human visual system.

Many different sensor characteristics may be considered to increase the fidelity of a sensor model. A list of typical infrared imaging system parameters is given in Fig. 8. When considering performance in the system sense, the groupings are often blocks in the system diagram, each with separate transfer functions. This approach works well unless the linear shift invariance assumption is not valid.

General Characteristics of Infrared Imagers

There are a number of imaging characteristics that make infrared systems a little different than conventional visible imaging systems. These include source flux levels, detector charge storage and sensitivity, detector size, diffraction blur, sampling, and uniformity characteristics.

First, we consider the source flux levels. The daytime and night-time flux levels (in photons per second per square centimeter) on Earth in the visible (0.3 to 0.7 μm) is 1.5×10^{17} and around 1×10^{10} , respectively. In the mid-wave (3 to 5 μm), the daytime and night-time flux levels are 4×10^{15} and 2×10^{15} μm where the flux is a combination of emitted and solar reflected flux. In the longwave, the flux is primarily emitted where both day and night yield a 2×10^{17} μm frequency, the AMOP drops rapidly to zero. The effect is to extend the prediction of the MRT beyond the half sample rate and introduce a parameter called minimum temperature difference perceived (MTDP) to replace MRT in the range performance estimate and also in the lab to characterize an undersampled system.

The concept of MTDP follows from observation of standard 4-bar images as the pattern frequency passes the imager half sample rate. The image changes from four bars to three, two, and finally one (can be perceived). The phase of the target image on the detector array may need to be adjusted to observe this progression. The standard MRT measurement requires that all four bars be resolvable by the observer during the measurement process. For use in the lab, the MTDP is defined as the minimum temperature difference at which two, three, or four bars can be resolved in the image of the standard 4-bar test pattern by an observer, with the test pattern positioned at optimum phase. The optimum phase is the phase at which two, three, or four bars are 'best perceived'. TRM3 uses the standard definition of MRT as for adequately sampled imagers. The TRM3 Approach Model is implemented if the imager is considered undersampled as defined when the prefilter MTF at the half-sample frequency is larger than 10%. The MTDP equation is given by

$$\text{MTDP}(f) = \frac{\frac{\pi}{2} \text{SNR}_{\text{th}} \Psi(f)}{\text{AMOP}(f)} \quad (3)$$

where SNR_{th} is a threshold signal-to-noise ratio and Ψ is the total system filtered noise. MTDP is used in the same manner as MRTD in range calculations.

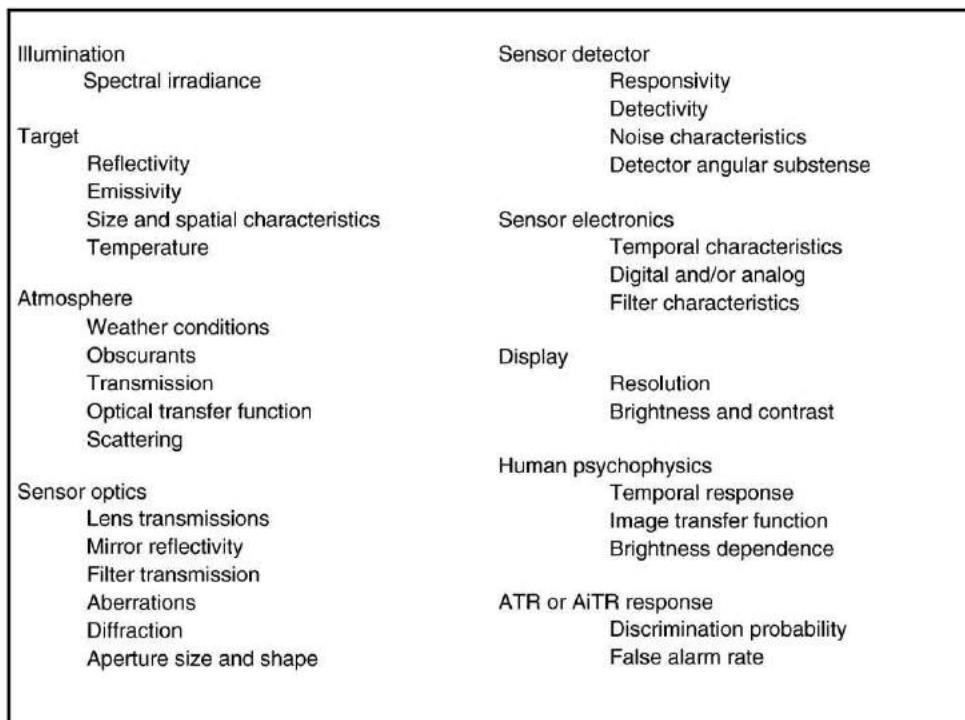


Fig. 8 Factors in infrared sensor modeling.

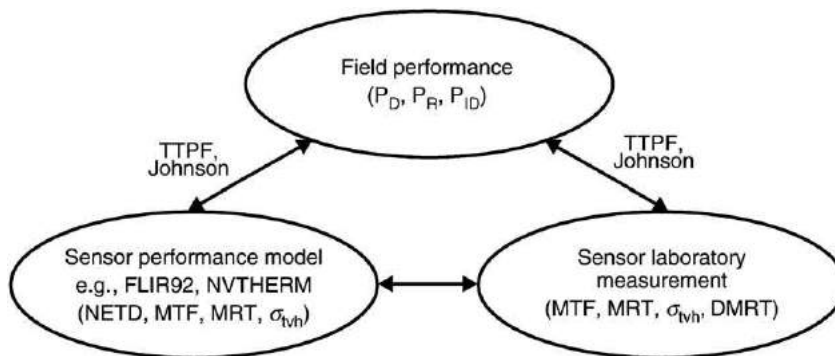


Fig. 9 Modeling, testing and performance triangle.

Another approach is the triangle orientation discrimination (TOD) threshold. In the TOD, the test pattern is a (nonperiodic) equilateral triangle in four possible orientations (apex up, down, left, or right), and the measurement procedure is a robust 4AFC (four alternative forced-choice) psychophysical procedure. In this procedure, the observer has to indicate which triangle orientation he sees, even if he is not sure. Variation of triangle contrast and/or size leads to a variation in the percentage correct between 25% (complete guess) and 100%, and by interpolation the exact 75% correct threshold can be obtained. A complete TOD curve (comparable to an MRTD curve) is obtained by plotting the contrast thresholds as a function of the reciprocal of the triangle angular size (Fig. 9).

The TOD method has a large number of theoretical and practical advantages: it is suitable for under-sampled and well-sampled electro-optical and optical imaging systems in both the thermal and visual domains, it has a close relationship to real target acquisition, and the observer task is easy. The results are free from observer bias and allow statistical significance tests. The method may be implemented in current MRTD test equipment with little effort, and the TOD curve can be used easily in a target acquisition (TA) model such as ACQUIRE. Two validation studies with real targets show that the TOD curve predicts TA performance for under-sampled and well-sampled imaging systems very well. Currently, a theoretical sensor model to predict the TOD (comparable to NVTherm or TRM3) is under development. The lab measurement and field performance appear sound, but the model is not yet available.

Testing Infrared Imagers: NETD, MTF, and MRTD (Sensitivity, Resolution and Acuity)

In general, imager sensitivity is a measure of the smallest signal that is detectable by a sensor. Sensitivity is determined using the principles of radiometry and the characteristics of the detector. For infrared imaging systems, noise equivalent temperature difference (NETD) is a measure of sensitivity. The system intensity transfer function (SITF) can be used to estimate the NETD, which is the system noise rms voltage over the noise differential output. NETD is the smallest measurable signal produced by a large target (extended source).

Eq. (4) describes NETD as a function of noise voltage and the system intensity transfer function. The measured values are determined from a line of video stripped from the image of a test target, as depicted in Fig. 10. A square test target is placed before a black-body source. The ΔT is the difference between the black-body temperature and the mask. This target is then placed at the focal point of an off-axis parabolic mirror. The mirror serves the purpose of a long optical path length to the target, yet relieves the tester from concerns over atmospheric losses to the temperature difference. The image of the target is shown in Fig. 10(b). The SITF slope for the scan line in Fig. 10(c) is the $\Delta S / \Delta T$ where ΔS is the signal measured for a given ΔT . The N_{rms} is the background signal on the same line.

$$NETD = \frac{N_{rms}[\text{volts}]}{SITF_Slope[\text{volts/K}]} \quad (4)$$

Resolution is a general term that describes the size of resolvable features in an imager's output. With infrared imaging systems, the resolution is described by the system modulation transfer function (MTF). Consider Fig. 11 for the determination of MTF, from either point spread function or edge spread function (psf or esf). In Fig. 11(a), a test target is placed at the focal point of a collimator. The target may be a point, edge, or line as shown in Fig. 11(b). The IR system images this target. The output of a detector scan, or a line of detectors in the staring array case, is taken for analysis, in Fig. 11(c). For the edge spread function, the spatial derivative is taken to get the psf. The Fourier transform is then calculated to give the system MTF as shown in Fig. 11(d). The point spread function is really the impulse response of the system, so a smaller psf is desirable, where a wide MTF is desirable. Such a psf/MTF combination gives a higher resolution.

The MRTD of an infrared system is defined as the differential temperature of a four bar target that makes the target just resolvable by a particular observer. It is a measure of observer visual acuity when a typical observer is using the infrared imager. It results in a descriptor, which is a function not just of a single value. It is a plot of sensitivity as a function of spatial frequency. This parameter can be extended to field performance using Johnson's criteria.

As a laboratory measurement, it uses a 7:1 aspect ratio bar target. The procedure is shown in Fig. 12. A bar target mask, with a black-body illuminator, is placed at the focal plane of a collimator (Fig. 12(a) and 12(b)). The temperature is increased from a small value until the bars are just resolvable to a trained observer (Fig. 12(c)). Then the temperature difference is decreased until the bars are no longer visible. These are averaged for a particular target. The same procedure is performed with a negative contrast bar target. The average of the positive and negative values is the MRT for a particular spatial frequency. These data points are plotted

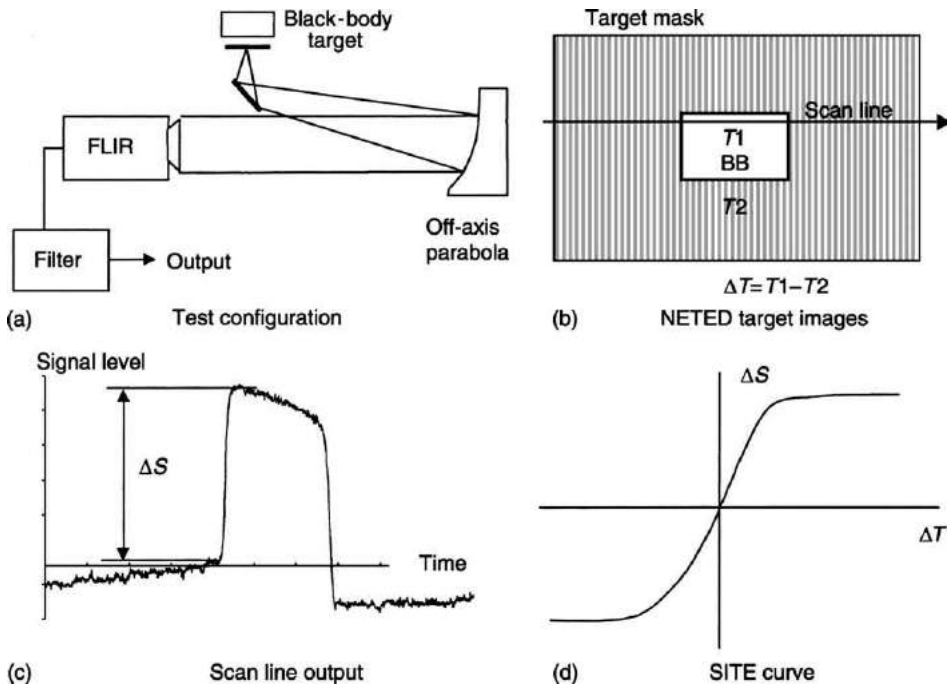


Fig. 10 NETD measurement.

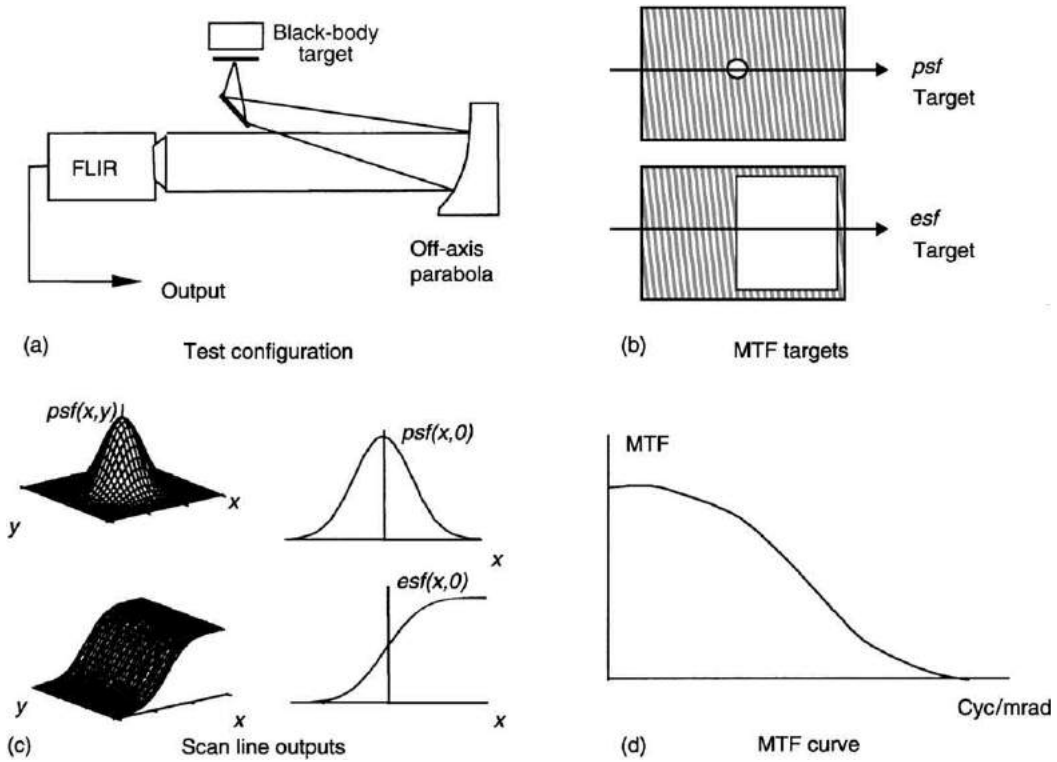


Fig. 11 MTF measurement.

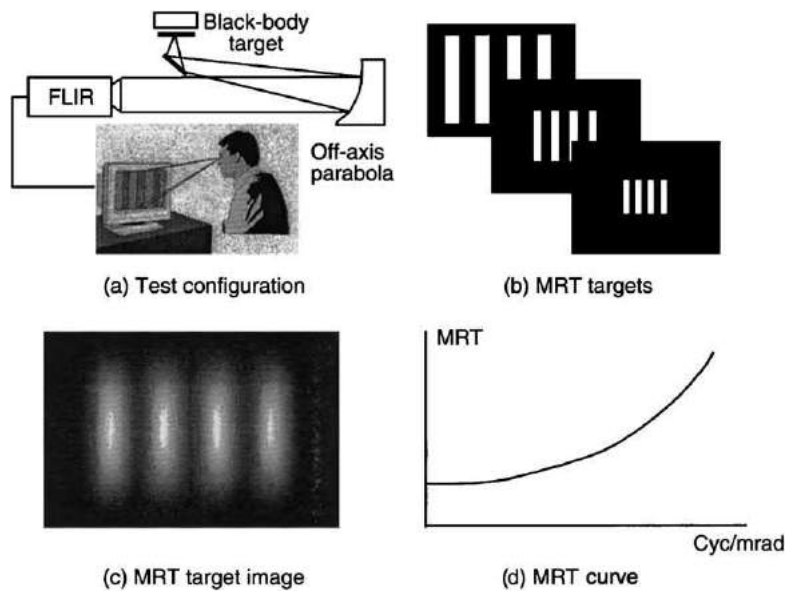


Fig. 12 MRTD measurement.

and a curve may be fitted to interpolate/extrapolate performance at other than the discrete spatial frequencies of the targets, as in Fig. 12(d).

Summary

We have provided a general description of infrared imaging systems in terms of characteristics, modeling, field performance, and performance measurement. The characteristics of infrared imagers continue to change. Significant changes have occurred in the

past five years, to include the development of higher performance uncooled imagers and ultra-narrow field of view photon systems. Large format detector arrays are commercially available in the mid-wave and are becoming more available in the longwave. Still in the research phase are dual-band focal plane level. At first sight, it appears that the longwave flux characteristics are as good as a daytime visible system; however, there are two other factors limiting performance. First, the energy bandgaps of infrared photons are much smaller than those photons in the visible, so the detectors suffer from higher dark current. The detectors are usually cooled to reduce this effect. Second, the reflected light in the visible is modulated with target and background reflectivities that typically range from 7 to 20%. In the infrared, all terrestrial objects are emitting with an emission temperature of around 300 Kelvin. Typically, a two-degree equivalent black-body difference in photon flux is considered a high contrast target to background flux. The flux difference between two blackbodies of 302 K compared to 300 K can be calculated in a manner similar to that shown in [Fig. 4](#). This is the difference in signal that provides an image, so note the difference in signal compared to the ambient background flux. In the longwave, the signal is three percent of the mean flux and in the mid-wave, it is six percent of the mean flux. This means that there is a large flux pedestal associated imaging in the infrared. Unfortunately, there are two problems accompanying this large pedestal. First, photon noise is the dominant noise and is determined by the mean of the pedestal and is compared to the small signal differences. Second, the charge well storage in infrared detectors is limited to around 10^7 charge carriers. A longwave system in a hot desert background would generate 10^{10} charge carriers in a 33 millisecond integration time. Smaller F-numbers, spectral bandwidths, and integration times are used so that the charge well does not saturate. This results in SNR of 10 to 30 times below the ideal. It has been suggested that some on-chip compression may be a solution to the well pedestal problem. In many scanning FLIR systems, the pedestal is eliminated by AC-coupling the detector signals. Infrared focal plane array (IRFPA) readout circuits have been previously designed and fabricated to perform charge skimming and charge partitioning.

Another major difference in infrared systems, from that of visible systems, is the size of the detector and diffraction blur. For infrared photon detectors, typical sizes range from 25 to 50 μm , where visible systems can be fabricated with 3 μm detectors. Also, the diffraction blur for the longwave is more than ten times larger than the visible blur and mid-wave blur is eight times larger than visible blur. Therefore, the image blur, due to diffraction and detector size, is much larger in an infrared system than a visible system. Also, it is common for infrared staring arrays to be sampling limited where the sample spacing is larger than the diffraction blur and the detector size. Dither and microscanning are frequently used to enhance performance.

Finally, both infrared photon and thermal detectors in staring arrays have responsivities that vary dramatically and it is current practice to correct for the resulting nonuniformity (nonuniformity correction or NUC). The nonuniformity can cause fixed pattern noise in the image that can limit the performance of the system even more than temporal noise (especially in static imaging or stabilized target acquisition imaging).

Modeling Infrared Imagers

Otto Schade Sr. developed the earliest imaging system models and performance measures. His work describing television in the 1950s and 1960s, pioneered the way for the characterizations of imaging sensors used by engineers today. His performance measure for still pictures, moving pictures and television systems was based on an observer resolving a three bar periodic target.

John Johnson developed the technique of relating the acquisition of military targets in the field to laboratory measurements of resolvable bar targets. He divided the discrimination tasks into four categories: detection, orientation, recognition, and identification. This technique was termed an ‘equivalent bar pattern approach’. It related performance on a simple test target to performance with complex object targets. Johnson viewed scale models and bar targets in the laboratory against a bland background. The smallest discernible barchart target yielded the maximum resolvable bar pattern frequency. These results, tabulated in [Table 3](#) as cycles across the minimum dimension, were the basis for the discrimination methodology in widespread use today. Note that detection took only one cycle, but as the tasks got more specific or difficult, the requirement went as high as 8 cycles across the target minimum dimension.

Table 3 Johnson’s criteria

<i>Target</i>	<i>Resolution per minimum dimension</i>			
	<i>Detection</i>	<i>Orientation</i>	<i>Recognition</i>	<i>Identification</i>
<i>Broadside view</i>				
Truck	0.9	1.25	4.5	8
M-48 Tank	0.75	1.2	3.5	7
Stalin Tank	0.75	1.2	3.5	6
Centurion Tank	0.75	1.2	3.5	6
Half-track	1	1.5	4	5
Jeep	1.2	1.5	4.5	5.5
Command car	1.2	1.5	4.3	5.5
Soldier (standing)	1.5	4.8	3.8	8
105 howitzer	1	1.5	4.8	6
Average	1 ± 0.25	1.4 ± 0.35	4.0 ± 0.8	6.4 ± 1.5

Table 4 Moser's data

Discrimination task	Line pairs/critical dimension
Detect ship	Aperiodic treatment
Classify as combatant	4
Recognize type	10

Recognizing that typical vehicle aspect ratios were somewhat limited, Johnson and Lawson conducted further experiments. Paul Moser conducted some of his own, reanalyzed Johnson and Lawson's, and developed the concept that this task was related to the average or critical dimension instead of the minimum dimension. This led to the conversion of line pairs to resolution elements, introducing the second dimension. Additionally, he performed a similar experiment on marine targets, which is summarized in [Table 4](#). Lloyd and Sendall developed the MRT metric, which combines both resolution and sensitivity characteristics. This measurement of MRT is described in detail in a following article, but there is a theoretical model that is used to evaluate new sensor designs and concepts. It relates the minimum temperature difference between bar pairs at which they are just resolvable as a function of spatial frequency. MRT describes the infrared imager sensitivity as a function of resolution. More specifically, MRT is a measure of thermal contrast sensitivity as a function of spatial frequency.

Ratches *et al.* developed the NVL (US Army Night Vision and Electronic Sensors Directorate formerly known as the Night Vision Laboratory) Static Performance Model, which predicted the end-to-end performance. It started with the target signature and carried through to the observer. The MRT, a laboratory measurement or a modeled value, was related to the probabilities of discrimination through the target transfer probability function, cumulative percentages related to Johnson's criteria.

D'Agostino and Webb created the 3D noise model, both spatial and temporal considerations, for analyzing system noise. This technique divided the noise components into manageable and understandable components. It also simplified the integration of these effects into models.

The introduction of focal plane array (FPA) imagers has created a requirement to improve the model performance to account for the difference between scanning arrays and staring arrays. This change has made sampling an important area for improvement. A new model called NVTHERM is a result of the NVESD model improvement effort, which addresses undersampled imaging system performance.

Sensor characterization is seen in three ways: theoretical models, field performance, and laboratory measurements ([Fig. 9](#)). Theoretical models are developed that describe sensor sensitivity, resolution, and human performance for the purpose of evaluating new conceptual sensors. Acquisition models, and other models, are developed to relate the theoretical models to field performance. This link allows theoretical models to be converted to field performance quantities (e.g., probabilities of detection, recognition, and identification). Field performance is measured in the field so that the theoretical models can be refined and become more accurate with advanced sensor developments. Since field performance activities are so expensive, methods for the direct measurement of sensor performance are developed for the laboratory. Field performance testing of every infrared sensor built, including buy-off, acceptance, and life-cycle testing, is ridiculous and out of the question. Laboratory measurements of sensor performance are developed such that, given these measurements, the field performance of a system can be predicted. Sensor characterization programs require accurate sensor models, field performance estimates with acquisition models, and repeatable laboratory measurements.

There are a few alternatives to the US NVTherm model. One candidate approach to undersampled imager modeling is Germany's TRM3, or the MDTP approach. The impact of undersampling in the image of a 4-bar target can readily be seen by simply observing the change in the image as a function of spatial frequency and phase. The spatial frequency is defined as line pairs per angular extent of the target, and the phase specifies the relative location of the target image to the detector array raster. These effects are obviously not independent of each other, but for each target there can be found an optimal phase where the observer can see the maximum amplitude modulation in the image of the target. MRT measurements in the past have utilized this variation with phase by allowing the observer to optimize the displayed image through target or system line-of-sight changes during the measurement process. For undersampled imagers, TRM3 addresses the problem of the MRT calculation's inability to predict beyond the half sample rate of the sensor, by replacing the MTF in the denominator of the MRT equation with an appropriately scaled term called the average modulation at optimum phase (AMOP). AMOP is the average signal difference in the image of the 4-bar standard pattern, with the test pattern positioned at optimum phase. AMOP oscillates between the pre-sample MTF and the bar modulation. Beyond a frequency of 0.8 times the sample rate or 1.6 times the half sample arrays and quantum well detector systems. These systems will find their place in applications to include both military and commercial sectors.

Further Reading

- Bijl, P., Valeton, M.J., 1998. Triangle orientation discrimination: the alternative to minimum resolvable temperature difference and minimum resolvable contrast. *Optical Engineering* 37 (7), 1976–1983.
- D'Agostino, J., Webb, C., 1991. 3-D analysis framework and measurement methodology for imaging system noise. *Proceedings of SPIE* 1488, 110–121.
- Driggers, R.G., Cox, P., Edwards, T., 1998. *Introduction to Infrared and Electro-Optical Systems*. Boston, MA: Artech House, p. 8.
- Holst, G.C., 1995. *Electro-Optical Imaging System Performance*. Winter Park, FL: JCD Publishing, p. 347.
- Holst, G.C., 1995. *Electro-Optical Imaging System Performance*. Winter Park, FL: JCD Publishing, p. 432.

- Johnson, J., 1958. Analysis of image forming systems. Proceedings of IIS. 249–273.
- Johnson J and Lawson W (1974) Performance modeling methods and problems. *Proceedings of IRIS*.
- Ratches JA (1973) *NVL Static Performance Model for Thermal Viewing Systems*. USA Electronics Command Report ECOM 7043, AD-A011212.
- Schade, O.H., 1948. Electro-optical characteristics of television systems. RCA Review IX (1–4),
- Sendall, R., Lloyd, J.M., 1970. Improved specifications for infrared imaging systems. Proceedings of IRIS 14 (2), 109–129.
- Vollmerhausen, R.H., Driggers, R.G., 1999. NVTHERM: next generation night vision thermal model. Proceedings of IRIS Passive Sensors 1.
- Wittenstein, W., 1998. Thermal range model TRM3. Proceedings of SPIE 3436.

Interferometric Imaging

DL Marks, University of Illinois at Urbana-Champaign, Urbana, IL, USA

© 2005 Elsevier Ltd. All rights reserved.

Interferometric imaging is image formation by measuring the interference between electromagnetic signals from incoherently emitting or illuminated sources.

Interferometric imaging attempts to improve the resolution of images by interferometrically combining the signals from several apertures. By doing so, a resolution can be achieved that is superior to that achievable through standard image formation by any of the individual apertures. This is especially useful in astronomy, where resolving more distant and smaller stellar objects is always desirable. It is also finding increasing use in microscopy. It has several advantages over conventional noninterferometric imaging methods. The method can combine the light gathered from several imaging instruments to form an image superior to that which can be formed by any one of the instruments. It can also obviate the need to produce a single very large aperture to achieve the equivalent resolution, which in many cases is of impractical size. In addition, because it can produce a phase-resolved measurement, the data can be processed more flexibly to form a computed image estimate. There are also considerable disadvantages to interferometric imaging compared to conventional imaging methods. The amount of signal gathered is far less than could be gathered with an equivalent single aperture. It is difficult to mechanically or feedback stabilize the time delay between two widely separated apertures to obtain an accurate phase measurement. Often only a small bandwidth of the source light can be utilized, further reducing the available signal. The interference component of the combined signal can be very small. Typically an interferometric image must be computed from the data, rather than being directly measured on photographic film or an electronic focal plane array. Even with these disadvantages, interferometric imaging is frequently more feasible than building a single large aperture to achieve the highest available resolution images.

In contrast to interferometric imaging, a standard image forming instrument such as a telescope has a resolution limited by the telescope mirror aperture in the absence of aberrations and atmospheric turbulence. As telescope mirrors become very large, they become bulky, extremely heavy, and difficult to manipulate, as well as deform under their own weight and temperature gradients. However, adaptive optics is being more frequently utilized to dynamically correct for instrument aberrations as well as atmospheric turbulence. The images captured by telescopes are typically intensity-only images, which are amenable to image processing and deconvolution, but nevertheless lack phase information. Interferometric imaging is an alternative, where the resolution of a large aperture can be attained by measuring the interference between individual points of the electromagnetic field within an area equivalent to the large aperture extent. The interferometric combining of light from sub-apertures to achieve the resolution of a larger aperture is called aperture synthesis.

Unlike image formation by a lens, interferometric images are never physically formed. Conventional images are formed when diverging spherical waves emanating from sources are focused by a lens to converging spherical waves on a photographic film or an electronic sensor. In this case, the image information is directly contained in the exposure at each position on the sensor. Interferometric measurements contain the image information as statistical correlations between the electromagnetic fields at pairs of spatial points. The statistical properties of electromagnetic waves are modeled by optical coherence theory. Interferometric imaging works by measuring the statistical correlations between various pairs of points in the electromagnetic field, and then uses optical coherence theory to infer what the image of the source of the radiation is.

The light that emanates from most natural and many artificial sources is incoherent. Incoherent radiation is optical random noise. It is analogous to the sound of static heard on a radio receiving no signal. Because it is random, the fluctuations of the electromagnetic field produced by an incoherent source are completely unrelated to the fluctuations of any other source. The origin of the randomness is usually from microscopic processes such as the thermal motions of electrons or the spontaneous emission of an excited atom. Because these microscopic processes tend to act independently of one another, the light produced by different sources tends to be mutually incoherent. Because incoherent sources do not radiate a deterministic field, the average amount of power they radiate is usually specified rather than the field itself, which is random.

However, as the fields propagate away from their sources, the fields can become partially coherent. To see this, consider the light waves propagating away from a single point-like incoherent source. The field produced at the source point consists of random fluctuations in time, but these fluctuations travel outward in a spherical wave at the speed of light. The fluctuations of the field of any two points an equal distance from the source will be the same, or perfectly correlated, because they arrive from the source at the same time. This is illustrated in Fig. 1(a), where a random wave is emanated from a point incoherent source. Because of the complete correlation, we can say that these two points are fully spatially coherent. So while the source itself produces random noise, the fluctuations of the field at points distant from the source become correlated because they originate from the same source.

Partially coherent waves occur when the measured field contains a superposition of the fields from more than one incoherent source. If there are now two point sources instead of one, each field point in space will receive radiation from both sources. The amount of the contribution from each source to each point in space depends on the distance between the field point and the source. In general, because both sources can be different distances from the two field points, the fluctuations at the field points will not be the same. For example, two field points that are the same distances from both sources will be perfectly correlated, or

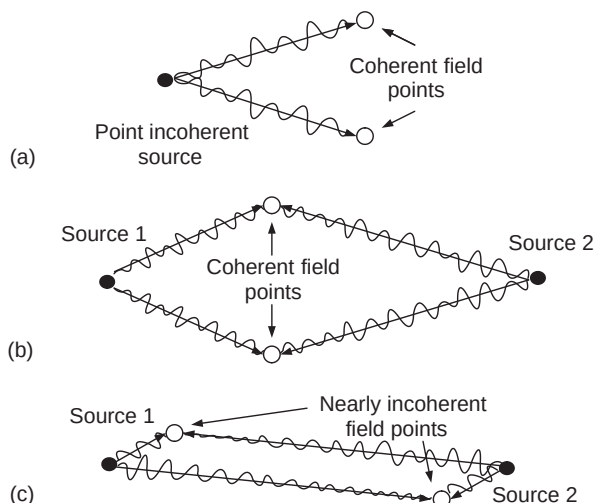


Fig. 1 The distance between incoherent sources and two field points determines the coherence between the field points. (a) Because there is only one source, the field everywhere is determined by the radiation of that source. Therefore, the same fluctuations of the wave arrive at two points an equal distance from the source at the same time, and produce identical fields at both points, and so are completely coherent. (b) Two field points are now the same distance from two incoherent sources. Because both sources contribute identical fields to both points (because the fluctuations of the wave from both sources arrive at both points at the same time), the superposition of the fields from both sources at both field points are identical, and so the fields are coherent. (c) Two field points are now near two sources. Because each field point receives radiation mostly from its nearby source and very little from the other source, the fields at each field point are almost completely mutually incoherent.

coherent, because they receive the same combination of fields from both sources. This situation is illustrated in [Fig. 1\(b\)](#). However, if one considers two points that are near both respective sources, the points will usually receive light from the nearby source. Because the sources are incoherent, and each point is mostly receiving radiation from its closer source, these points will be nearly uncorrelated, as shown in [Fig. 1\(c\)](#).

A spatial distribution of incoherent sources radiating various amounts of power will create a pattern of correlations in the field remote from the source. By using interferometry, the correlation between pairs of field points remote from the sources can be measured. The interference between two points in the electromagnetic field is achieved by relaying the two fields (e.g., with mirrors) to a common location that has an equal travel time for the fields from both points, where the fields are superimposed. The power of the superimposed beams is then detected, by a photocathode, for example. If the two signals being combined are incoherent, then no interference will take place, and the total power measured on the photodetector is simply the sum of the power of the two signals. However, if the two signals are partially coherent, there will be a deviation in the measured power from the sum of the power of the two signals. The magnitude of this deviation relative to the power of the constituent signals indicates the degree of partial coherence of the signal.

An interferometer, such as depicted in [Fig. 2](#), gathers the light from two areas of the electromagnetic field remote from an object and combines them together at the same point to form an interferogram. In this example, two telescopes of a known separation gather light from a star. The primary mirror of each telescope is of insufficient size to resolve the star as anything but a point. However, a telescope is used in practice to collect enough light to produce a measurable interference signal. Each telescope primary mirror forms a point image of the star. At the image plane of each telescope a mirror directs the point image to a measurement plane equidistant between the two telescopes. This line that joins the two telescopes is called the baseline, and it is the baseline length rather than the individual telescope mirror size that determines the minimally resolvable size. The image from each telescope is superimposed by mirrors (or perhaps beamsplitters) and imaged onto a focal plane, to get an image similar to that of [Fig. 3](#). If the light arriving from one telescope is blocked, then the image formed looks like an Airy ring pattern formed by an image, a point-like object, which is the incoherent image of [Fig. 3](#). However, when both signals are allowed to interfere, the image will have vertical stripes superimposed on it. If these stripes are very prominent, so that the image is fully darkened inside the stripes, then an image similar to the rightmost image of [Fig. 3](#) is obtained, complete interference is occurring and the two field points at which the telescopes are situated are coherent or completely correlated. If the stripes disappear, then no interference is occurring and the two field points at which the telescopes reside are incoherent. These examples are shown in [Fig. 3](#), with images of two superimposed points with varying degrees of coherence on the top row, and a line plot of the measured field through the center of the images in the bottom row. The degree to which the contrast of the stripes is enhanced by coherence is referred to as the 'modulation' and indicates the degree or magnitude of partial coherence.

In addition to having a degree of coherence, the correlation between two electromagnetic field points also has a relative phase shift. This phase shift is given by the relative position of the interference maxima, as shown in [Fig. 4](#). As the relative phase between the fields collected by the telescopes changes (which is a function of the path length difference between the light arriving at the two telescopes), the peaks of the interferogram between the two telescope images shift position. By observing the depth of the formed

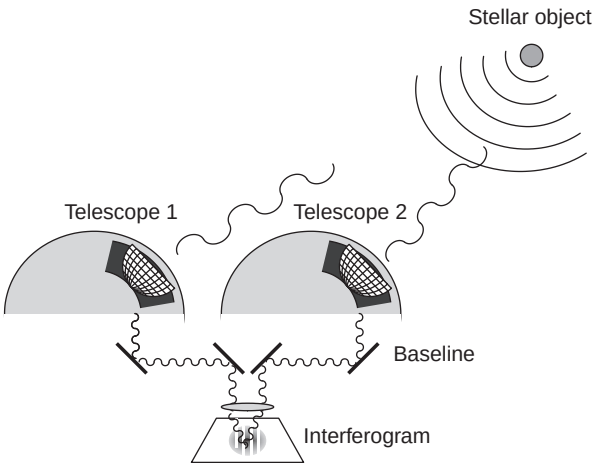


Fig. 2 Two telescopes receive radiation from a stellar object. The light is collimated and directed down the baseline to a point halfway between the telescopes, where the two images are interfered together onto a focal plane.

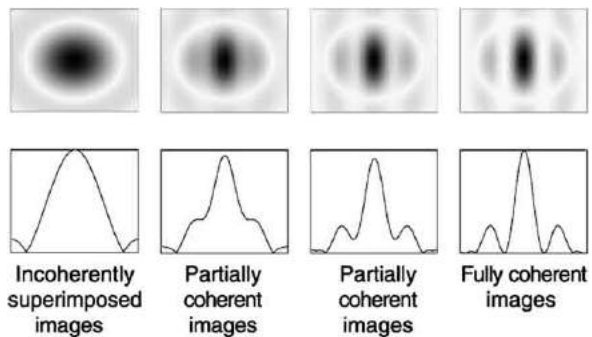


Fig. 3 The image of two interfering wavefronts from a point object with various states of coherence.

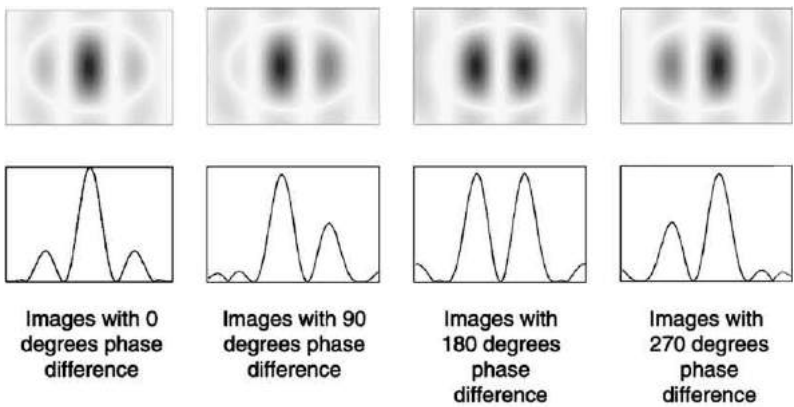


Fig. 4 The image of two interfering wavefronts from a point object with various phase differences between the two wavefronts.

interference fringes and the position of the fringes, both the amplitude and the phase of the correlation between the two field points can be simultaneously measured. The amplitude of the modulation and the phase of the interference are used to compute the complex degree of coherence that is used in the computation of interferometric images.

In practice, the phase of the complex degree of coherence is much more difficult to measure than its magnitude. This is because while the amplitude tends to change very slowly as the positions of the two correlated points change, the phase is sensitive to changes in relative path length on the order of an optical wavelength. Even though a long baseline produces better resolution, the baseline length must be known to a small fraction of a wavelength to obtain a meaningful phase measurement. Minute vibrations

in the positions of the mirrors on the baseline can easily cause fluctuations in the path length greater than a wavelength. This is observed as wild oscillations in the position of the fringes on the image plane. In addition, atmospheric turbulence can also produce large random phase shifts. Achieving a stable phase measurement is one of the most challenging engineering aspects of interferometric imaging.

To model realistic sources, we must determine the coherence produced by sources of a finite size. Finite size incoherent radiators can be regarded as an arrangement of an infinite number of point-like incoherent sources radiating various amounts of power. When an interferogram is made of a field which has two superimposed incoherent sources, the power of the interferograms that each would make alone is added together, rather than the fields of the interferograms. This is because when the fields of the two sources interfere, the phases of both sources are varying randomly over time. The two fields will randomly vary between constructively and destructively interfering with each other. Over a long time, the constructive and destructive interference averages out to no interference. We can then calculate what the interferogram of many incoherent sources combined is by computing the interferogram of a point source and adding the power of many such interferograms together.

To figure out what the power of an interferogram of an arbitrary object is, first consider what the interferogram of a single point source is. Consider a situation depicted in Fig. 5, but with a single point illuminator at the source plane positioned at (x, y, z) with radiant power I . We will have a receiving plane at a distance z from the source plane with two telescopes located symmetrically about the origin at $(\Delta x/2, \Delta y/2, 0)$ and $(-\Delta x/2, -\Delta y/2, 0)$. The fields received at the telescopes from the point will be interfered together at the center of the baseline, which is at the origin. We will assume that the point source produces a narrow band of radiation centered at wavelength λ . The point source produces a spherical wave that is received at both telescopes. The interference power is proportional to the two intensities measured at each telescope with the relative phase at the two telescopes causing constructive or destructive interference:

$$P \propto I_1 + I_2 + 2\sqrt{I_1 I_2} \cos \phi \quad (1)$$

where I_1 and I_2 are the intensities of the field at each telescope, and ϕ is the phase difference between the two fields at each telescope. We will assume that the point (x, y, z) is very far away from the telescopes, and so the intensities at the two telescopes are equal and proportional to the point intensity I , so that $P \propto I(1 + \cos \phi)$. The phase ϕ is given by the difference in distance from the the source point at (x, y, z) to the receivers at $(\Delta x/2, \Delta y/2, 0)$ and $(-\Delta x/2, -\Delta y/2, 0)$. The phase is given by the difference in the Euclidean distances:

$$\phi = \frac{2\pi}{\lambda} (\sqrt{z^2 + (x + \Delta x/2)^2 + (y + \Delta y/2)^2} - \sqrt{z^2 + (x - \Delta x/2)^2 + (y - \Delta y/2)^2}) \approx \frac{2\pi(x\Delta x + y\Delta y)}{z\lambda} \quad (2)$$

where the second term gives the difference in the distance to a first-order approximation in x and y . The interference intensity relation is then $P \propto I(1 + \cos(2\pi(x\Delta x + y\Delta y)/z\lambda))$.

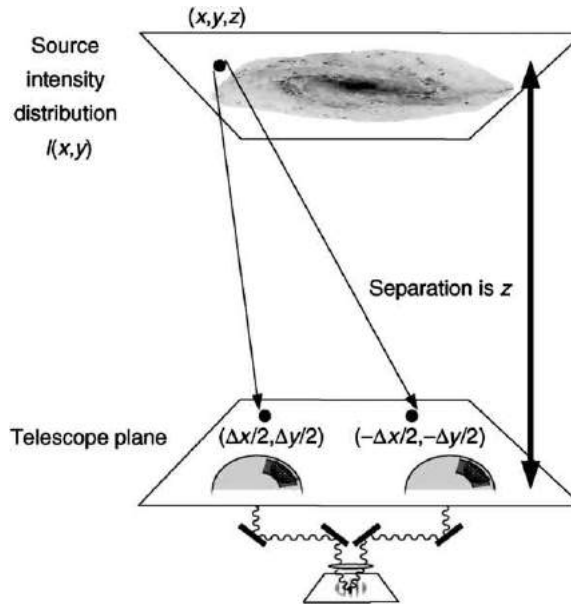


Fig. 5 Diagram of source and telescope plane used for derivation of Van Cittert-Zernike theorem. A source at a distance z from the telescope plane has a power density $I(x, y)$. The telescope plane has two movable apertures symmetric about the origin of the telescope plane separated by $(\Delta x, \Delta y)$. The light from both apertures is directed to the center of the baseline at the origin of the telescope plane, where it is interfered on the sensor.

If the source is now of a finite size, we can regard it as radiating an intensity (power per unit area) $I(x, y)$. To compute the total intensity of the interferogram, we integrate over all the sources in the source plane:

$$P(\Delta x, \Delta y) \propto \int I(x, y) \left(1 + \cos(2\pi(x\Delta x + y\Delta y)/z\lambda) \right) dx dy = \int I(x, y) dx dy + \int I(x, y) \cos(2\pi(x\Delta x + y\Delta y)/z\lambda) dx dy \quad (3)$$

This result can be interpreted as follows. The intensity of the interference when the two telescopes are separated by a distance $(\Delta x, \Delta y)$ is proportional to the real part of the two-dimensional Fourier transform of the intensity distribution $I(x, y)$ evaluated at spatial frequencies $(\Delta x/z\lambda, \Delta y/z\lambda)$, plus a constant intensity. This result, related to the Van Cittert–Zernike theorem, is a commonly used result in interferometric imaging.

By moving the telescopes to various horizontal and vertical separations and measuring the interference intensity, the real part of the Fourier transform of the intensity is inferred. In practice the imaginary part can also be found by varying the baseline optical path length difference by a quarter wavelength. With these samples of the Fourier transform, the inverse Fourier transform computes the image intensity distribution $I(x, y)$. This is how simple computational image formation occurs with interferometric data.

One application of the Van Cittert–Zernike theorem and interferometric imaging is the measurement of the diameter of stars. When observing coherence, the correlations between field points tend to decrease as the separation between the points grows. By observing this decrease of the coherence, the angular size of a star can be measured. This method was pioneered by Albert Michelson in 1890 with the invention of the Michelson Stellar Interferometer. It is needed because the aperture size required to measure the angular diameter of most nearby stars in the Milky Way is larger than the size of most telescope apertures. Stars usually have a circular profile that produces a circularly symmetric pattern of correlations. The coherence between field points as the distance between them is increased does not decrease monotonically, but disappears at certain separations, and has many revivals. The first separation distance at which interference disappears (and the degree of coherence is zero) for a circular profile object is $d_0 = 0.61\lambda/\alpha$, where λ is the wavelength of the light measured from the star, α is the angular diameter (in radians), and d_0 is the first separation from zero at which the interference disappears. In practice, each wavelength produces a null in interference at a different separation, so all but a small bandwidth of the star's light is filtered out before measurement to achieve a suitably strong interference null. For example, for the star Sirius, which has an angular diameter of 0.0068 arc seconds viewed from the Earth, measuring at a wavelength of $\lambda = 500$ nm (blue-green in the visible spectrum), the first separation distance at which no interference will be observed is at 18.5 m. This is larger than the largest available telescope aperture, and therefore an image cannot be directly formed that can resolve the size of Sirius.

Modern examples of stellar interferometry include the European Southern Observatory in Paranal, Chile, the CHARA Array at Mt Wilson, California, USA, and the NASA Jet Propulsion Laboratory Keck Interferometer in Hawaii, USA. The ESO main telescope system is the Very Large Telescope Interferometer (VLTI) consisting of four 8.2 m diameter telescopes as well as several smaller ones that can be moved independently, and will achieve milliarcsecond resolution through the interferometric combination of the telescopes. These telescopes are connected by underground tunnels through which the collected light from each telescope is combined. To measure the phase of the interference, extreme positioning precision is required in the relative delay between the light collected by the telescopes along the baseline, of 50 nm over 120 m, or 1 part in 2400 million, equivalent to 1.6 cm over the circumference of the Earth. Achieving this requires combinations of electronic and mechanical feedback systems with laser stabilized interferometric position measurement. The CHARA array has a very large 330-meter baseline, with 1-meter telescope apertures. The Keck Interferometer consists of two 10 m diameter telescopes with a baseline separation of 85 m. One of the main goals of achieving such high resolution is to be able to observe planets and planet formation around distant stars, to identify possible signs of life in other solar systems.

A similar principle is used for interferometric image formation in radio astronomy. The chief differences are the methods used to collect, detect, and correlate the electromagnetic radiation. Radio astronomical observatories, such as the Very Large Array in New Mexico, USA, and MERLIN at the Jodrell Bank Observatory in Cheshire, UK, use collection dishes and radio antennas to directly measure the electric field, which is not feasible at optical frequencies. These fields are converted to electrical voltages, which can then be directly correlated using electronic mixers, rather than detecting the correlation indirectly through interference. Because the wavelength is much larger, the stability requirements are greatly relaxed. Radio and optical frequency interferometric images are formed in essentially the same way, using the Van Cittert–Zernike theorem.

Interferometric imaging methods have also become of interest for microscopy applications. Aperture synthesis is not needed in microscopy, because the objects of interest are very small and therefore single apertures can be used to achieve diffraction-limited resolution. However, interferometric detection of partially coherent light affords greater flexibility because an image may be computed from the data rather than being directly detected. Typically this is achieved by creating an incoherent hologram of the object, typically through shearing interferometry. For example, by using a Rotational Shearing Interferometer, one can measure an incoherent hologram that can be used to produce a representation of an object with a greatly increased depth-of-field. In addition, direct detection of the partial coherence of sources can allow the methods of computed tomography to be used to produce three-dimensional images of volumes of incoherently emitting objects. Because of the versatility that interferometric detection can provide, interferometric imaging will surely be increasingly used outside of astronomy.

See also: Optical Coherence Tomography and Its Application to Imaging of Skin and Retina

Further Reading

- Baldwin, J.E., Haniff, C.A., 2002. The application of interferometry to optical astronomical imaging. *Philosophical Transactions of the Royal Society of London A* 360, 969–986.
- Beckers, J.M., 1991. Interferometric imaging with the very large telescope. *Journal of Optics* 22, 73–83.
- Born, M., Wolf, E., 1997. *Principles of Optics*. New York: Cambridge University Press.
- Goodman, J.W., 1996. *Introduction to Fourier Optics*. New York: McGraw-Hill Inc.
- Goodman, J.W., 2000. *Statistical Optics*. New York: Wiley-Interscience.
- Mandel, L., Wolf, E., 1995. *Optical Coherence and Quantum Optics*. New York: Cambridge University Press.
- Marks, D.L., Stack, R.A., Brady, D.J., Munson Jr, D.C., Brady, R.B., 1999. Visible cone-beam tomography with a lensless interferometric camera. *Science* 284, 2164–2166.
- Monnier, J.D., 2003. Optical interferometry in astronomy. *Reports on Progress in Physics* 66, 789–857.
- Potluri, P., Fetterman, M.R., Brady, D.J., 2001. High depth-of-field microscopic imaging using an interferometric camera. *Optics Express* 8, 624–630.
- Thompson, A.R., Moran, J.M., Swenson, G.W., 2001. *Interferometry and Synthesis in Radio Astronomy*. New York: Wiley-Interscience.

Multiplex Imaging

A Lacourt, Université de Franche-Comté, Besancon, France

© 2005 Elsevier Ltd. All rights reserved.

Introduction

The ability to project an image onto a screen in a darkened room has been known since antiquity and Leonardo da Vinci (1515) precisely described principles of the Camera oscura. When one exposes a paper sheet directly to sunlight, it is uniformly lighted. However, if one pierces a little hole through the paper sheet then an image of the sun can be displayed on a screen beyond the hole (Fig. 1). Similarly, the shape of the light spot we can see in a dark room, from the rays passing through a small aperture of a closed shutter, is not that of the aperture. Generally, the spot has a circular shape, as an image of the sun. Actually, it is neither a shadow of the hole nor an image of the sun but a combination of both. Light incident on an aperture contains the entire information about the light distribution of the object, but this information has to be 'deblurred', or decoded. The aperture acts as an elementary spatial 2D-filter that selects and preserves the direction and intensity of light rays coming from each point of the scene. The projection on the screen of each pencil of light passing through the hole is a small light spot, the geometry of which is the same as the aperture, and the intensity is proportional to that of the emitting point. The smaller the hole, the more detail the image exhibits, but with less luminosity. Thus, we must adopt a compromise between the resolution and the illumination of the image.

Now, if the mask has several holes, the amount of light reaching the screen, being proportional to the whole aperture area, is extensively increased. On the other hand, as each hole displays an image on the screen, the overlapping of these multiple images gives a new but blurred image. However, the information exists and we should be able to retrieve it if we know the characteristics of the complex aperture. This is the basic concept of multiplex imaging. The crucial question is: what is a good aperture design and a good way to reconstruct images having high resolution and with high illumination levels? The suitable masks with this aim are named synthetic apertures.

Porta (1589) and Kepler (1611) found the perfect solution, by setting up a lens in front of a large hole. Indeed, one can regard a lens as a wide aperture made of continuously juxtaposed micropisms (Fig. 2). The lens obtains such refractive properties that the image displayed by each of these micropisms is moved toward a center. Moreover, as the angle of the micropisms varies continuously, it focuses on a single point, all the rays coming from a given object point. Thus, the blurring is suppressed and the illumination is increased. The way was opened toward the invention of optical instruments and photography and the 'pinhole' has fallen into oblivion for a long time.

Two new problems awake interest in synthetic apertures, around 1960. First, the domain of imaging has extended out of visible light: infrared and ultraviolet light, γ -rays, X-rays, ultrasound waves, and others. Materials suitable for making lenses or mirrors, operating with such waves, are expensive or often do not exist. For example, most of the transparent materials have a refractive index close to 1 for γ -rays and X-rays. However, some are opaque enough to enable the use of masks.

The second challenge appeared with progress in astronomy. In order to explore the universe more, further from Earth and closer to the Big Bang, astrophysicists needed telescopes providing very high luminosity and high resolution, since distant stars, nebulae, and other galaxies are extremely weak and small luminous objects. This exploration can be achieved by increasing the diameter of the primary mirror of telescopes. Indeed, due to diffraction of light by the pupil of the primary mirror, the theoretical limit of resolution of a telescope (i.e., the smaller distance between two points that one can discriminate), is inversely proportional to the diameter of the mirror. In addition, the illumination is proportional to its area. However, the working of large mirrors is a technical prowess that is very expensive and has high risks. Defect of a fraction of micrometer when polishing the surface can have dramatic consequences, e.g., Hubble. Also, such mirrors, being heavy, can bend out of shape under their own weight. Nevertheless, outstandingly large telescopes have been implemented, based on Adaptive optics: the deformations of the mirror are measured in

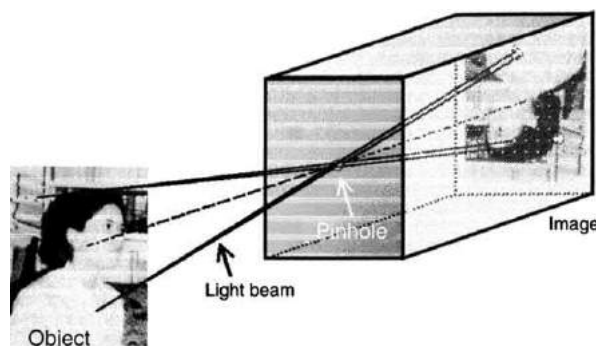


Fig. 1 Principle of darkroom.

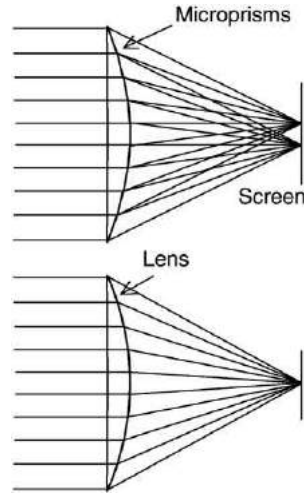


Fig. 2 A lens can be modeled as a juxtaposition of microprisms with continuously variable angles.

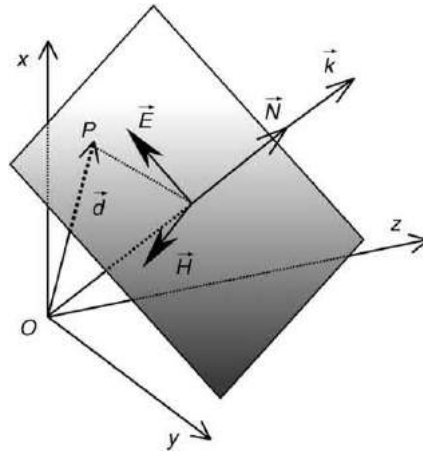


Fig. 3 Electromagnetic wave propagation.

real time by an interference method and compensated by activating a mosaic of jacks (New Technology Telescope – NTT – European program, 1989). Another way consists in replacing the solid-state mirror by a mosaic of 36 mirrors accurately directed (Keck telescope, USA). However, in either case, the mirror diameter does not exceed 8 to 10 meters. Synthetic apertures could be a solution to that difficulty, as earlier demonstrated by Fizeau (1868) and Michelson (1890).

In the first part of this article, any fundamentals are briefly recalled about properties of light and imaging. The second part describes recent experimental methods and results obtained in synthetic aperture imaging. The last part reports the important programs in progress today relating to stellar interferometry.

Background

Electromagnetic Waves Propagation

Light power, which is involved in image formation, is carried by the electric field of an electromagnetic wave (Fig. 3). The electric field $\vec{E}$ has properties of a vector orthogonal to the direction of propagation $\vec{N}$ of the light ray. It is characterized by its amplitude $\vec{E}_0$, the angular time frequency of vibration ω , and the wavevector $\vec{k}$ which specifies the propagating direction of the ray. Let $P(\vec{d})$ be a current point on a plane wavefront, defined by the position vector $\vec{d}$ from the origin of the reference frame. The electric field propagation of a monochromatic wave is expressed as:

$$\vec{E}(\vec{d}, t) = \vec{E}_0 \exp[i(\vec{k} \cdot \vec{d} - \omega t)] \quad (1)$$

The light power carried by the electromagnetic field is proportional to $I = \|\vec{E}_0\|^2 = \vec{E} \cdot \vec{E}^*$ where $\vec{E}^*$ is the complex conjugate of $\vec{E}$. The time frequency $\nu = \omega/2\pi$, or the wavelength $\lambda = c/\nu$, where c is the electromagnetic wave celerity, characterize the hue of the

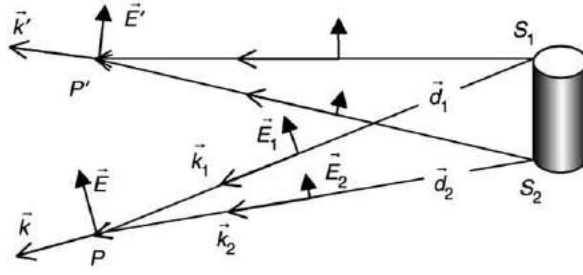


Fig. 4 Spatially coherent light.

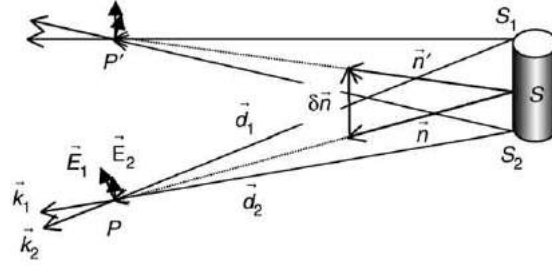


Fig. 5 Spatially incoherent light.

light. The wavevector $\vec{k} = \vec{N}\omega/c$ where $||\vec{N}|| = 1$ is an angular space frequency vector, as $\vec{n} = \vec{N}/\lambda = \vec{N}v/c$ is a space frequency vector. We have $\vec{k} = 2\pi\vec{n}$. As λ is a constant for a given monochromatic light, $\vec{n}$ defines the direction of propagation of a ray.

Spatially Coherent Light

Consider a very small source stated as the point source S_1 , emitting uniform lightwaves in any direction (Fig. 4). The wave reaching a farfield point, P , from S_1 , propagates along a direction defined by the wavevector $\vec{k}_1$. The vectors $\vec{k}_1$ and $\vec{d}_1 = \vec{S}_1P$ are collinear. Another point source, S_2 is characterized from P by the wavevector $\vec{k}_2$ such that $\vec{k}_2$ and $\vec{d}_2 = \vec{S}_2P$ are collinear. When the distances d_1 and d_2 are close to each other, the amplitudes $\vec{E}_{01}$ and $\vec{E}_{02}$ of the waves at P are proportional to those of the corresponding emitting sources, S_1 and S_2 . According to the principle of superposition, the resulting field at the point P , $\vec{E} = \vec{E}_1 + \vec{E}_2$ and the instant wavevector $\vec{k}$ are unique vectors.

If S_1 and S_2 are two synchronous sources, the amplitude $\vec{E}_0$ of $\vec{E}$ and the wavevector $\vec{k}$ are constant versus time at any point P . This is still the case when an extended object is made of an arrangement of synchronous points. That situation defines a spatially coherent object. Thus, one cannot recover multiple data about the object's distribution, from the field framework at a single point P , since the relationship between the two is multi-unequivocal. Nevertheless, due to diffraction, one can retrieve them by knowing the complex amplitude distribution of light on a set of wavevectors available over an extensive area of a screen. Let $\vec{E}_0(\vec{r})$ be the amplitude of the electric field at the object point $M(\vec{r})$ and let $\vec{E}_0(\vec{n})$ be the one diffracted by the object to farfield toward the direction corresponding to the space frequency $\vec{n}$. Fourier transforms give the connection between $\vec{E}_0(\vec{r})$ and $\vec{E}_0(\vec{n})$:

$$\begin{cases} \vec{E}_0(\vec{n}) = \int \vec{E}_0(\vec{r}) \exp[i2\pi\vec{n} \cdot \vec{r}] d\vec{r}, \\ \vec{E}_0(\vec{r}) = \int \vec{E}_0(\vec{n}) \exp[-i2\pi\vec{n} \cdot \vec{r}] d\vec{n}. \end{cases} \quad (2)$$

This is the situation encountered when an object is illuminated by a laser or by a point source such as a single star.

Spatially Incoherent Light

When the point sources S_1 and S_2 are independent – spatially incoherent – both wavevectors $\vec{k}_1$ and $\vec{k}_2$ coexist at the point P , each characterizing the angular position of S_1 and S_2 (Fig. 5). If S_1 and S_2 are part of an extended intensity distribution $I_0(\vec{r})$ fans of wavevectors and fields are mixed at P , each coming from an independent point source S . Together, they carry the whole data on the object distribution. Those data are partially uncorrelated at any two points P and P' on a screen, related to the different optical paths from any point S to P and P' and to the finite time of coherence of light. The correlation between data available to farfield in two directions $\vec{n}$ and $\vec{n}'$ from the object, is given by the mutual intensity function, $\Gamma(\delta\vec{n}) = \int \vec{E}(\vec{n}) \cdot \vec{E}^*(\vec{n} + \delta\vec{n}) d\vec{n}$ where $\delta\vec{n} = \vec{n}' - \vec{n}$. Fourier transforms give the relation between $I_0(\vec{r})$ and $\Gamma(\delta\vec{n})$:

$$\left\{ \begin{array}{l} \Gamma(\delta\vec{n}) = \int I(\vec{r}) \exp[i2\pi\delta\vec{n} \cdot \vec{r}] d\vec{r}, \\ I(\vec{r}) = \int \Gamma(\delta\vec{n}) \exp[-i2\pi\delta\vec{n} \cdot \vec{r}] d\delta\vec{n} \end{array} \right\} \quad (3)$$

Therefore, by drawing a map of $\Gamma(\delta\vec{n})$ one can recover the object distribution $I_0(\vec{r})$ by a Fourier transformation. Most of time, one uses the spatial degree of coherence function, $\gamma(\delta\vec{n}) = \Gamma(\delta\vec{n})/\Gamma(\vec{0})$. When the object illuminates a screen, at a large distance d , the degree of coherence of light between two points P and P' on the screen is $\gamma(\delta\vec{p}/\lambda d)$ with $\delta\vec{p} = \vec{PP'}$. The 2D extension of $\gamma(\delta\vec{p}/\lambda d)$ defines the area of coherence of light on the screen. The width of the area of coherence, $\Delta\rho$, and that of the object distribution, Δr , are inversely proportional:

$$\Delta\rho \cdot \Delta r \sim \lambda d \quad (4)$$

If a mask perforated by two pinholes, separated by a distance $\delta\vec{p}$ replaces the screen, one can observe Young fringes of interference beyond the mask, the visibility of them being precisely $\gamma(\delta\vec{p}/\lambda d)$. Therefore, one can build a map of $\gamma(\delta\vec{p}/\lambda d)$ versus $\delta\vec{p}$ from a set of 2D measurements of visibilities of fringes, then reconstruct the object distribution by a Fourier transformation.

Notice that the entire data of the object distribution is available from a single coherence area. That is, a pinhole covering exactly one area of coherence would select the entire available data from the object and would display to farfield the most accurate image possible. Generally, such a pinhole is so small that practically no light passes through it. However, we can regard each area of coherence as an independent channel of information. Then an imaging device based on a juxtaposition of such redundant channels can be considered as a multiplex imaging device.

Self-luminous objects such as extended or double stars, and artificial thermic sources are spatially incoherent sources.

Multiplex Imaging

Methods of multiplex – or lensless – imaging are performed either with coherent or incoherent light. The very different properties of the two types of light leads to completely different methods, even if all are related to holography. Holography with coherent light is an interference method for local recording of both the direction and modulus of the wavevectors and the amplitude of the field contained in many juxtaposed grains of speckle, scattered by the object. Each grain is an independent channel of imaging.

Synthetic Apertures with Spatially Incoherent Objects

As interferences are impossible by this method with incoherent light, we must improve other methods of imaging. Synthetic aperture is a concept consisting of using a redundant multichannel system to increase the luminosity of images without loss of resolution. The basic principle is that of a multiple holes darkroom. The processing is in two steps: recording and reading the image. The recording step consists in illuminating a photosensitive receptor with the object through a mask (Fig. 6). Generally, the receptor is a photographic plate. Obviously, the light distribution on the plate is blurred. Indeed, any elementary transparencies on the mask project on the receptor inverted images of the object located the shadow of the mask. Such a distribution is expressed by a convolution. Let $O(x,y)$ be the conical projection of the object on the receptor and $Z(x,y)$ that of the mask. The shadows distribution on the receptor is (the variable being normalized):

$$I'(x',y') = Z(x,y) * O(x,y) = \int Z(x,y) O(x' - x, y' - y) dx dy \quad (5)$$

The reading step is a decorrelation processing from the recorded intensity (Fig. 7). It consists in achieving a new correlation of I' with an arbitrary function T :

$$I''(x,y) = T(x,y) * [Z(x,y) * O(x,y)] = [T(x,y) * Z(x,y)] * O(x,y) \quad (6)$$

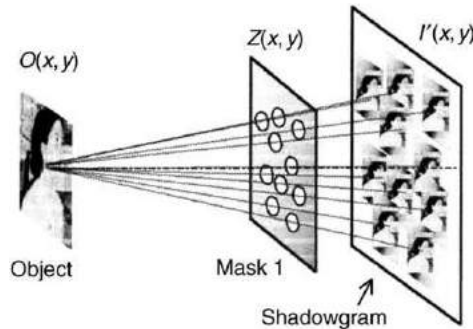


Fig. 6 Principle of synthetic aperture: recording.

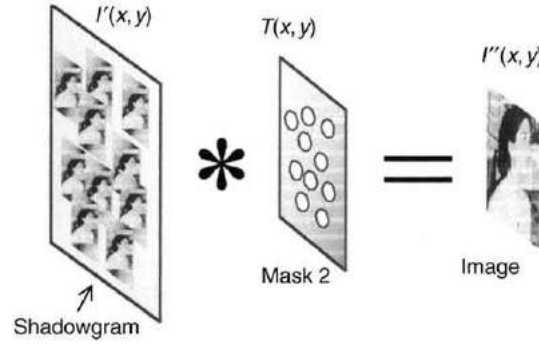


Fig. 7 Synthetic aperture: reading.

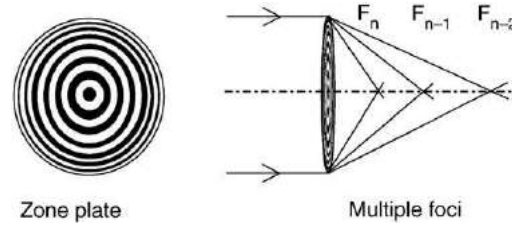


Fig. 8 Properties of FZP.

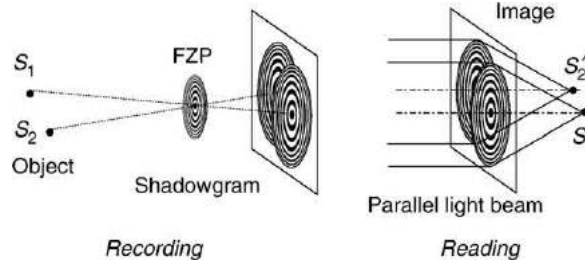


Fig. 9 Recording and analogical reading with a FZP.

$R[x, y] = [T(x, y) * Z(x, y)]$ is the system point spread function (SPSF) of the processing. It expresses the image distribution of a single-point. The perfect image is obtained when $R(x, y)$ is a 2D Dirac – or delta– distribution, $\delta(x, y)$:

$$R(x, y) = \delta(x, y), \quad \text{as} \quad \int \delta(x, y) O(x' - x, y' - y) dx dy = O(x', y') \quad (7)$$

Fresnel zone plates

First Mertz and Young used Fresnel zone plate (FZP) as a synthetic aperture for X-ray imaging purposes. FZPs are made of alternate transparent and opaque concentric crowns of equal area. The radius of the n th circle is $r_1 \sqrt{n}$, r_1 being the smallest. When illuminated by a cylindrical light beam, the FZP is a multiple foci diffractive lens (Fig. 8).

The reading step is either analogical or numerical. Through the analogical method, we consider the recorded plate as a pseudohologram (Fig. 9). When the 'shadowgram' is lighted with a parallel beam of laser light, the FZPs recorded around each point concentrates the light on its foci. Then an image of the object is reconstructed on the focal planes; a method that is simple and elegant. Nevertheless, the processing is not linear since we record an intensity distribution at the recording step, while the reading step by diffraction involves amplitude. Moreover, the S/N ratio decreases as the amount of data increases, recorded on the photographic plate because of its finite dynamic.

Mertz and Young had the idea of using a second FZP to perform the deconvolution processing by optical methods. In this case, the SPSF, $R(x, y)$, is the autocorrelation function of the FZP transparency (Fig. 10). It presents a narrow central peak, emerging from a pyramidal ground (Fig. 11). The optical setup is the same as the recording step. The FZP scale factor only must be adjusted to compensate for the magnification ratio generated by the conical projection at the recording step.

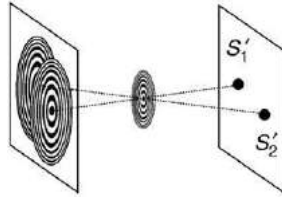


Fig. 10 Optical decorrelation.

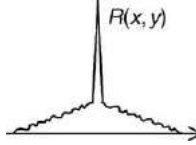


Fig. 11 System point spread function with a Fresnel zone plate.

In either digital or analogical deconvolution, the image reconstructed is accurate, with good sharpness, if the number of zones is sufficient. Ceglio *et al.* obtained relevant images of 8 μm targets with FZP having 100 zones. Nevertheless, the multiple foci of FZP induces decreasing of the S/N ratio which is an important limitation. Besides, we pointed out above that the larger the object, the more the decrease in the S/N ratio.

Random and uniformly redundant arrays

Dicke and Ables substituted a random array to the FZP. The mask is made of randomly distributed holes and the expected advantage is doubled. First, one can reduce the transparent area of the mask by less than 50% in order to increase the S/N ratio when the object is wide. Second, the autocorrelation function of such an array is theoretically a Dirac distribution when its width is infinite. Thus, one can hope to narrow the system point response. However, when the width is finite, the pyramidal ground of the point response mentioned above is present and the advantage is not clearly demonstrated. In order to correct this effect, Chris Brown used a 'mismatched' function, for the postprocessing deconvolution. This function is calculated, first by achieving the complement of the recording mask function, then subtracting from it. Thus, the 1 transparencies result in $+1$ while the 0 transparencies are changed by -1 . Brown shows images digitally reconstructed of X-ray sources, 30 sec. of arc wide. Fenimore and Cannon conceived a uniformly redundant array. The center random motif of the mask is half replicated on each side with only the center area being used for the recording step. The deconvolution is calculated by taking into account the whole area in a mismatched way. The result is an exact delta function. The authors show images of 200 μm -wide microspheres with a similar resolution but a signal 7100 times stronger than with a comparable pinhole camera.

Theta Rotating Interferometer

The theta shearing interferometer enables us to record complete data from the Fourier transform of a spatially incoherent object (Fig. 12). The source can be any natural one, but monochromatic. This is easily practicable using a good filter or monochromator.

The device is based on a Michelson interferometer, the plane mirrors being replaced by two roof prisms. One of the roof prisms is fixed so the other can rotate around the axis of the interferometer. As a result, we have two images of the object. Let θ be the angle between the two edges of the prisms. The angle between the images is 2θ . Since the original object is noncoherent, and two homologous points being mutually coherent, the light they emit interferes, resulting in fringes.

The intensity, space frequency, and direction of the fringes characterize the intensity and the position of the point source. One shows that the farfield visibility distribution of fringes $V(\delta\vec{n})$ and the object distribution $O(\vec{r})$ are related by Fourier transforms:

$$\begin{cases} V(\delta\vec{n}) = \int O(\vec{r}) \exp \left[i2\pi \frac{\delta\vec{n} \cdot \vec{r}}{2\sin\theta} \right] \frac{d\vec{r}}{2\sin\theta}, \\ O(\vec{r}) = \int V(\delta\vec{n}) \exp \left[-i2\pi \frac{\delta\vec{n} \cdot \vec{r}}{2\sin\theta} \right] d\delta\vec{n} \end{cases} \quad (8)$$

Then, one can see that the visibility of the interference pattern is an image of the degree of coherence $\gamma(\delta\vec{n})$ with a scale factor $1/2 \sin \theta$.

Such a pattern recorded on a photographic plate is a Fourier hologram of the noncoherent object. An image can be reconstructed by diffraction at infinity or near a focus (see Fig. 13, examples performed in 1972 by the author). The technique was successfully applied to the determination of the modulation transfer function of optical systems. C Roddier and F Roddier applied

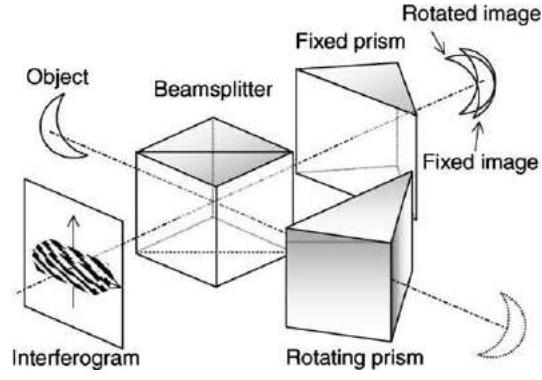


Fig. 12 Theta shearing interferometer.

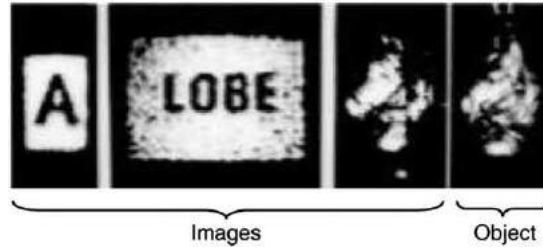


Fig. 13 Incoherent light holograms.

the method to astronomy. For instance, they examined Alpha Orionis and discovered the evolution of its dust envelope and the possible existence of a companion.

High Resolution Imaging in Astronomy

In astronomy, the limit of resolution depends chiefly on two factors. The first is a technical one. Adaptive optics and very large telescopes (VLT) brought spectacular progress to clarity and resolution of telescopes. The 8.2 m VLT of Kueyen set up in 1989, shows images easily resolved of HIC 69495 Centauri, separated by 0.12 sec. of arc. Nevertheless, working of large mirrors for telescopes reaches its limits. Therefore, their clarity remains too weak. On the other hand, atmospheric turbulences make the refractive index of air unstable and the position of the images in the focal plane of telescopes unsettled. Thus, the point response is a granular spot named 'speckle'. Short exposure times relieve that inconvenience, but they prohibit the observation of weak stars. In 1970, Labeyrie proposed a method of speckle interferometry consisting of a statistical analysis of a set of speckles. The result is a display of a diffraction-limited image. He resolved many tens of single and double stars. Only Hubble, above the atmosphere, is able to do it by direct imaging. However, the primary mirror of Hubble being only 2.40 m wide, has clarity much lower. Stellar interferometry is a very promising alternative solution.

Principles of Stellar Interferometry

Stellar interferometry is based on Eq. (3). The light emitted by double or wide stars is noncoherent, so that one can access an image of them through 2D measurements of the degree of coherence of light. We showed above that one could achieve this by using interference devices such as Young holes. Eq. (3) shows that if we measure the visibility of Young fringes at increasing distances $\delta\rho$ between the pinholes, the distance $\Delta\rho$ from which the visibility is zero gives the angular width of the star, $\Delta\alpha$, under the formula:

$$\Delta\rho \cdot \Delta\alpha \sim \lambda \quad (9)$$

where $\Delta\alpha = \Delta r/d$. In the case of a double star, the visibility varies periodically versus $\delta\rho$ when the pinholes are aligned along the axis of two stars. Let $\Delta\alpha$ be the angular distance between the stars: the smallest distance $\Delta\rho$ for which the visibility is zero is given by:

$$\Delta\rho \cdot \Delta\alpha = \lambda/2 \quad (10)$$

The larger the uttermost distance between the holes, the more the limit of resolution of the processing is potentially sharp.

In order to exploit this idea, Stephan placed and rotated a mask with two holes in front of the objective of a refractive telescope. He established that the diameter of the stars he observed was much less than 0.16 sec. of arc. Then, Michelson built up, on top of a

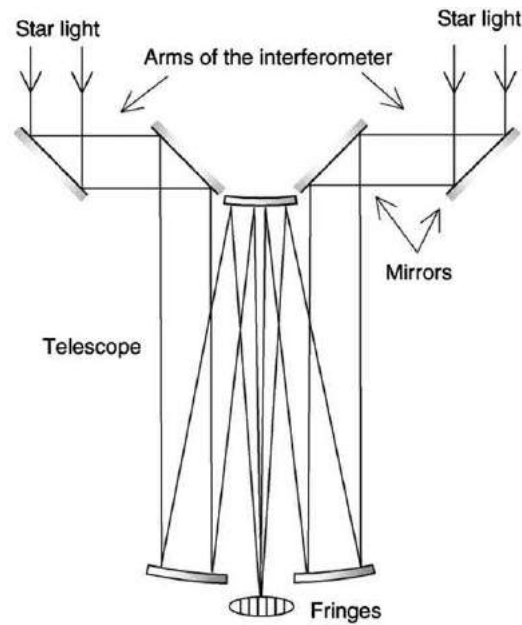


Fig. 14 Stellar interferometer of Michelson.

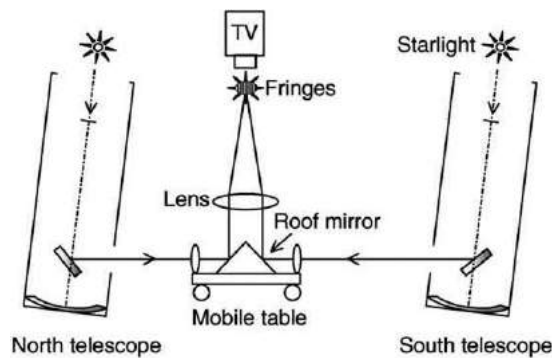


Fig. 15 The two telescopes interferometer of Labeyrie.

30 cm refractive telescope, a stellar interferometer consisting of two arms catching the light at 7 m of distance from each other (Fig. 14). He measured correctly the diameter of the satellites of Jupiter. Hale carried the distance of the base up to 15 m in 1920.

The Two Telescopes Interferometer of Labeyrie

In 1975, Labeyrie improved the idea by setting up two separate 25 cm diameter Cassegrain-coudé telescopes on a 12 m baseline, oriented north-south (Fig. 15). The setting and mechanical stability of the device are extremely sensitive. The tolerances do not exceed any μm . A mobile table near the common focal plane of the two telescopes compensates for the optical path differences due to stars tracking. In spite of the acute difficulties, he succeeded in observing fringes on Vega. He obtained a resolution of 0.01 sec. of arc versus 0.5 sec. with a single 25 cm telescope. Later, one telescope was rendered mobile bringing the base width variable between 4 and 67 m. The resolution reaches up to 0.00015 sec. of arc. Note that the rotation of Earth also enables east-west analysis with resolution of 0.005 sec., and today, the 25 cm telescopes are replaced by 1.5 m ones.

The Very Large Telescope Interferometer (VLTI)

The two telescopes interferometer of Labeyrie is a first step toward a very ambitious project: the telescope at synthetic aperture and very large base. Proofs of reliability of such a principle was already given by the very large array (VLA) experimented in radio astronomy. Labeyrie with Lena, and the team of the ESO (Ecole Supérieure d', Paris), imagined a very large telescope interferometer (VLTI) made of 6 telescopes of 2 m, in a 2D structure on a base of 100 m. The European VLTI was built at Cerro Paranal. It is made of four 8.2 m telescopes corrected in real time by adaptive optics (NTT technology) and three 1.8 m telescopes. The

fourth 8.2 m telescope was put in service in 2001. Some days, each of the 8.2 m telescopes used separately gives a resolution close to that of Hubble. However, the clarity is much better because of the larger primary mirror diameter (8.2 m against 2.4 m). In May 2003, the team of J-G Cuby observed a galaxy the farthest away ever seen, only 900 millions years old since the Big Bang. Obviously, a drastic gain of resolution (by a factor 10) is expected from the interference configuration.

Further Reading

Labeyrie, A., 1976. High resolution techniques in optical astronomy. In: Wolf, E. (Ed.), *Progress in Optics*, vol. XIV. New York: North-Holland, pp. 49–85.
Mertz, L., 1965. *Transformations in Optics*. Amsterdam, Oxford: John Wiley & Sons, Inc.

Photon Density Wave Imaging

V Toronov, University of Illinois at Urbana-Champaign, Urbana, IL, USA

© 2005 Elsevier Ltd. All rights reserved.

Nomenclature

μ_a [cm^{-1}] Absorption coefficient
Alternating part of PDW, alternating current (ac)
 g_1 Average cosine of the scattering angle
 ω [s^{-1}] Circular modulation frequency
Computed tomography (CT)
 $\{\Delta\mu_a(\mathbf{r}), \Delta\mu'_s(\mathbf{r})\} = \{\mu_a(\mathbf{r}), \mu'_s(\mathbf{r})\} - \{\mu_a(\mathbf{r}), \mu'_s(\mathbf{r})\}_e$
 $[\text{cm}^{-1}]$ Difference between the estimate and actual optical properties
 $D = v/(3\mu_a + 3\mu'_s)$ [cm^2/s] Diffusion coefficient
Diffusion equation (DE)
 $S(\mathbf{r}, \hat{\Omega}, t)$ [$\text{W cm}^{-3} \text{steradian}^{-1}$] Distribution of light source power
 ∇ [cm^{-1}] Divergence operator
 $\{\mu_a(\mathbf{r}), \mu'_s(\mathbf{r})\}_e$ [cm^{-1}] Estimate of optical properties
 F General nonlinear operator
 J Jacobian
 Δ [cm^{-2}] Laplace operator
Magnetic resonance imaging (MRI)
 $\{\mu_a^m, \mu_s^m\}$ [cm^{-1}] Measured bulk optical properties (absorption and reduced scattering coefficients)
 ν [s^{-1}] Modulation frequency
Non-alternating part of PDW, direct current (dc)

$\{\mu_a(\mathbf{r}), \mu'_s(\mathbf{r})\}$ [cm^{-1}] Optical properties of the medium (absorption and reduced scattering coefficients) as functions of spatial coordinate
 $\phi(\mathbf{r})$ Phase as a function of source–detector distance
 $f(\hat{\Omega}, \hat{\Omega}')$ Phase function
 $U(\mathbf{r}, t)$ [J cm^{-3}] Photon density
Photon density wave imaging (PDWI)
Photon density wave (PDW)
 l [cm] Photon mean free path
 $L(\mathbf{r}, \hat{\Omega}, t)$ [$\text{W cm}^{-2} \text{steradian}^{-1}$] Radiance
 $\Phi(\mathbf{r}, t)$ [W cm^{-2}] Radiation fluence
 $J(\mathbf{r}, t)$ [W cm^{-2}] Radiation flux
 μ'_s [cm^{-1}] Reduced scattering coefficient
 μ_s [cm^{-1}] Scattering coefficient
 G Set of measured values (ac, dc, and phase)
 G_e Set of predicted values (ac, dc, and phase)
 $Q(\mathbf{r}, t)$ [$\text{J cm}^{-3} \text{s}^{-1}$] Source term
 r [cm] Source–detector distance
 $\mathbf{r}$ [cm] Spatial coordinate
 v [cm/s] Speed of light in medium
 $Y_{nm}(\hat{\Omega})$ Spherical harmonics
Transport equation (TE)
 $h(\mathbf{r})$ Weight function
 $x = \omega/(\nu\mu_a)$

Photon density wave imaging (PDWI) employs diffused light, usually in the near-infrared band, for imaging inhomogeneous structures in a highly scattering medium. The other names for PDWI are diffuse optical tomography and photon migration imaging. The term photon density wave (PDW) refers to the wave of light density spreading in the highly scattering medium from the source whose power changes in time. Therefore, rigorously speaking, photon density wave imaging uses intensity-modulated light sources, such as pulsed lasers, or sources whose power is harmonically modulated. However, many applications employ steady-state light sources. Most applications of PDWI are in the field of imaging of biological tissues, including medicine and cognitive neuroscience.

As in the other imaging methods, such as x-ray, computed tomography (CT) or ultrasound imaging, the problem of PDWI consists in the reconstruction of the internal structure of the medium from the measurements made outside the medium (on its surface). In the case of PDWI the internal structure of the medium is represented by the spatial distribution of its optical properties. The reconstruction procedure is essentially an iterative mathematical algorithm, which includes the modeling of the light propagation inside the medium assuming a certain distribution of the optical properties, and the correction of this distribution to match the model prediction with the actual measurement. The modeling part of the imaging problem is usually referred to as the forward problem, and the correction procedure is known as an inverse problem.

Forward Problem: Photon Density Waves

Unlike optically transparent media, in a highly scattering medium the coherence of light from the coherent source very rapidly decays with distance. This happens due to the interference of many randomly and multiply scattered light waves at every point of the medium that is distanced from the source much farther than the photon mean free path l , which is the mean distance at which a light wave undergoes a single scattering event. For biological tissues, where light scattering occurs on cell membranes and other heterogeneities of the index of refraction, l is typically about 1 mm. Although high scattering eliminates the coherence of the electromagnetic waves at the short time intervals corresponding to oscillation at optical frequencies ($\sim 10^{-15}$ s), in the case of an intensity-modulated light source, the macroscopic wave of the density of light energy, called 'photon density wave', forms in the highly scattering medium. The properties of PDW depend on the type of

the source modulation and on the local optical properties of the medium, which are the scattering coefficient μ_s [cm^{-1}] and the absorption coefficient μ_a [cm^{-1}].

Because of the complexity and incoherence of the macroscopic electromagnetic field in highly scattering media, i.e. media with $\mu_a \ll \mu_s$, the fundamental Maxwell equations of electrodynamics do not provide a convenient approach to the solution of the problem of light propagation in such media at distances much larger than l . Therefore, statistical approaches have been developed, such as the photon transport model and its diffusion approximation. In the photon transport model the light is characterized by the radiance $L(\mathbf{r}, \hat{\Omega}, t)$ [$\text{W cm}^{-2} \text{steradian}^{-1}$] (where the steradian is the unit solid angle), which is the power of radiation near the point $\mathbf{r}$ propagating within a small solid angle around the direction $\hat{\Omega}$ at time t . The photon transport model accounts for the possible nonisotropy of the radiation flux. For the given spatio-temporal distribution of the optical properties, μ_a and μ_s , and for the given spatial and angular distribution of light source power $S(\mathbf{r}, \hat{\Omega}, t)$ [$\text{W cm}^{-3} \text{steradian}^{-1}$] the radiance $L(\mathbf{r}, \hat{\Omega}, t)$ can be found from the linear radiation transfer equation, also called the photon transport equation, or simply the transport equation (TE):

$$\frac{1}{v} \frac{\partial L(\mathbf{r}, \hat{\Omega}, t)}{\partial t} + \nabla \cdot L(\mathbf{r}, \hat{\Omega}, t) \hat{\Omega} + (\mu_s + \mu_a) L(\mathbf{r}, \hat{\Omega}, t) = \mu_s \int L(\mathbf{r}, \hat{\Omega}', t) f(\hat{\Omega}, \hat{\Omega}') d\hat{\Omega}' + S(\mathbf{r}, \hat{\Omega}, t) \quad (1)$$

where $f(\hat{\Omega}, \hat{\Omega}')$ is the phase function, which represents the portion of light energy scattered from the direction $\hat{\Omega}$ to the direction $\hat{\Omega}'$ and v is the speed of light in the medium. TE is equivalent to the Boltzmann transport equation describing the transport of particles undergoing scattering.

Eq. (1) is difficult to solve even numerically. Therefore, a hierarchy of approximations to the PTM was developed. This hierarchy, known as the P_N approximation, is based on the expansion of the radiance in a series of spherical harmonics $Y_{nm}(\hat{\Omega})$ truncated at $n=N$. In the P_1 approximation the coefficients of the series are proportional to the radiation fluence

$$\Phi(\mathbf{r}, t) = \int L(\mathbf{r}, \hat{\Omega}, t) d\hat{\Omega} \quad (2)$$

and the flux

$$J(\mathbf{r}, t) = \int L(\mathbf{r}, \hat{\Omega}, t) \hat{\Omega} d\hat{\Omega} \quad (3)$$

both having units of [W cm^{-2}]. By substituting the P_1 series for $L(\mathbf{r}, \hat{\Omega}, t)$ and $S(\mathbf{r}, \hat{\Omega}, t)$ into Eq. (1) and assuming isotropic light sources and sufficiently slow change of source intensity, one can obtain the equation that includes only $\Phi(\mathbf{r}, t)$ as the unknown. Usually, this equation

$$v\mu_a U(\mathbf{r}, t) + \frac{\partial U(\mathbf{r}, t)}{\partial t} - D\Delta U(\mathbf{r}, t) = Q(\mathbf{r}, t) \quad (4)$$

is written in terms of the value $U(\mathbf{r}, t) = \frac{1}{v} \Phi(\mathbf{r}, t)$ [J cm^{-3}], which has units of energy density, and is called the photon density. In Eq. (4) $D = v/(3\mu_a + 3\mu'_s)$ [cm^2/s] is the light diffusion coefficient, and $\mu'_s = \mu_s(1 - g_1)$ [cm^{-1}] is the reduced scattering coefficient; g_1 is the average cosine of the scattering angle. The source term $Q(\mathbf{r}, t)$ [$\text{J cm}^{-3} \text{s}^{-1}$] in the right-hand side of the Eq. (4) describes the spatial distribution of light sources (in the isotropic approximation, i.e., neglecting the dipole and higher momenta of $S(\mathbf{r}, \hat{\Omega}, t)$) and the temporal modulation of source intensity.

Eq. (4) is mathematically equivalent to the equation describing the density of the medium consisting of the particles undergoing linear diffusion. Therefore, it is called the diffusion equation (DE), and the P_1 approximation is also known as the diffusion approximation. The diffusion approximation is valid under the following three conditions: (a) the light sources are isotropic or the distance between the source and the detector is much larger than l ; (b) characteristic source modulation frequencies ν are low enough that $\nu\mu'_s/(2\pi\nu) \gg 1$; and (c) $\mu_a \ll \mu'_s$. In the biomedical applications of PDWI the condition (a) is fulfilled when the source–detector distance is larger than 1 cm. The condition (b) is valid when the characteristic modulation frequencies of the light source are less than 1 GHz. The condition (c) is fulfilled in most biological tissues, for which the scattering is highly isotropic, i.e. $g_1 \ll 1$. Thus, the forward problem of PDWI, i.e., the modeling of light propagation, can be solved using the diffusion approximation in a large variety of situations.

PDWs described by Eq. (4) are the scalar damped traveling waves, characterized by the coherent changes of light energy (photon density) in different locations on time intervals from microseconds to nanoseconds. Usually pulsed or harmonically modulated light sources are used to generate PDWs. The properties of the PDWs from a harmonic source are fundamental, because the PDW from the source of any type of modulation can be considered as a superposition of harmonic waves from sources having different frequencies using the Fourier transform.

In order to get an idea of the basic properties of PDW, one can consider the analytical solution of (4) for the case of a harmonically modulated point-like source placed in an infinite homogeneous scattering medium. This solution is

$$U(r, t) = U_{dc}(r) + U_{ac}(r) \exp[-i\omega t + i\phi(r)] \quad (5)$$

where U_{ac} , U_{dc} , and ϕ are the ac, dc and phase, respectively, of the wave measured at the distance r from the source, and $\omega = 2\pi\nu$. U_{ac} , U_{dc} , and ϕ depend on the medium optical properties, source intensity Q_0 , modulation amplitude A_0 , and initial phase ψ_0 as

$$U_{ac}(r) = \frac{Q_0 A_0}{4\pi D} \frac{\exp\left[-r\left(\frac{\nu\mu_a}{2D}\right)^{1/2} \left((1+x^2)^{1/2} + 1\right)^{1/2}\right]}{r} \quad (6a)$$

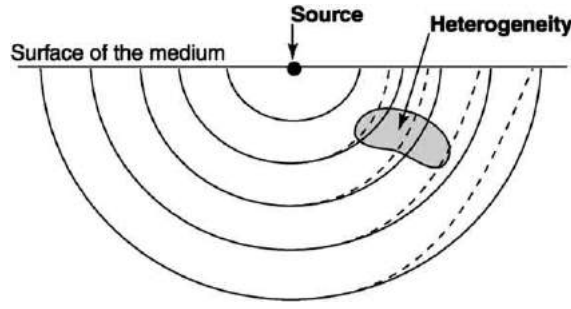


Fig. 1 Equiphase surfaces of the PDW induced by a point source in a semi-infinite homogeneous medium (solid lines) and in a medium including heterogeneity (dashed lines).

$$U_{dc}(r) = \frac{Q_0}{4\pi D} \frac{\exp\left[-r\left(\frac{\nu\mu_a}{D}\right)^{1/2}\right]}{r} \quad (6b)$$

$$\phi(r) = r\left(\frac{\nu\mu_a}{2D}\right)^{1/2} \sqrt{(1+x^2)^{1/2} - 1} \quad (6c)$$

where $x = \omega/(\nu\mu_a)$. Eqs. (5)–(6) describe a spherical wave whose wavelength at $\mu_a = 0.1 \text{ cm}^{-1}$ and $\mu'_s = 8 \text{ cm}^{-1}$ (typical for biological tissues) and at modulation frequency $\nu = 100 \text{ MHz}$, is close to 35 cm, so that the phase of the wave changes approximately by 10° per cm. In the same time, the wave decays exponentially at a distance of about 3 cm, so that PDWs are strongly damped. A similar analytical solution in the form of a spherical wave can be obtained for the case of a semi-infinite homogeneous medium and point source placed on the surface of the medium (see Fig. 1). These analytical solutions are employed to measure μ_a and μ'_s of the homogeneous medium. The method is based on fitting measured ac, dc, and phase of the PDW as functions of source–detector distance by the analytical solution.

Inhomogeneities in the optical properties of a turbid medium cause distortion of the PDW from sphericity (see Fig. 1). PDWs in homogeneous media of limited size or with strong surface curvature also usually have complex structure. For a limited number of configurations one can obtain solutions to the DE using analytical methods of mathematical physics, such as the method of Green's functions. However, in practice the complexity of the inhomogeneities and surfaces usually requires numerical solution of the DE. This can be achieved using numerical methods for the solution of partial differential equations, such as the finite element method and the finite difference method. The finite difference method can also be used to solve the TE. This may be necessary in situations when the DE is not valid, for example to accurately account for the influence of the cerebrospinal fluid (which has relatively low scattering) on light transport in the human skull. However, the solution of the TE requires formidable computational resources, which increase proportionally to the volume of the medium. Therefore, hybrid DE-TE methods were developed, in which DE is used to describe light propagation in most of the medium except small low-scattering areas described by the TE.

One should note that although the term 'photon' suggests quantum properties of light, no consistent application of quantum theory to problems related to PDWI has been developed so far. Also, it should be noted that the notion of a photon as a particle of light having certain energy, momentum, and localization at the same time is not completely consistent with quantum electrodynamics. However, since the theoretical models of light transport in a scattering medium (such as TE and DE) are mathematically equivalent to the models describing systems of randomly moving particles, the notion of a photon as a 'particle of light' may be accepted as adequate and is widely used in the limits of PDWI.

Inverse Problem and Imaging Algorithm

Mathematically the forward problem can be expressed in the form of a general nonlinear operator F :

$$G = F[\{\mu_a(\mathbf{r}), \mu'_s(\mathbf{r})\}] \quad (7)$$

where G is the set of measured values (ac, dc, and phase), $\mathbf{r}$ is the coordinate vector of a medium element (voxel), and $\{\mu_a(\mathbf{r}), \mu'_s(\mathbf{r})\}$ denotes the set of tissue optical properties corresponding to whole collection of voxels. The problem of finding the spatial distribution of medium optical properties can be formalized as

$$\{\mu_a(\mathbf{r}), \mu'_s(\mathbf{r})\} = F^{-1}[G] \quad (8)$$

Neither the transport model nor the diffusion model allow inversion in a closed analytical form for a general boundary configuration and spatial distribution of optical properties. The basic reason for the complexity of the inverse problem is in the complexity of the light paths in the highly scattering medium. This situation is opposite to that in CT, where the x-ray radiation propagates through the medium along straight lines almost without scattering, which significantly simplifies the forward problem and allows its analytical inversion. From the mathematical point of view, the complexity of the inverse problem of PDWI is due to its nonlinearity and ill-posedness. Only approximate methods of inversion have been developed so far.

The simplest method is called backprojection by analogy with the method of the same name used in CT. The method attempts to reconstruct the actual spatial distribution of the optical properties μ_a and/or μ'_s from the 'measured' bulk values μ_a^m and μ'_s^m obtained at a variety of locations assuming homogeneity of the medium. The measured values are considered as a result of the volumetric 'projection'

$$\{\mu_a^m(\mathbf{r}), \mu'_s^m(\mathbf{r})\} = \int \{\mu_a(\mathbf{r}'), \mu'_s(\mathbf{r}')\} h(\mathbf{r} - \mathbf{r}') d^3\mathbf{r}' \quad (9)$$

of the actual distribution, where $\mathbf{r}$ and $\mathbf{r}'$ are the locations of the measurement and reconstruction, respectively. A weight function $h(\mathbf{r})$ accounts for the contribution of the particular region to a measured value and should be derived from the forward model. Different versions of the backprojection method use different weight functions $h(\mathbf{r})$. In the backprojection method the values of μ_a and/or μ'_s are obtained by inverting the convolution (9). In practice the convolution integral is usually estimated by a sum over a limited number of voxels for a limited number of measurements. Therefore, the inversion of the convolution reduces to the solution of the set of algebraic equations corresponding to (9). The advantage of the method is high speed, which allows real-time imaging of the dynamic processes, for example, functional hemodynamic changes in the brain. However, the backprojection method provides a phenomenological rather than a self-consistent inversion of the forward problem, and its quantitative accuracy is low.

A more consistent method is based on the perturbation approach to the solution of the inverse problem (8). If we have an estimate $\{\mu_a(\mathbf{r}), \mu'_s(\mathbf{r})\}_e$ that is close to the ideal solution, then $G_e = F[\{\mu_a(\mathbf{r}), \mu'_s(\mathbf{r})\}_e]$ (see Eq. (7)) is close to S . Then the difference $G - G_e$ can be written as a Taylor series:

$$G - G_e = J[\{\mu_a(\mathbf{r}), \mu'_s(\mathbf{r})\}_e] \{\Delta\mu_a(\mathbf{r}), \Delta\mu'_s(\mathbf{r})\} + \dots \quad (10)$$

where $J[\dots]$ is the Jacobian, and $\{\Delta\mu_a(\mathbf{r}), \Delta\mu'_s(\mathbf{r})\} = \{\mu_a(\mathbf{r}), \mu'_s(\mathbf{r})\} - \{\mu_a(\mathbf{r}), \mu'_s(\mathbf{r})\}_e$. Neglecting terms after the first one in (10), the problem of finding the correction $\{\Delta\mu_a(\mathbf{r}), \Delta\mu'_s(\mathbf{r})\}$ to the estimate $\{\mu_a(\mathbf{r}), \mu'_s(\mathbf{r})\}_e$ reduces to the calculation of the Jacobian and to the solution of a linear algebraic equation.

This approach provides a basis for the PDWI algorithm, which is the following. First one makes an initial guess at the optical properties of the medium. Then one calculates the ac, dc, and phase of the PDW using a model of light propagation through the tissue and the geometry of the source–detector configuration. The calculated values are compared with the measured ones. If the error is above a set tolerance level, one calculates corrections to the current values of μ_a and μ'_s by solving Eq. (10). The algorithm is completed when the required error level is achieved. Thus, the PDWI algorithm iteratively applies both the forward and inverse problems. Since at each of many iterations one should solve the forward problem, which in the simplest case is a partial differential equation, the algorithm typically requires considerable computational resources.

One should note that the success of the method is largely dependent on the closeness of the initial guess to the actual values of the optical properties. In many studies the 'first guess' of the tissue optical properties was obtained by measuring the bulk absorption and reduced scattering coefficients. However, because of the ill-posedness of the inverse problem, homogeneous initial conditions did not always lead to the correct reconstruction. A more productive approach to the problem of the initial conditions is to assign approximate initial values of the optical properties known for example from *in vitro* measurements to different structures revealed by nonoptical imaging techniques, such as magnetic resonance imaging (MRI) or CT.

Applications

PDWI was proposed and developed as a noninvasive biomedical imaging technique. Although the spatial resolution of PDWI cannot compete with that of CT or MRI, PDWI has the advantages of low cost, high temporal resolution, and high biochemical specificity. The latter is based on the optical spectroscopic approach. Biological tissues are relatively transparent to light in the near-infrared band between 700 and 1000 nm (a near-infrared window). The major biological chromophores in this window are oxy- and deoxyhemoglobin, and water. One can separate contributions of these chromophores into light absorption by employing light sources at different wavelengths. Multi-wavelength PDWI can also be used to measure the redox state of the cytochrome oxidase, which is the terminal electron acceptor of the mitochondrial electron transport chain and responsible for 90% of cellular oxygen consumption.

One of the hot topics in cognitive neuroscience is the hemodynamic mechanism of the oxygen supply to neurons and removal of products of cerebral metabolism. High temporal resolution and biochemical specificity make PDWI a unique noninvasive tool for studies of functional cerebral hemodynamics, although light can penetrate no deeper than the very surface of the adult brain. Fig. 2 displays a typical arrangement of light sources and detectors used in PDWI of functional cerebral hemodynamics, and the images of the activated area in the brain.

Because of its low cost, noninvasiveness, and compact size of the optical sensors, PDWI is also considered to be a promising tool for mammography and structural imaging of anomalies in the brain. PDWI is an especially promising tool for neonatology, because the infant tissues have low absorption.

PDWI instruments employ lasers, light-emitting diodes and gas-discharge tubes as sources of light. Among the detectors there are photomultiplier tubes, avalanche photodiodes, and CCD cameras. Both sources and detectors can be coupled with the tissue by means of optical fibers. A number of commercially produced near-infrared instruments for PDWI are available on the market.

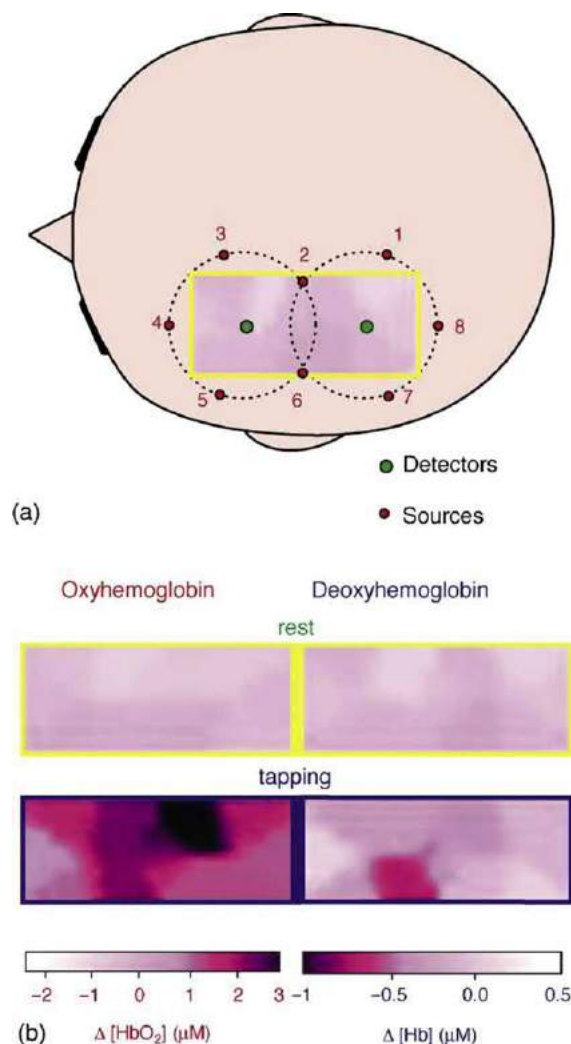


Fig. 2 (a) Geometrical arrangement of the 16 sources and two detectors on the head of a human subject. Each one of the eight small red circles (numbered 1–8) represents one pair of source fibers at 758 and 830 nm. Two larger green circles represent detector fiber bundles. The rectangle inside the head outline is the 4×9 cm² imaged area. (b) Maps of oxyhemoglobin ([HbO₂], left panels) and deoxyhemoglobin ([Hb], right panels) concentration changes at rest (top) and at the time of maximum response during the third tapping period (bottom). The measurement protocol involved hand-tapping with the right hand (contralateral to the imaged brain area). Adapted with permission from Franceschini MA, Toronov V, Filiaci M, Gratton E, Fantini S (2000) On-line optical imaging of the human brain with 160-ms temporal resolution. *Optics Express* **6**(3), 49–57 (Figs. 1 and 4).

According to the type of source modulation, there are three groups of instruments: frequency domain (harmonically modulated sources), time domain (pulsed sources), and steady state. Steady-state instruments are simpler and cheaper than frequency-domain and time-domain ones, but their capabilities are much lower. Time-domain instruments can theoretically provide a maximum of information about the tissue structure, but technically it is difficult to build a time-domain instrument with high temporal resolution and high signal-to-noise ratio. Frequency-domain instruments provide temporal resolution up to several milliseconds and have high quantitative accuracy in the determination of optical properties.

Further Reading

- Arridge, S.R., 1999. Optical tomography in medical imaging. *Inverse Problems* **15**, R41–R93.
- Boas, D.A., 1996. Diffuse Photon Probes of Structural and Dynamical Properties of Turbid Media: Theory and Biomedical Applications. PhD Thesis, University of Pennsylvania.
- National Research Council (US), Institute of Medicine (US), 1996. Mathematics and Physics of Emerging Biomedical Imaging. Washington, DC: National Academy Press.
- O’Leary, M.A., 1996. Imaging with Diffuse Photon Density Waves. PhD Thesis, University of Pennsylvania.
- Special Issue, 1997. Near-infrared spectroscopy and imaging of living systems. *Philosophical Transactions of the Royal Society B* **352** (1354).
- Tuchin, V., 2002. Handbook of Optical Biomedical Diagnostics. Bellingham, WA: SPIE Press.
- Tuchin, V.V., Luo, Q., Ulyanov, S.S., 2001. Imaging of Tissue Structure and Function. In: *Proceedings of SPIE* 4427. Bellingham, WA: SPIE Press.

- Villringer, A., Chance, B., 1997. Non-invasive optical spectroscopy and imaging of human brain function. *Trends in Neuroscience* 20, 435–442.
- Villringer, A., Dirnagl, U., 1993. *Optical Imaging of Brain Function and Metabolism*. New York: Plenum Press.
- Villringer, A., Dirnagl, U., 1997. *Optical Imaging of Brain Function and Metabolism 2: Physiological Basis and Comparison to Other Functional Neuroimaging Methods*. New York: Plenum Press.

Three-Dimensional Field Transformations

R Piestun, University of Colorado, Boulder, CO, USA

© 2005 Elsevier Ltd. All rights reserved.

Introduction

Conventional optical imaging is understood as an isomorphic mapping from an object space onto an image space. The object space is composed of scattering or emitting objects. The scattering objects can be opaque, transparent, or partially absorbing, while emitting objects can be as diverse as stars in astronomical imaging or fluorescent molecules in microscopic imaging. The image space is an optical field representation of the object.

In most classical applications of optical imaging the goal is to optimize the mapping between one plane in the object space and one plane in the image space. Modern optical systems, though, are aimed at acquiring three-dimensional information regarding the object. The image is thus a three-dimensional field representation of the object.

While many imaging systems are aimed at obtaining visually appealing images, many modern imaging systems register the field information in some type of sensitive device such as film or CCD cameras for later processing. This processing is typically performed electronically by a computer and the final image is ultimately a mathematical digital representation of the object. This type of imaging is called computational imaging. An example of computational imaging is optical tomography, which involves the registration of a set of images that do not necessarily resemble the object, and off-line digital processing to obtain a three-dimensional representation of the object.

In this article we will focus on imaging in the conventional sense but will also focus on the three-dimensional aspects of the image.

Geometrical Optics Transformations

In many optical problems it is appropriate to consider the wavelength of the electromagnetic wave as infinitely small. This approximation is the origin of the field of geometrical optics in which the propagation of light is formulated in terms of geometrical laws. Accordingly, the energy is transported along light rays and their trajectories are determined by certain differential or integral equations that can be directly derived from Maxwell's equations. One of these formulations uses Fermat's principle that states the principle of the shortest optical path. The optical path between two points P_1 and P_2 is defined as:

$$OP = \int_{P_1}^{P_2} n(P) ds \quad (1)$$

where n is the refractive index of the medium. Fermat's principle states that the optical path of a real ray is always a local minimum.

In ideal optical imaging systems, every point P_o of the so-called object space will generate a sharp image point P_i , that is a point where infinite rays emerging from P_o pass through P_i . Perfect imaging occurs when every curve defined by object points is geometrically similar to its image. Perfect imaging can occur between three-dimensional spaces or between surfaces.

An interesting result of geometrical optics is Maxwell's theorem. This theorem states that if an optical instrument produces a sharp image of a three-dimensional domain, the optical length of any curve in the object is equal to the optical length of its image. More explicitly, if the object space has index n_o and the image space index n_i , the theorem states that:

$$\int_{C_o} n_o ds_o = \int_{C_i} n_i ds_i \quad (2)$$

where C_i is the image of the object curve C_o .

An important consequence of Maxwell's theorem is a theorem by Carathéodory which states that any imaging transformation has to be a projective transformation, an inversion, or a combination of them. Another consequence of this theorem is that if the two media are homogeneous, perfect imaging can only occur if the magnification is equal to the ratio of the refractive indices in the object and image space. If the media have the same index, the imaging is trivial with unit magnification. The simplest example of a perfect imaging instrument is a planar mirror. Therefore, if we want to achieve nontrivial imaging between homogeneous spaces with equal index, we must give up either the perfect imaging (similarity) or the sharpness of the image.

The geometrical optics approximation is very useful in the design of optical imaging systems. However, it fails to provide a complete description of wave optical effects, notably diffraction. Therefore, the following sections provide a description of field transformations based on wave optics.

Systems Approach to 3D Imaging

In this article we deal with linear imaging systems. Hence the relation between three-dimensional (3D) input (object) signals and 3D output (image) signals can be completely specified by the 3D impulse response or point spread function (PSF) of the system $h(x, y, z; x', y', z')$. This function specifies the response of the system at image space coordinates (x, y, z) to a δ -function input (point source) at object coordinates (x', y', z') . The optical system object and image signals are related by the superposition integral:

$$i(x, y, z) = \int \int \int_{\Omega} o(x', y', z') h(x, y, z; x', y', z') dx' dy' dz' \quad (3)$$

where Ω is the object domain. It is clear that for a generic linear imaging system the image signal is completely specified by the response to input impulses located at all possible points of the image domain.

If the system is shift invariant, the PSF due to a point source located anywhere in the object domain will be a linearly shifted version of the response due to a reference point source. This approximation is in general not valid and the system's PSF is a function of the position of the point source in the object domain. Shift invariance is traditionally divided into radial and axial shift invariance. Radial shift variance is the change in the image of a point source as a function of its radial distance from the optical axis. Radial shift variance is caused by lens aberrations. Axial shift variance is the change in the image of a point source as a function of its axial position, z' , in the object space. Axial shift variance is inherent to any rotationally symmetric lens system because they can only have one pair of conjugate surfaces, i.e., object-image corresponding surfaces.

In most practical cases the image is detected at a single plane, $z = z_1$ transverse to the optical axis Oz . In those situations, we need to specify the impulse response on this plane:

$$h(x, y, z_1; x', y', z') = h_z(x, y; x', y') \quad (4)$$

The condition for lateral shift invariance is then expressed as $h_z(x, y; x', y') = \tilde{h}_z(x - x', y - y')$. In this case, the superposition integral is a two-dimensional (2D) convolution and it is possible to define the transfer function of the system that is the 2D Fourier transform of the impulse response $\tilde{h}_z(x - x', y - y')$.

Point Spread Function for Coherent, Incoherent, and Partially Coherent Illumination

The input-output relations of an optical imaging system relate different magnitudes according to the type of illumination. If the illumination is coherent, the system is linear in the complex amplitude, while if the illumination is fully incoherent the system is linear in intensity. The 3D impulse response of the incoherent (h_{Inc}) and coherent (h_{Coh}) systems are related as follows:

$$h_{\text{Inc}}(x, y, z; x', y', z') = |h_{\text{Coh}}(x, y, z; x', y', z')|^2 \quad (5)$$

Notwithstanding, many optical systems cannot be considered fully coherent or fully incoherent. In these situations the system is said partially coherent. If the optical path differences within the optical system are much shorter than the coherence length of the illumination, the light effectively has complete temporal coherence and the magnitude of interest is the mutual intensity. The mutual intensity is defined as the statistical correlation of the complex wavefunction $U(\mathbf{r}, t)$ between different points in space:

$$J(\mathbf{r}_1, \mathbf{r}_2) = \langle U^*(\mathbf{r}_1, t) U(\mathbf{r}_2, t) \rangle \quad (6)$$

where $\mathbf{r}_i = (x_i, y_i, z_i)$, $i = 1, 2$. The illumination is then called quasi-monochromatic. For the case of partially coherent quasi-monochromatic illumination, the system is linear in the mutual intensity. The partially coherent impulse response (h_{PC}) is also related to the coherent impulse response as follows:

$$h_{\text{PC}}(\mathbf{r}_1, \mathbf{r}_2; \mathbf{r}'_1, \mathbf{r}'_2) = h_{\text{Coh}}^*(\mathbf{r}_1; \mathbf{r}'_1) h_{\text{Coh}}(\mathbf{r}_2; \mathbf{r}'_2) \quad (7)$$

The 3D superposition integral that relate object and image mutual intensities is now:

$$J_{\text{image}}(\mathbf{r}_1, \mathbf{r}_2) = \int \int \int_{\Omega} h_{\text{PC}}(\mathbf{r}_1, \mathbf{r}_2; \mathbf{r}'_1, \mathbf{r}'_2) \times J_{\text{object}}(\mathbf{r}'_1, \mathbf{r}'_2) d\mathbf{r}'_1 d\mathbf{r}'_2 \quad (8)$$

Therefore, we conclude that knowledge of the coherent PSF is enough to characterize the optical system for illumination of various degrees of coherence.

Spatial Frequency Content of the 3D Point Spread Function

As previously stated, the coherent impulse response of the optical system completely defines its imaging characteristics. Therefore, it is justified to wonder what the limitations are in the specification of this response. In this section we analyze the frequency domain of coherent and incoherent impulse responses.

The impulse response of the optical system will satisfy the corresponding wave equations regardless of the particulars of the optical system. In particular, assuming a homogeneous isotropic medium in the imaging domain, the coherent response $h(\mathbf{r})$ to an

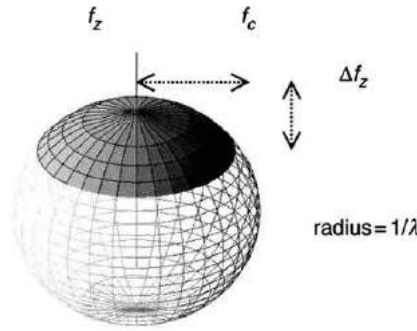


Fig. 1 Three-dimensional spatial-frequency domain of the coherent point spread function.

impulse at $\mathbf{r}'$ will satisfy the homogeneous Helmholtz equation in the image domain:

$$\nabla^2 h(\mathbf{r}) + k^2 h(\mathbf{r}) = 0 \quad (9)$$

where $h(\mathbf{r})$ is the complex amplitude and $k = 2\pi/\lambda$, with λ the wavelength in the medium. If the field at a given origin of coordinates $z = 0, f(x, y)$, is known, $h(\mathbf{r})$ can be calculated using the angular spectrum representation as follows:

$$h(x, y, z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(f_x, f_y) \exp[i2\pi(f_x x + f_y y + f_z z)] df_x df_y \quad (10)$$

where $f_z = \sqrt{\lambda^{-2} - (f_x^2 + f_y^2)}$ and $F(f_x, f_y) = \text{FT}\{f(x, y)\}$; FT represents a 2D Fourier transformation and f_x, f_y are the spatial frequencies along the x and y axes respectively.

Eq. (10) can be written as a 3D integral as follows:

$$h(x, y, z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(f_x, f_y) \cdot \delta\left[f_z - \sqrt{\lambda^{-2} - f_x^2 - f_y^2}\right] \exp[i2\pi(f_x x + f_y y + f_z z)] df_x df_y df_z \quad (11)$$

where δ represents the dirac impulse. Next we represent the 3D impulse response $h(x, y, z)$ as a 3D Fourier decomposition:

$$h(x, y, z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H(f_x, f_y, f_z) \exp[i2\pi(f_x x + f_y y + f_z z)] df_x df_y df_z \quad (12)$$

where $H(f_x, f_y, f_z) = \text{FT}^{3D}\{h(x, y, z)\}$. Comparing **Eqs. (11)** and **(12)** we can infer that:

$$H(f_x, f_y, f_z) = F(f_x, f_y) \cdot \delta\left[f_z - \sqrt{\lambda^{-2} - f_x^2 - f_y^2}\right] \quad (13)$$

This relation implies that the domain of the Fourier transform of the impulse response is confined to a spherical shell of radius $1/\lambda$. Alternatively, $H(f_x, f_y, f_z)$ is zero everywhere, except from the surface of the sphere.

The previous derivation did not consider any limitation in the range of spatial frequencies allowed through the system. We now assume a finite aperture that will introduce an effective radial cutoff frequency $f_c < 1/\lambda$ and will further limit the spatial spectral domain of $h(x, y, z)$. This is depicted in **Fig. 1**. Note that the transverse and longitudinal spatial frequencies are related, hence the longitudinal spatial frequencies are located within a bandpass of width $\Delta f_z = 1 - \sqrt{\lambda^{-2} - f_c^2}$.

Let us now consider the frequency content of possible incoherent PSFs. From **Eq. (5)** we derive the 3D Fourier transform of the incoherent impulse response:

$$\begin{aligned} H_{\text{Inc}}(f_x, f_y, f_z) &= \text{FT}\{h_{\text{Inc}}(x, y, z; x', y', z')\} \\ &= \text{FT}\{|h_{\text{Coh}}(x, y, z; x', y', z')|^2\} \\ &= H \otimes H(f_x, f_y, f_z) \end{aligned} \quad (14)$$

where $\otimes$ represents a 3D correlation. The domain of H_{Inc} now becomes a volume of revolution as shown in **Fig. 2**. Moreover, because h_{Inc} is real, H_{Inc} is conjugate symmetric, i.e., $H_{\text{Inc}} = H_{\text{Inc}}^*$. Now, the transverse cutoff frequency is $f_c^{\text{Inc}} = 2f_c$ while the longitudinal cutoff frequency is $f_z^{\text{Inc}} = 2\Delta f_z$.

Diffraction Calculation of the 3D Point Spread Function

The 3D-image field created by a point source in the object space can be predicted by proper modeling of the optical system. In the absence of aberrations, the distortion of the PSF is originated in the diffraction created by the limiting aperture of the optical system. Diffraction is thus the ultimate limitation to the resolution of the optical system. Under the design conditions of the

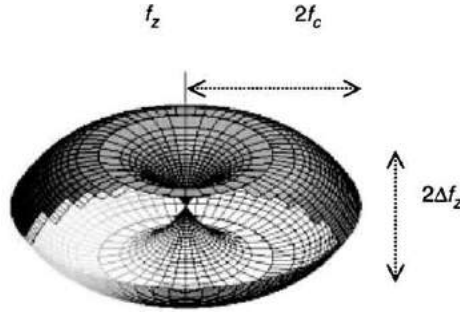


Fig. 2 Three-dimensional spatial-frequency domain of the incoherent point spread function.

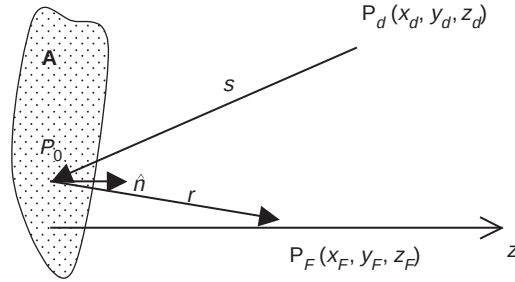


Fig. 3 Diffraction of a spherical wave by a planar aperture.

system, i.e., assuming all aberrations are corrected, the image of a point source is the diffraction pattern produced when a spherical wave converging to the (geometrical) image point is diffracted by the limiting aperture of the system. This is called a diffraction limited image.

There are many formulations to describe the 3D diffraction pattern obtained with a converging wave through an aperture. Here we review the Kirchhoff formulation. Let us consider the aperture A and the field at a point P_d originated from a converging spherical wave focused at point P_F (see Fig. 3). We define the vectors $\mathbf{r}$ joining a generic point on the aperture P_0 and P_F , and $\mathbf{s}$ joining P_0 and P_d . The Kirchhoff formulation states that the complex amplitude $U(P_d/P_F)$ is the integral over the aperture of the complex amplitude at each point propagated to P_d by a spherical wave $\exp(iks/s)$ modified by the directional factor $\cos(\hat{\mathbf{n}}, \mathbf{r}) - \cos(\hat{\mathbf{n}}, \mathbf{s})$ – $\cos(\hat{\mathbf{n}}, \mathbf{s})$. $\hat{\mathbf{n}}$ is the unit vector normal to the aperture:

$$U(P_d/P_F) = \frac{-iC}{2\lambda} \int \int_A \frac{\exp[-ik(r-s)]}{rs} [\cos(\hat{\mathbf{n}}, \mathbf{r}) - \cos(\hat{\mathbf{n}}, \mathbf{s})] dS \quad (15)$$

where $r = \|\mathbf{r}\|$, $s = \|\mathbf{s}\|$, and C is the amplitude of the spherical wave.

The directional factor varies slowly and in many situations can be approximated by a constant equal to 2 for small image-field angles. The amplitude over the aperture is also slowly varying and can be approximated by a constant, i.e., $C/r \approx C_0$. Different formulations of the Kirchhoff integral differ in the estimates of the factor $1/s$ and the phase $k(r-s)$. For example, one such formulation leads to the following expression for the 3D diffraction pattern for a focal point on the axis of an optical system with circular aperture:

$$U(P_d/P_F) = \frac{-ika^2 C_0}{z_d} \exp \left[ik \left(z_d - z_F + \frac{r_d^2}{2z_d} \right) \right] \times \int_0^1 J_0 \left(\frac{k a \rho r_d}{z_d} \right) \exp \left[\frac{-ika^2 \rho^2 (z_d - z_F)}{2z_d z_F} \right] \rho d\rho \quad (16)$$

where a is the radius of the aperture, r_d is the radial coordinate of P_d , and ρ is the radial coordinate of P_0 .

To illustrate the use of Eq. (16), we calculated the intensity around the focus of a spherical wave diffracted through a circular aperture of radius $a = 0.5$ cm, focusing at distance $z_F = 4$ cm from the screen. We assumed the wavelength to be $\lambda = 0.5 \mu$. Fig. 4 depicts the magnitude of the complex amplitude in a meridional plane near the focus of the converging spherical wave.

Models that take into account off-axis focal points have also been developed. For the most accurate calculations, it is necessary to consider the full electromagnetic nature of the optical field.

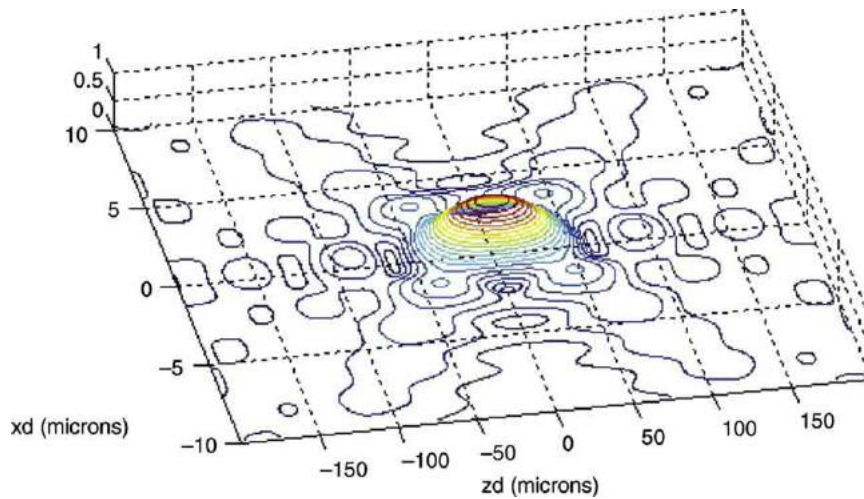


Fig. 4 Lines of equal intensity in a meridional plane near the focus of a converging spherical wave diffracted at a circular aperture ($\lambda = 0.5 \mu$, $a = 0.5$ cm, $z_F = 4$ cm).

Conclusions

Classical imaging theory can be applied to the 3D imaging transformation from object to image. The systems approach to the imaging problem implies that knowledge of the 3D coherent PSF suffices to predict the 3D response of the system under illumination of varying degrees of coherence. The PSF response can be predicted using well-known formulations of diffraction.

Further Reading

- Born, M., Wolf, E., 1999. *Principles of Optics*, 7th edn. New York: Cambridge University Press.
- Gibson, S.F., Lanni, F., 1989. Diffraction by a circular aperture as a model for three-dimensional optical microscopy. *Journal of the Optical Society of America A* 6, 1357–1367.
- Goodman, J.W., 1996. *Introduction to Fourier Optics*. New York: McGraw-Hill.
- Li, Y., Wolf, E., 1984. Three-dimensional intensity distributions near the focus in systems of different Fresnel numbers. *Journal of the Optical Society of America A* 1, 801–808.
- Piestun, R., Shamir, J., 2002. Synthesis of 3D light fields and applications. *Proceedings of the IEEE* 90, 222–244.
- Saleh, B.E.A., Teich, M.C., 1991. *Fundamentals of Photonics*. New York: Wiley.
- Stamnes, J.J., 1986. *Waves in Focal Regions*. Bristol: Hilger.

Volume Holographic Imaging

G Barbastathis, Massachusetts Institute of Technology, Cambridge, MA, USA

© 2005 Published by Elsevier Ltd.

Volume holographic imaging (VHI) refers to a class of imaging techniques which use a volume hologram in at least one location within the optical path. The volume hologram acts as a 'smart lens,' which processes the optical field to extract spatial information in three dimensions (lateral as well as longitudinal) and spectral information.

Fig. 1 is a generic VHI system. The object is either illuminated by a light source (e.g., sunlight or a pump laser) as shown in the figure, or it may be self-luminous. Light scattered or emitted by the object is first transformed by an objective lens and then illuminates the volume hologram. The role of the objective is to form an intermediate image which serves as input to the volume hologram. The volume hologram itself is modeled as a three-dimensional (3D) modulation $\Delta\epsilon(\mathbf{r})$ of the dielectric index within a finite region of space. The light entering the hologram is diffracted by $\Delta\epsilon(\mathbf{r})$ with efficiency η , defined as

$$\eta = \frac{\text{Power diffracted by the volume hologram}}{\text{Power incident to the volume hologram}} \quad (1)$$

We assume that diffraction occurs in the Bragg regime. The diffracted field is Fourier-transformed by the collector lens, and the result is sampled and measured by an intensity detector array (such as a CCD or CMOS camera).

Intuitively, we expect that a fraction of the illumination incident upon the volume hologram is Bragg-matched and is diffracted towards the Fourier-transforming lens. The remainder of the incident illumination is Bragg-mismatched, and as result is transmitted through the volume hologram un-diffracted. Therefore, the volume hologram acts as a filter which admits the Bragg-matched portion of the object and rejects the rest. When appropriately designed, this 'Bragg imaging filter' can exhibit very rich behavior, spanning the three spatial dimensions and the spectral dimension of the object.

To keep the discussion simple, we consider the specific case of a transmission geometry volume hologram, described in **Fig. 2**. The volume hologram is created by interfering a spherical wave and a planewave, as shown in **Fig. 2(a)**. The spherical wave originates at the coordinate origin. The planewave is off-axis and its wavevector lies on the xz plane. As in most common holographic systems, the two beams are assumed to be at the same wavelength λ , and mutually coherent. The volume hologram results from exposure of a photosensitive material to the interference of these two beams.

First, assume that the object is a simple point source. The intermediate image is also approximately a point source, that we refer to as 'probe,' located somewhere in the vicinity of the reference point source. Assuming the wavelength of the probe source is the same as that of the reference and signal beams, volume diffraction theory shows that:

- (i) If the probe point source is displaced in the y direction relative to the reference point source, the image formed by the volume hologram is also displaced by a proportional amount. Most common imaging systems would be expected to operate this way.
- (ii) If the probe point source is displaced in the x direction relative to the reference point source, the image disappears (i.e., the detector plane remains dark).
- (iii) If the probe point source is displaced in the z direction relative to the reference point source, a defocused and faint image is formed on the detector plane. 'Faint' here means that the fraction of energy of the defocused probe transmitted to the detector plane is much smaller than the fraction that would have been transmitted if the probe had been at the origin.

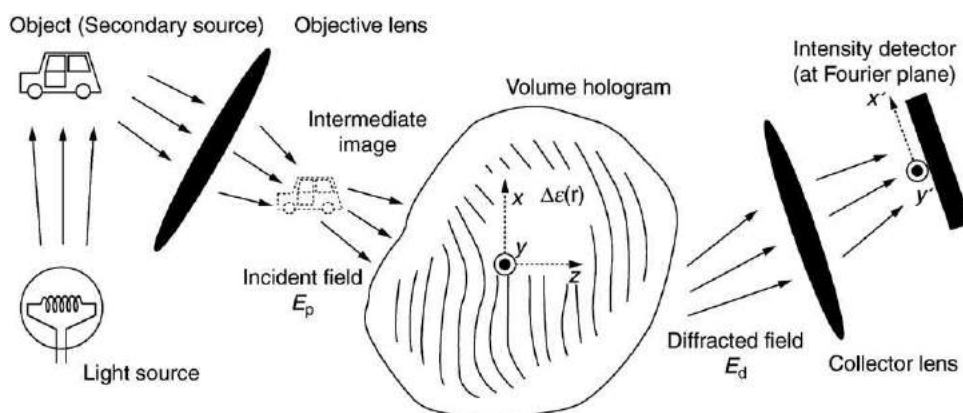


Fig. 1 Volume holographic imaging (VHI) system.

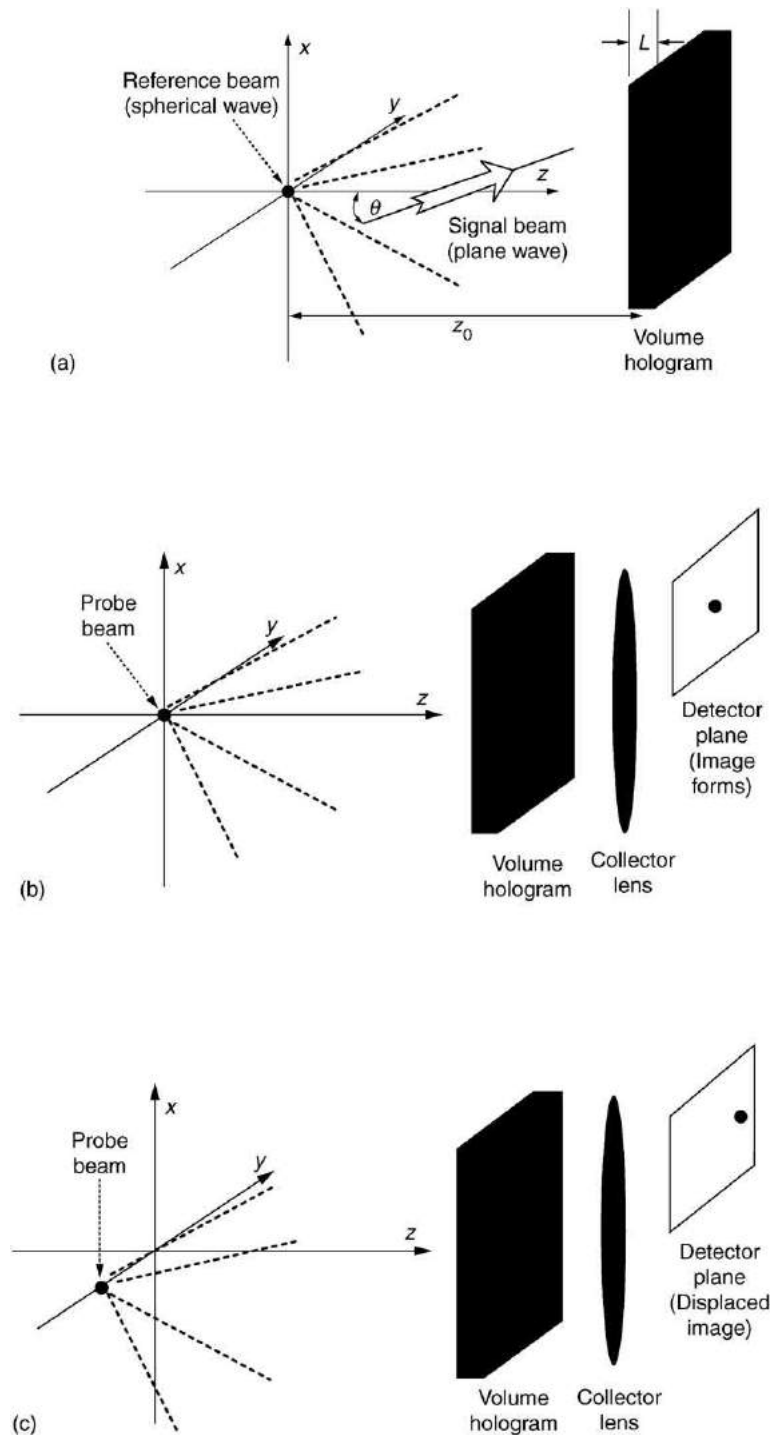


Fig. 2 (a) Recoding of a transmission-geometry volume hologram with a spherical wave and an off-axis plane wave with its wave-vector on the xz plane. (b) Imaging of a probe point source that replicates the location and wavelength of the reference point source using the volume hologram recorded in part (a). (c) Imaging of a probe point source at the same wavelength but displaced in the y direction relative to the reference point source. (d) Imaging of a probe point source at the same wavelength but displaced in the x direction relative to the reference point source. (e) Imaging of a probe point source at the same wavelength but displaced in the z direction relative to the reference point source. (f) 'Bragg slitting:' Imaging of an extended monochromatic, spatially incoherent object using a volume hologram recorded as in (a). (g) Joining Bragg slits from different colors to form rainbow slices: Imaging of an extended polychromatic object using a volume hologram recorded as in (a). (h) Multiplex imaging of several slices using a volume hologram formed by multiple exposures.

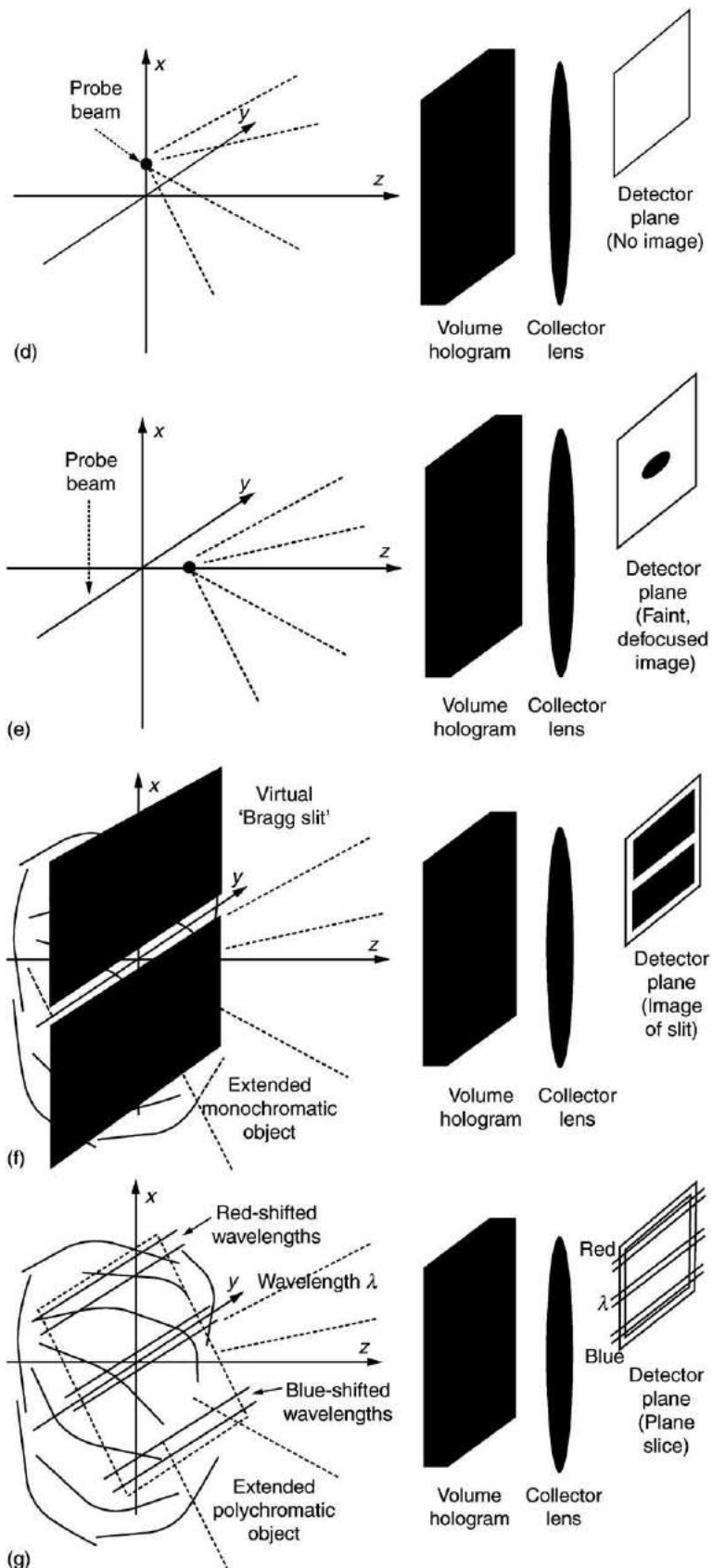


Fig. 2 (Continued).

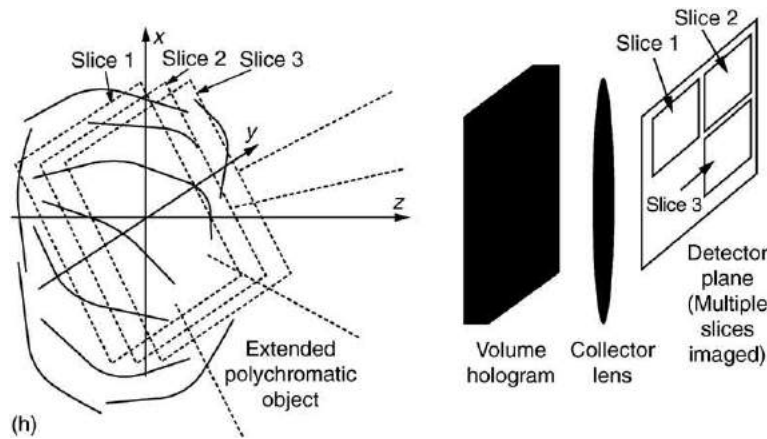


Fig. 2 (Continued).

Now consider an extended, monochromatic, spatially incoherent object and intermediate image, as in [Fig. 2\(f\)](#). According to the above description, the volume hologram acts as a ‘Bragg slit’ in this case: because of Bragg selectivity, the volume hologram transmits light originating from the vicinity of the y axis, and rejects light originating anywhere else. For the same reason, the volume hologram affords depth selectivity (like the pinhole of a confocal microscope). The width of the slit is determined by the recording geometry, and the thickness of the volume hologram. For example, in the transmission recording geometry of [Fig. 2\(a\)](#), where the reference beam originates a distance z_0 away from the hologram, the planewave propagates at angle θ with respect to the optical axis z (assuming $\theta \ll 1$ radian), and the hologram thickness is L (assuming $L \ll z_0$), the width of the slit is found to be

$$\Delta x \approx \frac{\lambda z_0}{L\theta} \quad (2)$$

The imaging function becomes richer if the object is polychromatic. In addition to its Bragg slit function, the volume hologram exhibits then dispersive behavior, like all diffractive elements. In this particular case, dispersion causes the hologram to image simultaneously multiple Bragg slits, each at a different color and parallel to the original slit at wavelength λ , but displaced along the z axis. Light from all these slits finds itself in focus at the detector plane, thus forming a ‘rainbow image’ of an entire slice through the object, as shown in [Fig. 2\(g\)](#).

To further exploit the capabilities of volume holograms, we recall that in general it is possible to ‘multiplex’ (superimpose) several volume gratings within the same volume by successive exposures. In the imaging context, suppose that we multiplex several gratings similar to the grating described in [Fig. 2\(a\)](#) but with spherical reference waves originating at different locations and plane signal waves at different orientations. When the multiplexed volume hologram is illuminated by an extended polychromatic source, each grating forms a separate image of a rainbow slice, as described earlier. By spacing appropriately the angles of propagation of the plane signal waves, we can ensure that the rainbow images are formed on nonoverlapping areas on the detector plane, as shown in [Fig. 2\(h\)](#). This device is now performing true four-dimensional (4D) imaging: it is separating the spatial and spectral components of the object illumination so that they can be measured independently by the detector array. Assuming the photon count is sufficiently high and the number of detector pixels is sufficient, this ‘spatio-spectral slicing’ operation can be performed in real time, without need for mechanical scanning.

The complete theory of VHI in the spectral and spatial domains is given in the Further Reading section below. Experimental demonstrations of VHI have been performed in the context of a confocal microscope where the volume hologram performs the function of a ‘Bragg pinhole’ as well as a real-time 4D imager. The limited diffraction efficiency η of volume holograms poses a major concern for VHI systems. Holograms, however, are known to act as matched filters. It has been shown that the matched filtering nature of volume holograms as imaging elements is superior to other filtering elements (e.g., pinholes or slits) in an information-theoretic sense. Efforts are currently underway to strengthen this property by design of volume holographic imaging elements which are more elaborate than described here.

See also: Overview: Holography

Further Reading

- Barbastathis, G., Brady, D.J., 1999. Multidimensional tomographic imaging using volume holography. *Proceedings of the IEEE* 87 (12), 2098–2120.
 Barbastathis, G., Sinha, A., 2001. Information content of volume holographic images. *Trends in Biotechnology* 19 (10), 383–392.
 Barbastathis, G., Balberg, M., Brady, D.J., 1999. Confocal microscopy with a volume holographic filter. *Optics Letters* 24 (12), 811–813.

- Barbastathis, G., Sinha, A., Neifeld, M.A., 2001. Information-theoretic treatment of axial imaging. In: OSA Topical Meeting on Integrated Computational Imaging Systems, pp. 18–20. Albuquerque, NM.
- Liu, W., Psaltis, D., Barbastathis, G., 2002. Real time spectral imaging in three spatial dimensions. *Optics Letters* 27 (10), 854–856.
- van-der Lugt, A., 1964. Signal detection by complex spatial filtering. *IEEE Transactions on Information Theory* IT-10 (2), 139–145.

Wavefront Sensors and Control (Imaging Through Turbulence)

CL Matson, Air Force Research Laboratory, Kirtland, NM, USA

© 2005 Elsevier Ltd. All rights reserved.

Introduction

In many situations, distant objects are imaged using optical or near-infrared imaging systems. Examples include terrestrial surveillance from space, tactical surveillance from airborne imaging systems, and ground-based astronomical imaging. Ideally, the resolution in these images is limited only by the wavelength of light λ used to create the images divided by the imaging system's effective diameter D . The ratio λ/D is called the diffraction limit of the imaging system. Unfortunately, when the imaging system is located in the Earth's atmosphere, it is often atmospheric turbulence fluctuations that limit the level of detail in the long-exposure images, not the diffraction limit of the telescope. For example, the Keck telescopes, located on top of Mauna Kea in Hawaii, are 10 m diameter telescopes that have a diffraction-limited angular resolution of $0.2 \mu\text{rad}$ for an imaging wavelength of $2 \mu\text{m}$. However, atmospheric turbulence limits the achievable resolution in long-exposure images to the order of $1\text{--}5 \mu\text{rad}$ at this wavelength.

The important discovery by Anton Labeyrie in 1970 that short-exposure images of objects retain meaningful diffraction-limited information about the object, where 'short' means that the exposure time is less than the atmospheric turbulence coherence time, has revolutionized imaging through turbulence. Atmospheric coherence times depend upon a number of factors, including wind speeds, wind altitudes, and observation wavelength. At astronomical observatories for visible observation wavelengths, a rule of thumb is that atmospheric turbulence stays essentially constant for exposure times of one to ten milliseconds under average clear-sky weather conditions. A variety of techniques that exploit this discovery have been developed. These techniques can be grouped into three different categories: techniques that remove blurring from the detected intensity images (post-detection compensation), techniques that modify the optical field prior to detecting the light (pre-detection compensation), and hybrid techniques that use both post-detection and pre-detection compensation. In this article, a variety of these techniques are presented that are representative of the myriad of methods that have been developed during the past decades. To simplify the discussion, it is assumed that the blurring due to atmospheric turbulence is the same at every point in the image (isoplanaticity), that the images are formed using incoherent light as emitted by the sun and stars, and that a single filled aperture is used to create the images. Interferometric techniques that use multiple apertures to estimate the mutual coherence function of the optical field and calculate an image from it.

The emphasis in this article is on identifying and removing atmospheric blurring in images. As part of this process, regularization techniques must be included in the image reconstruction algorithms in order to obtain the best possible resolution for an acceptable level of noise. Although regularization techniques are not discussed in this article, information can be found in the Further Reading section at the end of this article.

Imaging and Atmospheric Turbulence

In Fig. 1, a conceptual diagram illustrates the impact that atmospheric turbulence has upon resolution in classical imaging systems. A distant star is being imaged by a telescope/detector system. In the absence of atmospheric turbulence, the field emitted by the star has a planar wavefront over the extent of the lens since the star is a long distance from the imaging system and thus unresolved by the telescope. When imaged by the telescope, the diffraction pattern of the telescope is obtained, resulting in a diffraction-limited image (Fig. 2). However, in the presence of atmospheric turbulence near the telescope, the planar wavefront is randomly distorted by the turbulence, resulting in a corrugated wavefront with a spatial correlation scale given by the Fried parameter r_0 . For visible wavelengths, at good astronomical observing sites, r_0 ranges from 5 cm to 20 cm. When this corrugated wavefront is converted to an image by the telescope, a distorted image is obtained (Fig. 3). Because atmospheric turbulence is time-varying, the distortions in the image also change over time. If the imaging shutter on the telescope remains open for many occurrences of atmospheric turbulence, the resulting image will be an average of all the individually distorted images, resulting in what is known as the seeing disk (Fig. 4). It is the seeing disk that typically limits spatial resolution in long-exposure images, not the diffraction pattern of the telescope. As a result, the effective spatial resolution in long-exposure images in the presence of atmospheric turbulence is given by λ/r_0 , not λ/D .

Atmospheric turbulence distorts a field's wavefront by delaying it in a way that is spatially and temporally random. The origin of these delays is the random variation in the index of refraction of the atmosphere as a result of turbulent air motion brought about by temperature variations. A Fourier optics imaging model is useful in characterizing the effects of atmospheric turbulence on spatial resolution in images. Let $i(\vec{x})$ be an image of an object $o(\vec{x})$ where $\vec{x}$ is a two-dimensional spatial vector, and let $I(\vec{f})$ and $O(\vec{f})$ be their spatial Fourier transforms, where $\vec{f}$ is a two-dimensional spatial frequency vector. In the absence of atmospheric turbulence, the diffraction-limited Fourier transform $I(\vec{f})$ is given by

$$I_a(\vec{f}) = O(\vec{f})H_d(\vec{f}) \quad (1)$$

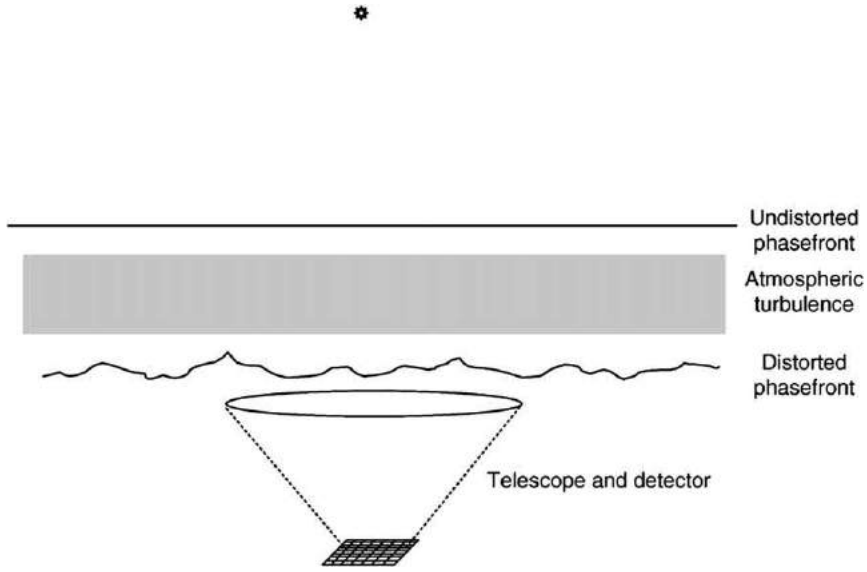


Fig. 1 Conceptual diagram of imaging through atmospheric turbulence.

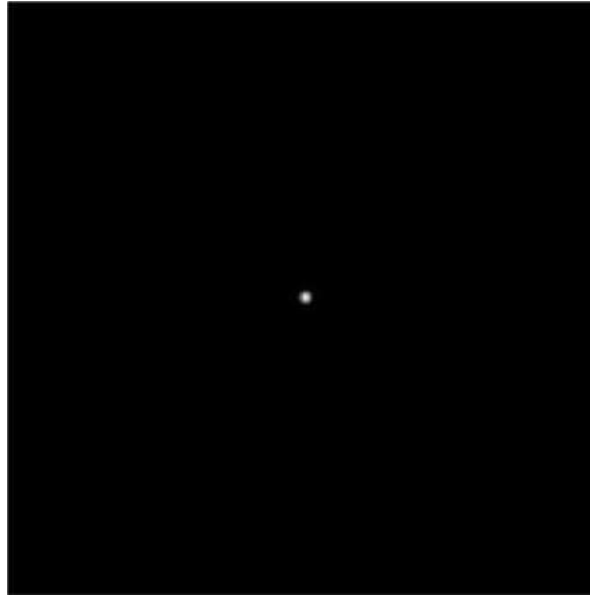


Fig. 2 Computer-simulated diffraction-limited image of a point source.

where the subscript 'd' denotes the diffraction-limited case and where $H_d(\vec{f})$ is the diffraction-limited telescope transfer function. In the presence of atmospheric turbulence, the Fourier transform of a long exposure image $i_{le}(\vec{x})$ is given by

$$I_{le}(\vec{f}) = O(\vec{f})H_d(\vec{f})H_{le}(\vec{f}) = I_d(\vec{f})H_{le}(\vec{f}) \quad (2)$$

where the subscript 'le' denotes long exposure. If the turbulence satisfies Kolmogorov statistics, as is typically assumed, then:

$$H_{le}(\vec{f}) = \exp \left[-3.44 \left(\frac{\lambda d |\vec{f}|}{r_o} \right)^{5/3} \right] \quad (3)$$

where d is the distance between the imaging system's exit pupil and detector plane. As can be seen from Eq. (2), the effect of atmospheric turbulence on a long-exposure image is a multiplication of the diffraction-limited image's Fourier transform by a transfer function due to atmospheric turbulence. In Fig. 5, an amplitude plot of a diffraction-limited transfer function is displayed along with an amplitude plot of a long-exposure transfer function for a 2.3 m diameter telescope, an imaging wavelength of

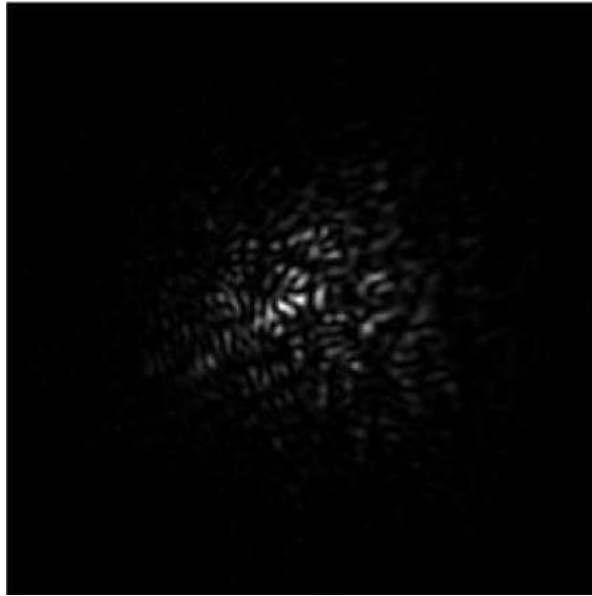


Fig. 3 Computer-simulated short-exposure image of a point source through atmospheric turbulence.

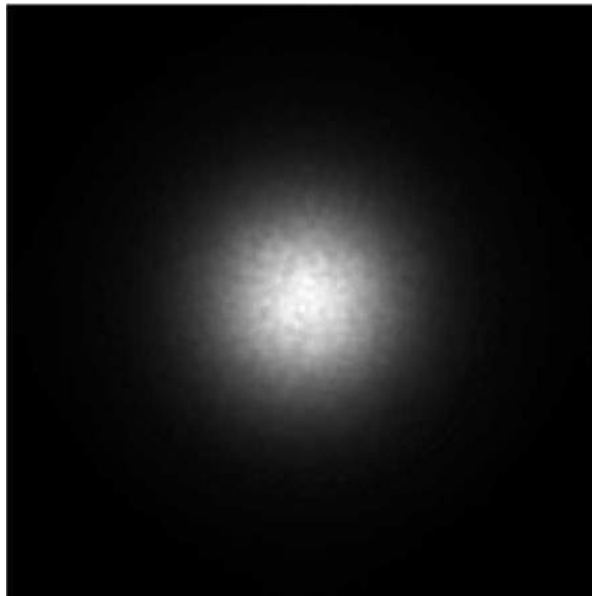


Fig. 4 Computer-simulated long-exposure image of a point source through atmospheric turbulence.

0.5 μm , and an r_0 value of 10 cm. Notice that the long exposure transfer function significantly reduces the amount of high spatial frequency information in an image. Diffraction-limited and long-exposure turbulence-limited images of a binary star are shown in [Figs. 6](#) and [7](#), illustrating the significant loss of spatial resolution brought on by atmospheric turbulence.

Post-Detection Compensation

Speckle Imaging

Labeyrie's 1970 discovery was motivated by his analysis of short-exposure star images. In these images (see [Fig. 3](#)), a number of speckles are present, where the individual speckle sizes are roughly the size of the diffraction-limited spot of the telescope. From this, he realized that high spatial frequency information is retained in short-exposure images, albeit distorted. He then showed that

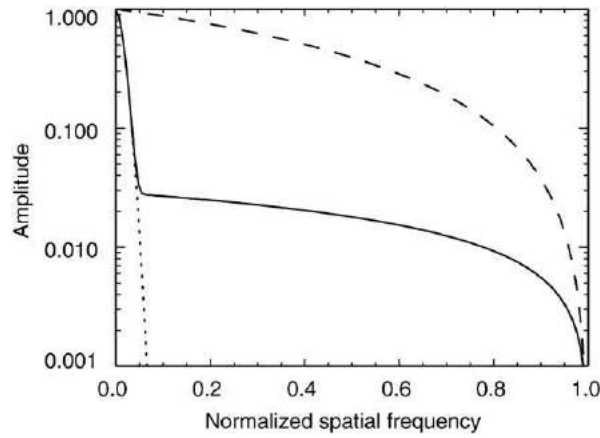


Fig. 5 Plots of the Fourier amplitudes of the diffraction-limited transfer function (dashed line), Fourier amplitudes of the long-exposure transfer function (dotted line), and the square root of the average short exposure energy spectrum. The spatial frequency axis is normalized to one at the diffraction limit of the telescope.

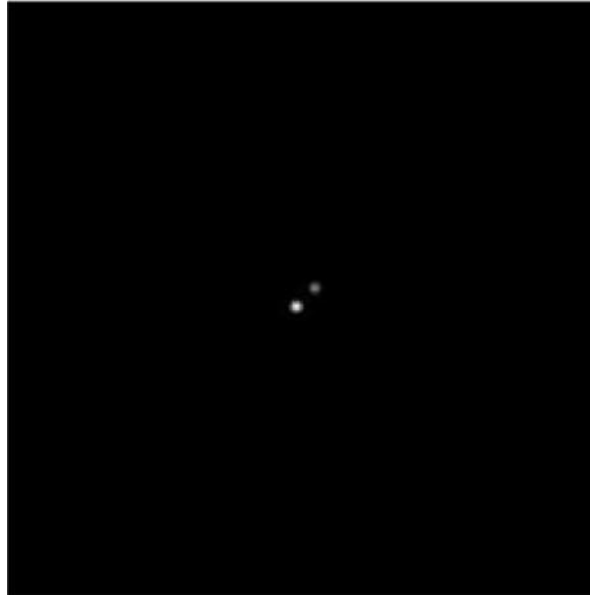


Fig. 6 Computer-simulated diffraction-limited image of a binary star.

the energy spectra $E_1(\vec{f})$ of the individual short-exposure images can be averaged to obtain an estimate of the average energy spectrum that retains high signal-to-noise ratio (SNR) information at high spatial frequencies. The square root of the average energy spectrum (the amplitude spectrum) of a star, calculated using Labeyrie's method (also known as speckle interferometry), is shown in Fig. 5, along with the diffraction-limited and long-exposure amplitude spectra. Notice that meaningful high spatial frequency information is retained in the average amplitude spectrum out to the diffraction limit of the telescope. Because the atmospherically corrupted amplitude spectrum estimate is significantly lower than the diffraction-limited amplitude, a number of short exposure images must be used to build up the SNRs at the higher spatial frequencies.

To reconstruct an image, both the Fourier amplitudes and Fourier phases of the image must be known. In fact, it is well known that the Fourier phase of an image carries most of the information in the image. Because only the Fourier amplitudes are available using Labeyrie's method, other techniques were subsequently developed to calculate the Fourier phase spectrum from a sequence of short-exposure images. Techniques that combine amplitude spectrum estimation using Labeyrie's method along with phase spectrum estimation are called speckle imaging techniques. The two most widely used phase spectrum estimation methods are the Knox-Thompson (or cross spectrum) and bispectrum algorithms. Both of these algorithms calculate Fourier-domain quantities that retain high spatial frequency information about the Fourier phase spectrum when averaged over multiple short-exposure images. The cross spectrum and bispectrum techniques calculate the average cross spectrum $C(\vec{f}, \Delta\vec{f})$ and bispectrum $B(\vec{f}_1, \vec{f}_2)$

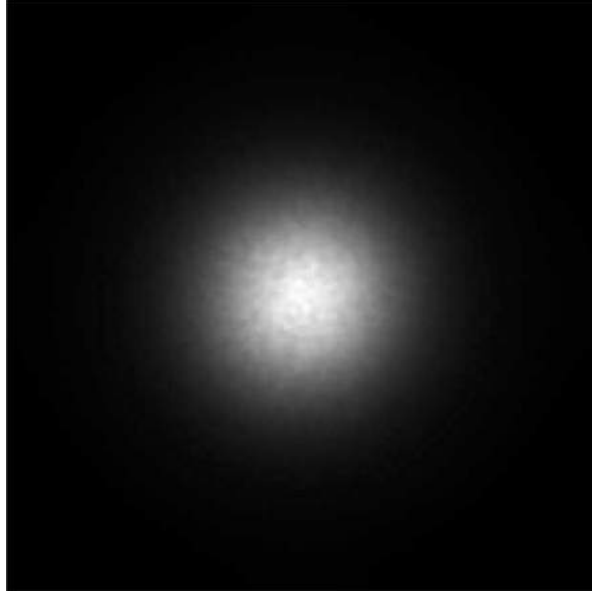


Fig. 7 Computer-simulated long-exposure image of a binary star.

respectively, that are given by

$$C(\vec{f}, \Delta\vec{f}) = I(\vec{f})I^*(\vec{f} + \Delta\vec{f}) \quad (4)$$

$$B(\vec{f}_1, \vec{f}_2) = I(\vec{f}_1)I(\vec{f}_2)I^*(\vec{f}_1 + \vec{f}_2) \quad (5)$$

where the superscript * denotes complex conjugate. Although the cross spectrum is defined for all values of $\vec{f}$ and $\Delta\vec{f}$ only for values of $\Delta\vec{f}$ or $\vec{f}$ less than r_o/λ are the cross spectrum elements obtained with sufficiently high SNRs to be useful in the reconstruction process. The equivalent constraint on the bispectrum elements is that either $\vec{f}_1, \vec{f}_2$ or $\vec{f}_1 + \vec{f}_2$ must be less than r_o/λ . Once the average cross spectrum or bispectrum is calculated from a series of short-exposure images, the phase spectrum $\phi_1(\vec{f})$ of the underlying image can be calculated from their phases using one of a number of algorithms based upon the following equations:

$$\phi_1(\vec{f} + \Delta\vec{f}) = \phi_1(\vec{f}) - \phi_c(\vec{f} + \Delta\vec{f}) \quad (6)$$

for the cross spectrum and:

$$\phi_1(\vec{f}_1 + \vec{f}_2) = \phi_1(\vec{f}_1) + \phi_1(\vec{f}_2) - \phi_b(\vec{f}_1 + \vec{f}_2) \quad (7)$$

for the bispectrum. Because the phases are only known for units of 2π , phasor versions of Eqs. (6) and (7) must be used. For the cross spectrum technique, each short-exposure image must be centroided prior to calculating its cross spectrum and adding it to the total cross spectrum. The bispectrum technique is insensitive to tilt, so no centroiding is necessary; however, the bispectrum technique requires setting two image phase spectrum values to arbitrary values to initialize the phase reconstruction process. Typically, the image phase spectrum values at $\vec{f} = (0, 1)$ and $\vec{f} = (1, 0)$ are set equal to zero, which centroids the reconstructed image but does not affect its morphology.

To obtain an estimated Fourier transform $I_e(\vec{f})$ of the image from the energy and phase spectrum estimates, the square root of the average energy spectrum is combined with the average phase spectrum calculated from either the bispectrum or cross spectrum. The result is given by

$$I_e(\vec{f}) = [E_1(\vec{f})]^{1/2} \exp[j\phi_1(\vec{f})] \quad (8)$$

where $j = \sqrt{-1}$. The result in Eq. (8) includes atmospheric attenuation of the Fourier amplitudes, so it must be divided by the amplitude spectrum of the atmosphere to obtain the desired object spectrum estimate. The atmospheric amplitude spectrum estimate is obtained by using Labeyrie's technique on a sequence of short-exposure images of a nearby unresolved star. No phase spectrum correction is necessary because it has been shown theoretically and verified experimentally that atmospheric turbulence has a negligible effect upon $\phi_1(\vec{f})$.

Blind Deconvolution

Speckle imaging techniques are effective as well as easy and fast to implement. However, it is necessary to obtain a separate estimate of the atmospheric energy spectrum using an unresolved star in order to reconstruct a high-resolution image. Usually, the star is not co-located with the object and the star images are not collected simultaneously with the images of the object. As a result,

the atmospheric energy spectrum estimate may be inaccurate. At the very least, collecting the additional star data expends valuable observation time. In addition, there is often meaningful prior knowledge about the object and the data, such as the fact that measured intensities are positive (positivity), that many astronomical objects are imaged against a black background (support), and that the measured data are band-limited, among others. Speckle imaging algorithms do not exploit this additional information. As a result, a number of techniques have been developed to jointly estimate the atmospheric blurring effects as well as an estimate of the object being imaged using only the short-exposure images of the object and available prior knowledge. These techniques are called blind deconvolution techniques. Two of the most common techniques are known as multi-frame blind deconvolution (MFBD) and projection-based blind deconvolution (PBBBD).

The MFBD algorithm is based upon maximizing a figure of merit to determine the best object and atmospheric short-exposure point spread function (PSF) estimates given the data and available prior knowledge. This figure of merit is created using a maximum likelihood problem formulation. Conceptually, this technique produces estimates of the object and the atmospheric turbulence short-exposure PSFs that most likely generated the measured short-exposure images. Mathematically, the MFBD algorithm obtains these estimates by maximizing the conditional probability density function (pdf) $p[i_m(\vec{x})|_o(\vec{x}); h_m(\vec{x})]$, $m = 1, \dots, M$, where $i_m(\vec{x})$ is the m th of M short exposure images and is given by

$$i_m(\vec{x}) = h_m(\vec{x}) * o(\vec{x}) + n_m(\vec{x}) \quad (9)$$

where $h_m(\vec{x})$ and $n_m(\vec{x})$ are the atmospheric/system blur and detector noise functions for the m th image, respectively, and the asterisk denotes convolution. The noise is assumed to be zero mean with variance $\sigma^2(\vec{x})$. The conditional pdf $p[i_m(\vec{x})|_o(\vec{x}); h_m(\vec{x})]$, $m = 1, \dots, M$ is the pdf of the noise with a mean equal to $h_m(\vec{x}) * o(\vec{x})$ and, as a function of $o(\vec{x})$ and $h_m(\vec{x})$, it is known as a likelihood function. For spatially independent Gaussian noise, as is the case for camera read noise, the likelihood function is given by

$$L_G[o(\vec{x}); h_m(\vec{x}), m = 1, \dots, M] = \prod_m \prod_{\vec{x}} \frac{1}{\sqrt{2\pi\sigma^2(\vec{x})}} \exp \left\{ -[i(\vec{x}) - (o * h_m)(\vec{x})]^2 / 2\sigma^2(\vec{x}) \right\} \quad (10)$$

For Poisson noise, the likelihood function is given by

$$L_P[o(\vec{x}); h_m(\vec{x}), m = 1, \dots, M] = \prod_m \prod_{\vec{x}} \frac{[(o * h_m)(\vec{x})]^{i(\vec{x})} \exp\{-[(o * h_m)(\vec{x})]\}}{i(\vec{x})!} \quad (11)$$

In practice, logarithms of Eqs. (10) and (11) are taken prior to searching for the estimates of $o(\vec{x})$ and $h_m(\vec{x})$ that maximize the likelihood function.

In general, there are an infinite number of estimates of $o(\vec{x})$ and $h_m(\vec{x})$ that reproduce the original dataset when convolved together. However, when additional prior knowledge is included to constrain the solution space, such as positivity, support, and the knowledge that $h_m(\vec{x})$ is band-limited, along with regularization, the MFBD algorithm almost always converges to the correct estimates of $o(\vec{x})$ and $h_m(\vec{x})$.

The PBBBD technique also uses prior knowledge constraints to jointly estimate $o(\vec{x})$ and $h_m(\vec{x})$. However, instead of creating and maximizing a likelihood function, the PBBBD technique uses the method of convex projections to find estimates of $o(\vec{x})$ and $h_m(\vec{x})$ that are consistent with the measured short-exposure images and all the prior knowledge constraints. The data consistency constraint and the prior knowledge constraints are mathematically formulated as convex sets and the PBBBD technique repeatedly takes the current estimates of $o(\vec{x})$ and $h_m(\vec{x})$ and projects them onto the convex constraint sets. As long as all the sets are truly convex, the algorithm is guaranteed to converge to a solution that is in the intersection of all the sets as long as the algorithm doesn't stagnate at erroneous solutions called traps. These traps are the projection equivalent of local minima in cost function minimization algorithms.

A key component of this algorithm is the ability to formulate the measured short-exposure images and prior knowledge constraints into convex sets. As an example, consider the convex set $C_{\text{pos}}[u(\vec{x})]$ that corresponds to the positivity constraint. This set is given by

$$C_{\text{pos}}[u(\vec{x})] \triangleq \{u(\vec{x}) \text{ such that } u(\vec{x}) \geq 0\} \quad (12)$$

It is straightforward to show that $C_{\text{pos}}[u(\vec{x})]$ is a convex set. Other constraints can be converted into convex sets in a similar manner. Although all constraints can be formulated into sets, most but not all of these sets are convex.

The projection process is straightforward to implement. For example, the projections of the current estimates of $o(\vec{x})$ and $h_m(\vec{x})$ onto $C_{\text{pos}}[u(\vec{x})]$ are carried out by zeroing out the negative values in these estimates. Similarly, the projections of $o(\vec{x})$ and $h_m(\vec{x})$ onto the convex sets that correspond to their support constraints are carried out by zeroing out all their values outside of their supports. In practice, some versions of PBBBD algorithms have been formulated to use minimization routines as part of the projection process in order to minimize the possibility of stagnation.

Deconvolution from Wavefront Sensing

A key distinguishing feature of post-detection compensation algorithms is the type of additional data required to estimate and remove atmospheric blurring in the measured short-exposure images. Blind deconvolution techniques need no additional data. Speckle imaging techniques estimate average atmospheric blurring quantities and need an estimate of the average atmospheric

energy spectrum. Because average quantities are calculated when using speckle imaging techniques, only an estimate of the average atmospheric energy spectrum is needed, not the specific energy spectra that corrupted the measured short-exposure images. As a result, the additional data need not be collected simultaneously with the short-exposure images. The technique described in this section, deconvolution from wavefront sensing (DWFS), seeks to deconvolve the atmospheric blurring in each short-exposure image by estimating the specific atmospheric field phase perturbations that blur each short-exposure image. Thus, the additional data must be collected simultaneously with the short-exposure images. The atmospheric field phase perturbations are calculated from data obtained with a wavefront sensor using a different region of the optical spectrum than used for imaging so that no light is taken from the imaging sensor. Most commonly, the wavefront sensor data consists of the first derivatives of the phasefront (phase differences) as obtained with Shack–Hartmann or shearing interferometer wavefront sensors, but curvature sensing wavefront sensors are also used that produce data that are proportional to the second derivatives of the phasefront.

The original algorithm to implement the DWFS technique produced an estimate $\hat{O}(\vec{f})$ of the Fourier transform $O(\vec{f})$ of the object being imaged using the following estimator:

$$\hat{O}(\vec{f}) = \frac{\sum_m I_m(\vec{f}) \tilde{H}_m^*(\vec{f})}{\sum_m |\tilde{H}_m(\vec{f})|^2} \quad (13)$$

where $\tilde{H}_m(\vec{f})$ is the overall system transfer function (including the telescope and atmospheric effects) for the m th short-exposure image estimated from the wavefront sensor data. To create $\tilde{H}_m(\vec{f})$ the atmospheric field phase perturbations corrupting the m th image must be calculated from the wavefront sensor measurements. Then $\tilde{H}_m(\vec{f})$ is obtained using standard Fourier optics techniques. Although Eq. (13) is easily implemented given an existing system that includes a wavefront sensor, it suffers from the fact that it is a biased estimator. An unbiased version of Eq. (13) is given by

$$\hat{O}(\vec{f}) = \frac{\sum_m I_m(\vec{f}) \tilde{H}_m^*(\vec{f})}{\sum_n H_{n,\text{ref}}(\vec{f}) \tilde{H}_{n,\text{ref}}^*(\vec{f})} \quad (14)$$

where $\tilde{H}_{n,\text{ref}}(\vec{f})$ and $\tilde{H}_{n,\text{ref}}^*(\vec{f})$ are estimates of the overall system transfer function obtained from data collected on a reference star using wavefront sensor data and image-plane data, respectively. Thus, to obtain an unbiased estimator, two more datasets must be collected. The number of frames in the reference star dataset need not be the same as for the imaging dataset. As for speckle imaging, the atmospheric statistics must be the same for the reference star and imaging datasets.

Another technique that uses at least one additional data channel along with the measured image is called phase diversity. In the additional data channels, images are also recorded, but with known field phase aberrations intentionally introduced to add known diversity to the unknown field phase corrupting the measurements. The most commonly introduced field phases aberration in an additional channel is defocus due to its ease of implementation. In the case of defocus, the additional data channel along with the imaging channel can be viewed as samples of the focal volume of the imaging system. Both the noise-free object and the unknown field phase aberrations can be recovered from the measured data using optimization techniques to enforce data consistency along with other prior knowledge constraints. These types of systems can use higher-order aberrations as well. The phase diversity concept has been generalized to an approach known as computational imaging, where an imaging system is designed by taking into account its intended use, the expected image degradation mechanisms, and the fact that the image will be the product of both an optical system and a postprocessing algorithm.

Pre-Detection Compensation

The primary pre-detection-only compensation technique in use today employs dynamical electro-optical systems to sense and remove the phases in the impinging wavefront due to atmospheric turbulence. This technique, known as adaptive optics, provides much higher-quality data than post-detection compensation techniques, when there is enough light from the object being imaged or a guidestar (artificial or natural) to drive the adaptive optics system at a bandwidth high enough to keep up with the temporal changes in atmospheric turbulence. It consists of a wavefront sensor that measures the portion of the phasefront in the pupil of the telescope that is due to atmospheric turbulence, a wavefront reconstructor that generates the deformable mirror drive signals from the wavefront sensor data, and a deformable mirror that applies the conjugate of the atmospheric phase distortions to a reimaged pupil. Some of the light reflected off the deformable mirror is then sent to the imaging sensor. Here, it suffices to say that no adaptive optics system produces imagery that is diffraction limited in quality, although, under the right conditions, the images can be near diffraction limited. However, atmospheric seeing conditions fluctuate significantly over a night as well as seasonally. Therefore, virtually all adaptive optics images benefit from applying post-compensation techniques, as is discussed next.

Hybrid Techniques

Adaptive Optics and Post-Processing

Often, an adaptive optics system does not fully correct for atmospheric turbulence. This can occur when atmospheric conditions become worse than the system is designed to handle. Also, an adaptive optics system may be intentionally designed to only

partially correct for atmospheric turbulence in order to lower the cost of the system. As a result, there still may be significant atmospheric blurring in systems that use adaptive optics. For this reason, additional post-processing of adaptive optics imagery can result in significant improvement in image quality. All of the techniques discussed in the section on post-detection compensation can be applied to adaptive optics imagery. Applying these techniques should never degrade image quality, and will improve the quality if the adaptive optics system doesn't fully correct for atmospheric turbulence. Generally speaking, the amount of improvement decreases as the performance of the adaptive optics system increases. Because the adaptive optics system's performance parameters can vary as atmospheric conditions vary, the blind deconvolution algorithms are especially applicable to processing adaptive optics imagery since they estimate the overall system transfer function from the imaging data. Because speckle imaging techniques require a separate estimate of the overall system transfer function, it can be difficult and time-consuming to obtain all of the additional data necessary to apply these techniques to adaptive optics data. Since these post-detection compensation techniques require short-exposure images, the adaptive optics system needs to have a camera capable of collecting short-exposure images. Finally, even if an adaptive optics system fully corrects for the atmosphere, deconvolution techniques that remove the diffraction-limited transfer function amplitude attenuation can improve image quality.

Partially Redundant Pupil Masking

Although adaptive optics systems can function very well, currently they are expensive and require significant maintenance. The primary benefit of adaptive optics systems is that they remove atmospheric turbulence phase distortions in the pupil of the telescope, thus permitting light from all areas of the pupil to interfere in phase when creating an image. An interferometric view of imaging reveals that multiple regions of the light in the pupil contribute to a single spatial frequency in the image. When the phasefront in the pupil is distorted by the atmosphere, these multiple regions tend to not interfere in phase, reducing the amplitude of the image's Fourier component at that spatial frequency. This amplitude reduction effect occurs for all spatial frequencies greater than the limit imposed by the atmosphere, r_o/λ . Therefore, techniques have been developed to modify the wavefront by masking part of the pupil prior to measuring the field with an imaging sensor. This class of techniques has various names including optical aperture synthesis, pupil masking, and aperture masking. Here this class of techniques will be referred to as partially redundant pupil masking (PRPM).

PRPM techniques function by masking the pupil in such a way as to minimize the amount of out-of-phase interference of light on the imaging detector. Early efforts used pupil masks that were opaque plates with r_o -sized holes in them. Because the atmospheric coherence length is r_o , all the light from an r_o -sized hole is essentially in phase. By choosing the holes in the mask so that only two holes contribute to a single spatial frequency in the image, out-of-phase interference problems are removed. This type of mask is called a nonredundant pupil mask. Of course, a significant amount of light is lost due to this masking, so only bright objects benefit from this particular implementation. Often, the pupil mask needs to be rotated in order to obtain enough spatial frequencies in the image to be able to reconstruct a high-quality estimate of the object. To increase the light levels in the image, masks have been made to allow some out-of-phase interference for a given spatial frequency, thus trading off the amplitude of the spatial frequencies allowed in the image with the amount of light in the image. Masks with this property are called partially redundant pupil masks.

One version of PRPM is called the variable geometry pupil (VGP) technique. It is envisioned to work in conjunction with an adaptive optics system that does not fully correct the wavefront for atmospheric phase distortions. It employs a spatial light modulator or a micro-electromechanical mirror between the deformable mirror and the imaging sensor. Either of these two devices is driven by signals from the wavefront sensor in the adaptive optics system in such a way that the portions of the pupil where the phasefront has a sufficiently large second derivative is blocked, while the remainder of the phasefront is transmitted.

All PRPM techniques benefit from post-compensation processing. For the case of nonredundant pupil masks, post-processing techniques are required to convert the discrete-spatial-frequency information into an image with reasonable quality. These techniques are essentially the same techniques as used in interferometric imaging, not those described in this article. When partially redundant pupil masks are used (including VGP masks), the image may include enough spatial frequency content that the methods described in the post-detection compensation section can be applied.

See also: Imaging Through Scattering Media

Further Reading

- Ayers, G.R., Northcott, M.J., Dainty, J.C., 1988. Knox–Thompson and triple-correlation imaging through atmospheric turbulence. *Journal of the Optical Society of America A* 5, 963–985.
- Bertero, M., Boccacci, P., 1998. *Introduction to Inverse Problems in Imaging*. Bristol, UK: Institute of Physics Publishing.
- Craig, I.J.D., Brown, J.C., 1986. *Inverse Problems in Astronomy*. Bristol, UK: Adam Hilger Ltd.
- Dainty, J.C., 1984. Stellar speckle interferometry. In: Dainty, J.C. (Ed.), *Topics in Applied Physics*, vol. 9: *Laser Speckle and Related Phenomena*, 2nd edn Berlin: Springer-Verlag, pp. 255–320.
- Goodman, J.W., 1985. *Statistical Optics*. New York: John Wiley and Sons.

- Haniff, C.A., Buscher, D., 1992. Diffraction-limited imaging with partially redundant masks I: Infrared imaging of bright objects. *Journal of the Optical Society of America A* 9, 203–218.
- Katsaggelos, A.K., Lay, K.T., 1991. Maximum likelihood identification and restoration of images using the expectation-maximization algorithm. In: Katsaggelos, A.K. (Ed.), *Digital Image Restoration*. Berlin: Springer-Verlag, pp. 143–176.
- Nakajima, T., Haniff, C.A., 1993. Partial adaptive compensation and passive interferometry with large ground-based telescopes. *Publications of the Astronomical Society of the Pacific* 105, 509–520.
- Roggemann, M.C., Welsh, B., 1996. *Imaging Through Turbulence*. Boca Raton, FL: CRC Press.
- Schulz, T.J., 1993. Multiframe blind deconvolution of astronomical images. *Journal of the Optical Society of America A* 10, 1064–1073.
- Tyson, R.K., 1991. *Principles of Adaptive Optics*. San Diego, CA: Academic Press.
- Youla, D.C., Webb, H., 1982. Image restoration by the method of convex projections: Part 1 – Theory. *IEEE Transactions on Medical Imaging* M1-1, 81–94.