HANDBOOK OF BIOMEDICAL OPTICS

EDITED BY

David A. Boas
Constantinos Pitris
Nimmi Ramanujam

CRC Press
Taylor & Francis Group

HANDBOOK OF
BIOMEDICAL OPTICS

HANDBOOK OF
BIOMEDICAL OPTICS

EDITED BY

David A. Boas
Constantinos Pitris
Nimmi Ramanujam

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

International Standard Book Number-13: 978-1-4200-9037-6 (eBook - PDF)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

This book is dedicated to Professor Britton Chance (1913–2010), a pioneer in biophotonics and a highly respected leader since the inception of this rapidly developing field of the biomedical sciences. Through his passion for research and education, he enriched the lives of generations of scientists. His legacy will live on through the numerous scientists who trained under his mentorship. This book commemorates his relentless spirit of innovation and discovery.

Contents

PART III Tomographic Imaging

PART IV Microscopic Imaging

Preface

There has been tremendous progress in medicine and biology over the past few decades. During this time, we have witnessed previously unimaginable advances in the understanding of human biology and physiology. One notable example is the human genome project, which has opened up new insights into disease diagnosis and therapy. Science and engineering have been instrumental in harnessing this new knowledge into clinically practical solutions. Within the field of medical imaging, the use of light plays an increasingly important role to reveal functional, structural, and molecular information nondestructively; also, interest in biomedical optics is growing rapidly. The use of light in medicine and biology offers great promise to deliver non- or minimally invasive diagnostics and targeted, customizable therapeutics more efficiently and safely. We are still far from the holy grail of "optical biopsies" and "optical therapeutics," but the strides we are making in biomedical optics foretell a future that may well deliver on the promise.

Biomedical optics is a broad discipline that covers the use of light in medicine and biology. It includes the use of light-based diagnostic methodologies as well as light-mediated therapeutics. Light can be used to both image microstructure and attain biochemical information such as the oxygenation of blood or the redox state of mitochondria. Optical imaging methodologies can cover from the microscopic to the macroscopic levels, with techniques including various forms of microscopy on the one hand and diffuse optical imaging on the other. In addition, diagnostically useful biochemical and functional information can be collected using spectroscopic techniques or molecular imaging. Laser surgery is now standard practice in certain medical disciplines, and photodynamic therapy offers treatment alternatives. More recently, tissue engineering and nanotechnology have entered the realm of biomedical optics, offering novel and potentially powerful possibilities both in the formation (e.g., with two-photon techniques) as well as the monitoring of tissue constructs.

The breadth and depth of the knowledge in biomedical optics has been expanding continuously and exponentially. We receive this progress with excitement; at the same time, it exemplifies the daunting amount of multidisciplinary information that is being amassed and the lack of a single source to serve as a reference and teaching tool for scientists in related fields. With this handbook, we aim to provide a comprehensive and detailed source of knowledge to serve as a research and teaching reference for both graduate and undergraduate students.

This handbook is organized into six parts: Background (Part I), Spectroscopy and Spectral Imaging (Part II), Tomographic Imaging (Part III), Microscopic Imaging (Part IV), Molecular Probe Development (Part V), and Phototherapy (Part VI). Part I provides introductory material on optics as well as the optical properties of tissue. Part II describes the various forms of spectroscopy and its application in medicine and biology, including methods that exploit intrinsic absorption and scattering contrast, dynamic contrast, as well as fluorescence and Raman contrast mechanisms. Tomographic imaging enables the formation of three-dimensional images despite the strong blurring of the highly scattered light within the tissue. Part III provides extensive coverage of tomography from the microscopic with optical coherence tomography to the macroscopic with diffuse optical tomography and photoacoustic tomography. Part IV discusses both conventional and cutting-edge technologies and their translation to biomedical applications in the basic sciences and clinical studies. Molecular imaging holds great promise for providing exquisite sensitivity to disease-specific markers. Its success is dependent on the development of exogenous agents that are detailed in Part V. Finally, Part VI provides examples of how light is being used to treat disease and injury.

We are indebted to the hard work of all of the authors to help create this comprehensive and cohesive handbook. We have learned a great deal about the process of compiling a handbook and thank Luna Han for all of her assistance and patience. We would also like to thank Christy Wanyo for her tremendous assistance at the end with handling the daunting tasks of formatting the chapters. Prof. Ramanujam would also like to thank Ms. Marlee Junker for managing many of the day to day tasks associated with compiling, reviewing, and handling portions of this handbook.

For MATLAB® and Simulink® product information, please contact:

The MathWorks, Inc.
3 Apple Hill Drive
Natick, MA, 01760-2098 USA
Tel: 508-647-7000
Fax: 508-647-7001
E-mail: info@mathworks.com
Web: www.mathworks.com

Editors

Dr. David A. Boas is an associate professor at the Harvard Medical School and associate physicist at Massachusetts General Hospital in Boston, Massachusetts. He received his bachelor's degree in physics from Rensselaer Polytechnic Institute, Troy, New York, in 1991 and his PhD in physics from the University of Pennsylvania, Philadelphia, Pennsylvania. His research interests include the following: photon migration in highly scattering media with an emphasis on diffuse optical tomography, clinical applications of diffuse optical tomography in brain and breast radiology, and fundamental studies of brain function and stroke using diffuse optical tomography and optical microscopy. Dr. Boas has been an associate editor of *Optics Express* and a guest editor of *Medical Physics* and the *Journal of Biomedical Optics*. He is a member of SPIE and the Optical Society of America (OSA) and has served as conference program chair for various OSA topical meetings.

Dr. Constantinos Pitris is an associate professor in the faculty of Electrical and Computer Engineering at the University of Cyprus, Nicosia, Cyprus. He completed his studies at the University of Texas at Austin (BS honors in electrical engineering, 1993, MS in electrical engineering, 1995), Massachusetts Institute of Technology (PhD in electrical and medical engineering, 2000), and Harvard Medical School (MD magna cum laude in medicine, 2002). He has worked as a research assistant at the University of Texas and Massachusetts Institute of Technology, and as a postdoctoral associate at the Wellman Laboratories of Photomedicine of the Massachusetts General Hospital and Harvard Medical School. His main research interests cover the areas of optics and biomedical imaging. The goal of this research is the introduction of new technologies in clinical applications for the improvement of diagnostic and therapeutic options. He is an active member of the OSA and a reviewer for *Optics Letters*, *Applied Optics*, and *Biomedical Optics*.

Dr. Nimmi Ramanujam is professor of biomedical engineering at Duke University. She received her PhD in biomedical engineering from the University of Texas, Austin, Texas, in 1995 and trained as an NIH postdoctoral fellow at the University of Pennsylvania from 1996 to 1999. Prior to her tenure at Duke, she was an assistant professor in the Department of Biophysics and Biochemistry at the University of Pennsylvania from 1999 to 2000 and in the Department of Biomedical Engineering at the University of Wisconsin, Madison, Wisconsin, from 2000 to 2005. Dr. Ramanujam's interests in the field of biophotonics are centered on research and technology development for applications to cancer. She is developing novel quantitative optical sensing and imaging tools for applications to breast, cervical, and head and neck cancers. She has been leading a multidisciplinary effort to translate these technologies into cancer patients. Dr. Ramanujam is a fellow of the Optical Society of America and is a member of the Department of Defense (DOD) Breast Cancer Research Program (BCRP) Integration Panel (IP). She has received several awards for her work in cancer research and technology development, including the Stansell Distinguished Research Award from the Pratt School of Engineering at Duke University, Era of Hope Scholar awards from the DOD, a Global Indus Technovator award from MIT, and the MIT TR100 innovator award.

Contributors

Thomas Ach
Department of Ophthalmology
University Hospital Heidelberg
Heidelberg, Germany

Samuel Achilefu
Department of Radiology
Washington University School of Medicine
St. Louis, Missouri

Walter J. Akers
Department of Radiology
Washington University School of Medicine
St. Louis, Missouri

Carlo Amadeo Alonzo
Wellman Center for Photomedicine
Massachusetts General Hospital
Boston, Massachusetts

Roman Amberger
Kirchhoff-Institute for Physics
Heidelberg University
Heidelberg, Germany

Simon R. Arridge
Department of Computer Science
University College London
London, United Kingdom

Costas Balas
Department of Electronic and Computer Engineering
Technical University of Crete
Chania, Greece

Alexey N. Bashkatov
Institute of Optics and Biophotonics
Saratov State University
Saratov, Russia

Andrew J. Berger
The Institute of Optics
University of Rochester
Rochester, New York

Gerrit Best
Applied Optics and Information Processing
Kirchhoff-Institute for Physics
Heidelberg University
Heidelberg, Germany

Brett E. Bouma
Wellman Center for Photomedicine
Massachusetts General Hospital
Boston, Massachusetts

Evgenia Bousi
Department of Electrical and Computer Engineering
University of Cyprus
Nicosia, Cyprus

Alexander Brunner
Applied Optics and Information Processing
Kirchhoff-Institute for Physics
Heidelberg University
Heidelberg, Germany

Sean A. Burgess
Department of Biomedical Engineering
Columbia University
New York, New York

Theresa M. Busch
Division of Medical Physics
University of Pennsylvania School of Medicine
Philadelphia, Pennsylvania

Colin M. Carpenter
Radiation Oncology
Stanford University School of Medicine
Stanford, California

Keith Cengel
Division of Medical Physics
University of Pennsylvania School of Medicine
Philadelphia, Pennsylvania

Albert E. Cerussi
Beckman Laser Institute and Medical Clinic
University of California, Irvine
Irvine, California

Jae Won Cha
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts

Aaron C.-H. Chen
Wellman Center for Photomedicine
Massachusetts General Hospital
and
Graduate Medical Sciences
Boston University School of Medicine
Boston, Massachusetts

Yu Chen
Fischell Department of Bioengineering
University of Maryland
College Park, Maryland

Ran Cheng
Center for Biomedical Engineering
University of Kentucky
Lexington, Kentucky

Heejin Choi
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts

So-Hyun Chung
Beckman Laser Institute and Medical Clinic
University of California, Irvine
Irvine, California

Carol J. Cogswell
Department of Electrical, Computer, and Energy Engineering
University of Colorado
Boulder, Colorado

Christoph Cremer
Institute for Pharmacy and Molecular Biotechnology
Heidelberg University
Heidelberg, Germany
and
The Jackson Laboratory
Bar Harbor, Maine
and
Institute for Molecular Biophysics
University of Maine
Bar Harbor, Maine

Anjul M. Davis
Department of Biomedical Engineering
Duke University
Durham, North Carolina

Dan Davis
Department of Life Sciences
Imperial College London
London, United Kingdom

Scott C. Davis
Thayer School of Engineering
Dartmouth University
Hanover, New Hampshire

Christian Depeursinge
Advanced Photonics Laboratory
Federal Polytechnic School of Lausanne
Lausanne, Switzerland

Charles A. DiMarzio
Department of Electrical and Computer Engineering
Department of Mechanical and Industrial Engineering
Northeastern University
Boston, Massachusetts

Stefan Dithmar
Department of Ophthalmology
University Hospital Heidelberg
Heidelberg, Germany

Nicoleta M. Dragomir
Australian Institute of Physics
Victoria University
Melbourne, Victoria, Australia

Christopher Dunsby
Department of Physics
Imperial College London
London, United Kingdom

Turgut Durduran
Institute of Photonic Sciences
Barcelona, Spain

Amanda Durkin
Beckman Laser Institute and Medical Clinic
University of California, Irvine
Irvine, California

Daniel S. Elson
Department of Surgery and Cancer
Imperial College London
London, United Kingdom

George Epitropou
Department of Electronic and Computer Engineering
Technical University of Crete
Chania, Greece

Michael Feld
George R. Harrison Spectroscopy Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts

Jarod C. Finlay
Division of Medical Physics
University of Pennsylvania School of Medicine
Philadelphia, Pennsylvania

Paul M.W. French
Wellman Center for Photomedicine
Massachusetts General Hospital
and
Department of Dermatology
Harvard Medical School
Boston, Massachusetts

James G. Fujimoto
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts

Neil Galletly
Department of Histopathology
Imperial College London
London, United Kingdom

Malte C. Gather
Wellman Center for Photomedicine
Massachusetts General Hospital
Boston, Massachusetts

Elina A. Genina
Institute of Optics and Biophotonics
Saratov State University
Saratov, Russia

Adam Gibson
Department of Medical Physics and Bioengineering
University College London
London, United Kingdom

Brian D. Goldberg
Axsun Technologies
Billerica, Massachusetts

Robert N. Graf
Department of Biomedical Engineering
Duke University
Durham, North Carolina

David Grant
Department of Physics
Imperial College London
London, United Kingdom

Michael R. Hamblin
Wellman Center for Photomedicine
Massachusetts General Hospital
and
Department of Dermatology
Harvard Medical School
Boston, Massachusetts

and

Harvard–MIT Division of Health Sciences and Technology
Cambridge, Massachusetts

Rainer Heintzmann
King's College London
London, United Kingdom
and
Institute of Photonic Technology
Jena, Germany
and
University of Jena
Jena, Germany

Andreas H. Hielscher
Departments of Biomedical Engineering, Radiology and Electrical Engineering
Columbia University
New York, New York

Elizabeth M.C. Hillman
Departments of Biomedical Engineering and Radiology
Columbia University
New York, New York

Ying-Ying Huang
Wellman Center for Photomedicine
Massachusetts General Hospital
and
Department of Dermatology
Harvard Medical School
Boston, Massachusetts

and

Aesthetic and Plastic Center of Guangxi Medical University
Nanning, People's Republic of China

Irving Itzkan
Beth Israel Deaconess Medical Center
Harvard University
Boston, Massachusetts

Joseph A. Izatt
Department of Biomedical Engineering
Duke University
Durham, North Carolina

Steven L. Jacques
Departments of Dermatology and Biomedical Engineering
Oregon Health & Science University
Portland, Oregon

Pratixa Joshi
Department of Biomedical Engineering
The University of Texas at Austin
Austin, Texas

Jeon Woong Kang
Wellman Center for Photomedicine
Massachusetts General Hospital
Boston, Massachusetts

Gordon Kennedy
Department of Physics
Imperial College London
London, United Kingdom

Chulhong Kim
Department of Biomedical Engineering
Washington University in St. Louis
St. Louis, Missouri

Daekeun Kim
Department of Biomedical Engineering
Dankook University
Yongin-si, Gyeonggi-do, Republic of Korea

Hyun Keol Kim
Department of Biomedical Engineering
Columbia University
New York, New York

Sharon V. King
Boulder Nonlinear Systems
Lafayette, Colorado

Alexander D. Klose
Department of Radiology
Columbia University
New York, New York

Sunil Kumar
Department of Physics
Imperial College London
London, United Kingdom

Peter M.P. Lanigan
Institute of Chemical Biology
Imperial College London
London, United Kingdom

Timothy Larson
Department of Biomedical Engineering
The University of Texas at Austin
Austin, Texas

Paul Lemmer
Kirchhoff-Institute for Physics
Heidelberg University
Heidelberg, Germany

Changhui Li
Department of Biomedical Engineering
Washington University in St. Louis
St. Louis, Missouri

Anthony Magee
Department of Life Sciences
Imperial College London
London, United Kingdom

Pierre Marquet
Neuropsychiatric Center
Lausanne University Hospital
Lausanne, Switzerland

Sasha McGee
George R. Harrison Spcetroscopy Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts

James McGinty
Blackett Laboratory
Imperial College London
London, United Kingdom

Jelena Mirkovic
George R. Harrison Spectroscopy Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts

Ian Munro
Department of Physics
Imperial College London
London, United Kingdom

Mark Neil
Department of Physics
Imperial College London
London, United Kingdom

Vasilis Ntziachristos
Institute for Biological and Medical Imaging
Technical University of Munich
and
Helmholtz Zentrum München
Munich, Germany

Dylan Owen
Department of Physics
Imperial College London
London, United Kingdom

Christos Pappas
Department of Electronic and Computer Engineering
Technical University of Crete
Chania, Greece

Raghuveer Parthasarathy
Department of Physics and Materials Science Institute
The University of Oregon
Eugene, Oregon

Nicolas Pavillon
Advanced Photonics Laboratory
Federal Polytechnic School of Lausanne
Lausanne, Switzerland

Lev T. Perelman
Beth Israel Deaconess Medical Center
Harvard University
Boston, Massachusetts

Constantinos Pitris
Department of Electrical and Computer Engineering
University of Cyprus
Nicosia, Cyprus

Ting-Chung Poon
Department of Electrical and Computer Engineering
Virginia Polytechnic Institute and State University
Blacksburg, Virginia

Chrysanthe Preza
Department of Electrical and Computer Engineering
The University of Memphis
Memphis, Tennessee

Le Qiu
Beth Israel Deaconess Medical Center
Harvard University
Boston, Massachusetts

Milind Rajadhyaksha
Dermatology Service
Memorial Sloan-Kettering Cancer Center
New York, New York

Rebecca Rae Richards-Kortum
Department of Bioengineering
Rice University
Houston, Texas

Francisco E. Robles
Medical Physics Department
Duke University
Durham, North Carolina

Darren Roblyer
The Beckman Laser Institute and Medical Clinic
University of California, Irvine
Irvine, California

Giuliano Scarcelli
Wellman Center for Photomedicine
Massachusetts General Hospital
Boston, Massachusetts

John C. Schotland
Department of Bioengineering
University of Pennsylvania
Philadelphia, Pennsylvania

Richard A. Schwarz
Department of Bioengineering
Rice University
Houston, Texas

Martin Schweiger
Department of Computer Science
University College London
London, United Kingdom

Juliette Selb
Martinos Center for Biomedical Imaging
Massachusetts General Hospital
Charlestown, Massachusetts

Colin Sheppard
Division of Bioengineering
National University of Singapore
Singapore

Melissa C. Skala
Department of Biomedical Engineering
Duke University
Durham, North Carolina

Peter T.C. So
Departments of Mechanical Engineering and Biological Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts

Konstantin Sokolov
Department of Biomedical Engineering
The University of Texas at Austin
Austin, Texas

and

Department of Imaging Physics
The University of Texas MD Anderson Cancer Center
Houston, Texas

Subhadra Srinivasan
Thayer School of Engineering
Dartmouth College
Hanover, New Hampshire

Gordon Stamp
Department of Histopathology
Imperial College London
London, United Kingdom

Melissa J. Suter
Wellman Center for Photomedicine
Massachusetts General Hospital
Boston, Massachusetts

Clifford Talbot
Department of Physics
Imperial College London
London, United Kingdom

Wendy Tanamai
Beckman Laser Institute and Medical Clinic
University of California, Irvine
Irvine, California

Yuankai K. Tao
Department of Biomedical Engineering
Duke University
Durham, North Carolina

Guillermo J. Tearney
Wellman Center for Photomedicine
Massachusetts General Hospital
Boston, Massachusetts

Kort Travis
Department of Physics
The University of Texas at Austin
Austin, Texas

Bebhinn Treanor
Department of Life Sciences
Imperial College London
London, United Kingdom

Bruce J. Tromberg
Beckman Laser Institute and Medical Clinic
University of California, Irvine
Irvine, California

Valery V. Tuchin
Institute of Optics and Biophotonics
Saratov State University
and
Institute of Precise Mechanics and Control
Russian Academy of Sciences
Saratov, Russia

Tuan Vo-Dinh
Departments of Biomedical Engineering and Chemistry
Duke University
Durham, North Carolina

Hsin-Neng Wang
Departments of Biomedical Engineering and Chemistry
Duke University
Durham, North Carolina

Lihong V. Wang
Department of Biomedical Engineering
Washington University in St. Louis
St. Louis, Missouri

William C. Warger, II
Wellman Center for Photomedicine
Massachusetts General Hospital
Boston, Massachusetts

Adam Wax
Department of Biomedical Engineering and Medical Physics Program
Duke University
Durham, North Carolina

Arjun G. Yodh
Department of Physics and Astronomy
University of Pennsylvania
Philadelphia, Pennsylvania

Guoqiang Yu
Center for Biomedical Engineering
University of Kentucky
Lexington, Kentucky

Seok-Hyun Yun
Wellman Center for Photomedicine
Massachusetts General Hospital
Boston, Massachusetts

Chao Zhou
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts

Timothy C. Zhu
Division of Medical Physics
University of Pennsylvania School of Medicine
Philadelphia, Pennsylvania

I

Background

1

Geometrical Optics

Ting-Chung Poon
Virginia Polytechnic Institute and State University

Geometrical optics is the study of light without diffraction or interference and is based on *Fermat's principle*. We treat light as particles of energy traveling through space. These particles follow trajectories that are called *rays*. Hence, geometrical optics is often called *ray optics*. Fermat's principle is a concise statement that contains all the physical laws, such as the *law of reflection* and the *law of refraction*, in geometrical optics (Poon and Kim 2006).

1.1 Fermat's Principle

Fermat's principle states that the path of a light ray follows is an extremum in comparison to nearby paths. The extremum may be a minimum, a maximum, or stationary with respect to variations in the ray path. However, the extremum is usually a minimum. For a simple example, as shown in Figure 1.1, the shortest distance (the minimum distance) between two points A and B is along a straight line (solid line) in a *homogeneous medium*, i.e., in a medium with a constant *refractive index*, instead of taking the nearby dotted line. Since the speed of light in a homogeneous medium is constant, the time it takes for the ray to traverse the solid line must be minimum. Hence Fermat's principle is often stated as a *principle of least time*. Under this context, the light ray would follow that path for which the time taken is minimum. For a more complicated example, we show the derivation of the well-known Snell's law of refraction. In Figure 1.2, θ_i and θ_t are the angles of incidence and transmission, respectively. The angles are measured from the normal NN′ to the interface MM′, which separated media 1 and 2, characterized by refractive indexes n_i and n_t, respectively. The total time taken to transit from point A to B is given by

$$t(z) = \frac{\mathrm{AO}}{v_1} + \frac{\mathrm{OB}}{v_2} = \frac{\sqrt{h_1^2 + z^2}}{v_1} + \frac{\sqrt{h_1^2 + (d-z)^2}}{v_2}, \qquad (1.1)$$

where v_1 and v_2 are the light velocities in media 1 and 2, respectively. According to Fermat's principle, we are required to minimize the total time. In order to minimize $t(z)$, we set

$$\frac{\mathrm{d}t(z)}{\mathrm{d}z} = \frac{z}{v_1\sqrt{h_1^2 + z^2}} - \frac{d-z}{v_2\sqrt{h_1^2 + (d-z)^2}} = 0,$$

which gives

$$\frac{\sin\theta_i}{v_1} = \frac{\sin\theta_t}{v_2}.$$

Now, $v_1 = c/n_i$ and $v_2 = c/n_t$, where c is the speed of light particles in vacuum. Hence the above equation becomes

$$n_i \sin\theta_i = n_t \sin\theta_t, \qquad (1.2)$$

which is called *Snell's law of refraction*.

1.2 Matrix Method in Paraxial Optics

1.2.1 Ray Transfer Matrix

We now consider how matrices may be used to describe ray propagation through optical systems comprising, for instance, a succession of lenses all centered on the same axis called the *optical axis*. We take the optical axis along the z-axis, which is the general direction in which the rays travel. We also consider those rays, called *paraxial rays*, that lie only in the x–z plane and that are close to the z-axis. To be precise, paraxial rays are rays with angles of incidence, reflection, and refraction at an interface, satisfying the small-angle approximation in that $\tan\theta \approx \sin\theta \approx \theta$ and $\cos\theta \approx 1$, where angle θ is measured in radians. *Paraxial*

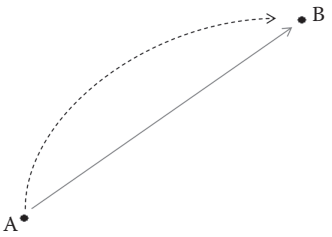

FIGURE 1.1 In a homogeneous medium, light ray takes the shortest distance, a straight line (solid line), between two points.

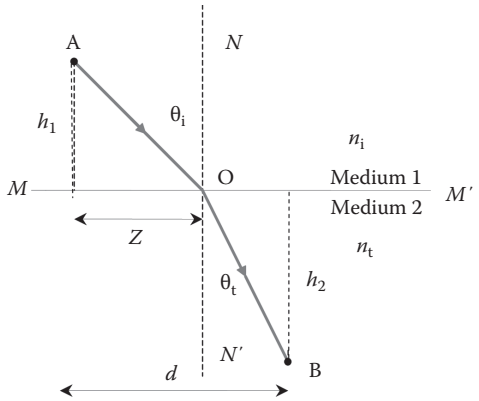

FIGURE 1.2 Law of refraction: incident (AO) and refracted (OB) rays.

optics deals with paraxial rays. Hence, in paraxial optics, Snell's law simplifies to

$$n_i \theta_i = n_t \theta_t. \tag{1.3}$$

We can now consider the propagation of a paraxial ray through an optical system shown in Figure 1.3, where a ray at a given x–z plane may be specified by its height x from the optical axis and by its angle θ or *slope*, which makes with the z-axis. The convention for the angle is anticlockwise positive measured from the z-axis. The height x of a point on a ray is taken positive if the point lies above the z-axis and negative if it is below the z-axis. The quantities (x, θ) represent the coordinates of the ray for a given z-plane. However, it is customary to replace the corresponding angle θ by $v = n\theta$, where n is the refractive index at the z-constant plane. Therefore, as shown in Figure 1.3, the ray at $z = z_1$ passes through the input plane with input ray coordinates $(x_1, n_1\theta_1)$ or (x_1, v_1). After the ray has gone through the optical system, the output ray coordinates at $z = z_2$ are $(x_2, n_2\theta_2)$ or (x_2, v_2). In the matrix formalism of paraxial optics, we relate the input coordinates to the output coordinates by a 2×2 matrix as follows:

$$\begin{pmatrix} x_2 \\ v_2 \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x_1 \\ v_1 \end{pmatrix}. \tag{1.4}$$

The above ABCD matrix is called the *ray transfer matrix*, which can be made up of many matrices to account for the effects of ray passing through various optical elements such as lenses. To write out Equation 1.4, we have

$$x_2 = Ax_1 + Bv_1 \tag{1.5a}$$

and

$$v_2 = Cx_1 + Dv_1. \tag{1.5b}$$

As an example, let us formulate Snell's law in terms of the matrix formulism. From Equation 1.3, we can write it as

$$\begin{pmatrix} x_2 \\ v_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ v_1 \end{pmatrix}, \tag{1.6}$$

where $v_2 = n_2\theta_t$ and $v_1 = n_1\theta_i$. Hence the ABCD matrix for Snell's law is $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, and Figure 1.4 summarizes the matrix for mulism for Snell's law. Note that the input and output planes in this case are the same plane at $z = z_1 = z_2$ and $x_1 = x_2$ as the heights of the input and output rays are the same. Other useful matrices are summarized in Figure 1.5. In Figure 1.5a, we have the *translation matrix* **T**, which describes the ray undergoing a translation of distance d in a homogenous medium characterized by n, and the matrix equation is given as follows:

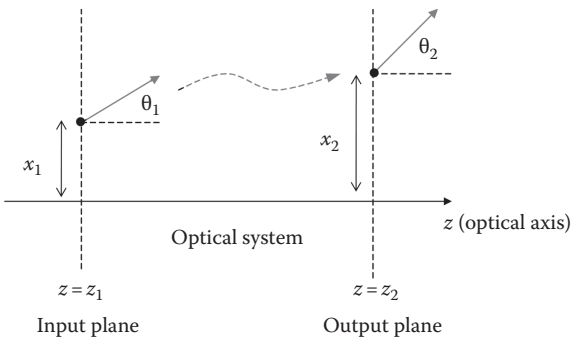

FIGURE 1.3 Input and output planes in an optical system.

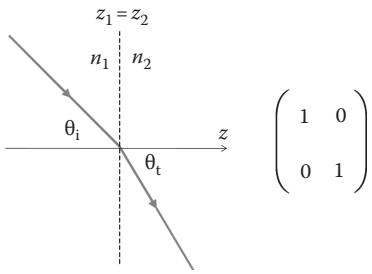

FIGURE 1.4 ABCD matrix for Snell's law.

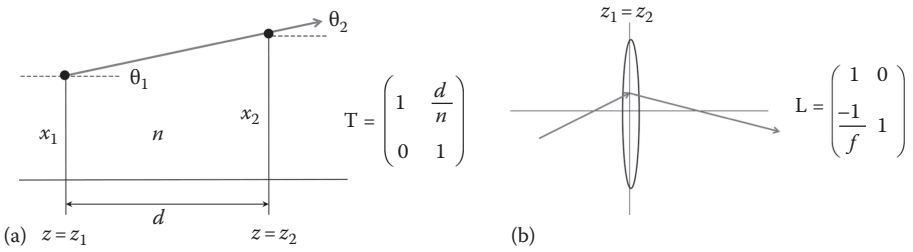

FIGURE 1.5 (a) ABCD matrix for ray translation in homogeneous medium: translation matrix **T**. (b) ABCD matrix for ray diffraction by thin lens of focal length *f*: lens matrix *L*.

$$\begin{pmatrix} x_2 \\ v_2 \end{pmatrix} = \begin{pmatrix} 1 & \frac{d}{n} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ v_1 \end{pmatrix} = \mathbf{T} \begin{pmatrix} x_1 \\ v_1 \end{pmatrix}, \tag{1.7a}$$

where $v_2 = n\theta_2$ and $v_1 = n\theta_1$. Hence the translation matrix for ray propagating a distance d in a homogenous medium of n is

$$\mathbf{T} = \begin{pmatrix} 1 & \frac{d}{n} \\ 0 & 1 \end{pmatrix}. \tag{1.7b}$$

Note that when the ray is undergoing translation in a homogeneous medium, $\theta_1 = \theta_2$, and therefore $v_1 = v_2$. When a thin converging lens of focal length f is involved, the matrix equation is

$$\begin{pmatrix} x_2 \\ v_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \frac{-1}{f} & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ v_1 \end{pmatrix} = \mathbf{L} \begin{pmatrix} x_1 \\ v_1 \end{pmatrix}, \tag{1.8a}$$

where **L** is the *thin-lens matrix* given by

$$\mathbf{L} = \begin{pmatrix} 1 & 0 \\ \frac{-1}{f} & 1 \end{pmatrix} \tag{1.8b}$$

Note that by definition, a lens is thin when the thickness of it is assumed to be zero, and hence $x_1 = x_2$, i.e., the input and output planes have become the same plane or $z_1 = z_2$, as shown in Figure 1.5b.

1.2.2 Ray Tracing through Thin Converging Lens

1.2.2.1 Input Rays Traveling Parallel to the Optical Axis

From Equation 1.8a, we recognize that $x_1 = x_2$ as the heights of the input and output rays are the same for the thin lens as illustrated in Figure 1.5b. Now, according to Equation 1.8a, $v_2 = -(1/f)x_1 + v_1$. For $v_1 = 0$, i.e., the input rays are parallel to the optical axis, $v_2 = -(1/f)x_1$. For positive x_1, $v_2 < 0$ as $f > 0$ for a converging lens. For negative x_1, $v_2 > 0$. Hence all input rays parallel to the optical axis converge behind the lens to the back focal point F (a distance of f away from the lens) of the lens as shown in Figure 1.6a. Note that for a thin lens, the front focal point is also a distance of f away from the lens.

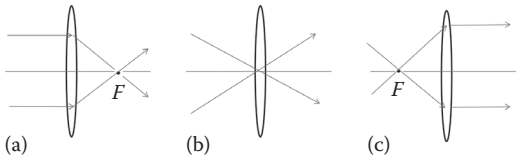

FIGURE 1.6 Ray tracing through a thin converging: (a) all parallel input rays converge to the back focal point *F*, (b) all input rays through the center of the lens pass undeviated, and (c) all input rays through the front focal point *F* give output rays parallel to the optics axis.

1.2.2.2 Input Rays Traveling through the Center of the Lens

For input rays traveling through the center of the lens, their input ray coordinates are $(x_1, v_1) = (0, v_1)$. The output ray coordinates, according to Equation 1.8a, are $(x_2, v_2) = (0, v_1)$ because $v_2 = v_1$, Hence we see all rays traveling through the center of the lens will pass undeviated as shown in Figure 1.6b.

1.2.2.3 Input Rays Passing through the Front Focal Point of the Lens

For this case, the input ray coordinates are $(x_1, x_1/f)$, and, according to Equation 1.8a, the output ray coordinates are $(x_2 = x_1, 0)$, which states that all output rays will be parallel to the optical axis ($v_2 = 0$), as shown in Figure 1.6c.

1.3 A Thin Converging Lens

1.3.1 Imaging

In this section, we show an example of using the transfer matrices, namely, the translation matrix **T** and the lens matrix **L**, to analyze the problem of a single thin lens. Figure 1.7 shows an object OO′ located at a distance d_0 in front of a thin lens of focal length *f*. We first construct a *ray diagram* for the imaging system, and the knowledge obtained from the last section should help

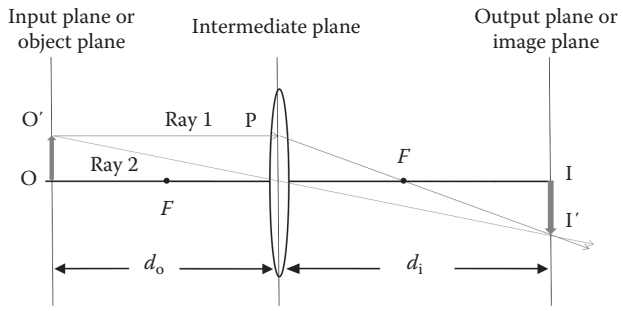

FIGURE 1.7 Imaging of a thin converging lens.

us to accomplish it. We send two rays from a point, say from O′, toward the lens. Ray 1 from O′ is incident parallel to the optical axis, and from part (a) from the last section, the input ray parallel to the optical axis converges behind the lens to the back focal point *F*. A second ray, i.e., ray 2 also from O′ may now be drawn through the center of the lens without bending—that is, the result from part (b) from the last section. The interception of the two rays on the other side of the lens will form an image point of O′. The image point of O′ is labeled at I′ in the diagram.

Now we investigate the imaging properties of the single thin lens using the matrix formalism. Let us consider the input plane, the immediate plane, and the output plane as shown in Figure 1.7. We also let (x_0, v_0), (x_p, v_p), and (x_i, v_i) represent the coordinates of the ray at O′, P (where P is on the immediate plane, and O′P defines the path where the parallel ray from O′ hitting the thin lens), and I′, respectively. We see that there are three matrices involved in the problem. From the input plane or the object plane to the immediate plane, it involves a translation matrix of distance d_0 (see Equation 1.7b, where we have assumed $n = 1$ for air):

$$\begin{pmatrix} x_p \\ v_p \end{pmatrix} = \begin{pmatrix} 1 & d_0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_0 \\ v_0 \end{pmatrix}, \tag{1.9}$$

and then when the incident ray to the lens is diffracted from the intermediate plane, we have

$$\begin{pmatrix} x'_p \\ x'_p \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \dfrac{-1}{f} & 1 \end{pmatrix} \begin{pmatrix} x_p \\ v_p \end{pmatrix}, \tag{1.10}$$

where (x_p, v_p) and (x'_p, v'_p) are the input ray and output ray coordinates due to the thin lens. Note that (x_p, v_p) and (x'_p, v'_p) are on the same plane—the intermediate plane as the lens is thin. Now finally, the ray exiting from the lens will translate for a distance d_i to reach the final output plane (or the image plane):

$$\begin{pmatrix} x_i \\ v_i \end{pmatrix} = \begin{pmatrix} 1 & d_i \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x'_p \\ v'_p \end{pmatrix}. \tag{1.11}$$

By substituting Equation 1.9 into Equation 1.10 and subsequently into Equation 1.11, we can relate the input coordinates (x_0, v_0) on the object plane to the output coordinates (x_i, v_i) on the image plane of the whole system as follows:

$$\begin{pmatrix} x_i \\ v_i \end{pmatrix} = \begin{pmatrix} 1 & d_i \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \dfrac{-1}{f} & 1 \end{pmatrix} \begin{pmatrix} 1 & d_0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_0 \\ v_0 \end{pmatrix} = \mathbf{S} \begin{pmatrix} x_0 \\ v_0 \end{pmatrix}, \tag{1.12a}$$

where

$$\mathbf{S} = \begin{pmatrix} 1 & d_i \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \dfrac{-1}{f} & 1 \end{pmatrix} \begin{pmatrix} 1 & d_0 \\ 0 & 1 \end{pmatrix} = \mathbf{T}_2\mathbf{L}\mathbf{T}_1 \tag{1.12b}$$

is called the *system matrix* of the entire imaging system. The overall system matrix **S** is expressed in terms of the product of three matrices $\mathbf{T}_2\mathbf{L}\mathbf{T}_1$ written in order from right to left as the ray goes from left to right along the optical axis. Let **A** and **B** be the 2×2 matrices as follows:

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} e & f \\ g & h \end{pmatrix}.$$

Then, the matrix product **AB** is

$$\mathbf{AB} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} ae + bg & af + bh \\ ce + dg & cf + dh \end{pmatrix}. \tag{1.13}$$

According to the rule of matrix multiplication in Equation 1.13, Equation 1.12b can be simplified to

$$\mathbf{S} = \begin{pmatrix} 1 - d_i/f & d_0 + d_i - d_0 d_i/f \\ -1/f & 1 - d_i/f \end{pmatrix}. \tag{1.14}$$

Hence Equation 1.12a becomes

$$\begin{pmatrix} x_i \\ v_i \end{pmatrix} = \begin{pmatrix} 1 - d_i/f & d_0 + d_i - d_0 d_i/f \\ -1/f & 1 - d_i/f \end{pmatrix} \begin{pmatrix} x_0 \\ v_0 \end{pmatrix} = \mathbf{S} \begin{pmatrix} x_0 \\ v_0 \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x_0 \\ v_0 \end{pmatrix}. \tag{1.15}$$

To investigate the conditions for imaging, let us concentrate on the ABCD matrix of **S** in the above equation. For imaging, the B element of **S** must be zero, which leads to $x_i = Ax_0 + Bv_0 = Ax_0$. This means that all rays passing through the input plane at the same object point x_0 will pass through the same image point x_i in the output plane—a condition of imaging. In addition, $A = x_i/x_0$ is the *lateral magnification* of the imaging system. Now,

in our case of thin-lens imaging, $B=0$ in Equation 1.15 leads to $d_0+d_i-d_0d_i/f=0$, which gives the *thin-lens formula*:

$$\frac{1}{d_0}+\frac{1}{d_i}=\frac{1}{f}. \tag{1.16}$$

The sign convention is that the object distance d_0 is positive (negative) if the object is to the left (right) of the lens. If the image distance d_i is positive (negative), the image is to the right (left) of the lens, and it is real (virtual). In Figure 1.7, we have $d_0>0$, $d_i>0$, and the image is therefore real, which means physically that light rays actually converge to the formed image.

Returning to Equation 1.15 with Equation 1.16, we have

$$\begin{pmatrix} x_i \\ v_i \end{pmatrix}=\begin{pmatrix} 1-d_i/f & 0 \\ -1/f & 1-d_i/f \end{pmatrix}\begin{pmatrix} x_0 \\ v_0 \end{pmatrix}, \tag{1.17}$$

which relates the input ray and output ray coordinates in the imaging system. The lateral magnification M of the imaging system, using Equations 1.16 and 1.17, is

$$M=A=\frac{x_i}{x_0}=\frac{1-d_i}{f}=-\frac{d_i}{d_0}. \tag{1.18}$$

If $M>0$, the image is erect and if $M<0$, the image is inverted. As shown in Figure 1.7, we have inverted image as both d_i and d_0 are positive.

When imaging of a volume, we need to consider longitudinal magnification as well. *Longitudinal magnification* M_z is the ratio of an image displacement along the axial direction δd_i to the corresponding object displacement δd_0: $M_z=\delta d_i/\delta d_0$. Using Equation 1.16 and treating d_i and d_0 as variables, we can take the derivative of d_i with respect to d_0 to obtain

$$M_z=\frac{\delta d_i}{\delta d_0}=-M^2. \tag{1.19}$$

Equation 1.19 states that the longitudinal magnification is equal to the square of the lateral magnification. The minus sign in front of the equation means that the decrease in the distance of the object from the lens $|d_0|$ will result in the increase in the image distance $|d_i|$. The situation of a magnified volume is shown in Figure 1.8, where a cube volume (abcd plus the dimension into the paper) is imaged into a truncated pyramid with a–b imaged into a′–b′ and c–d imaged into c′–d′.

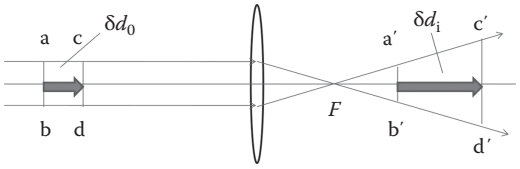

FIGURE 1.8 Longitudinal magnification: the image of a magnified volume is on the right side of the lens (side view).

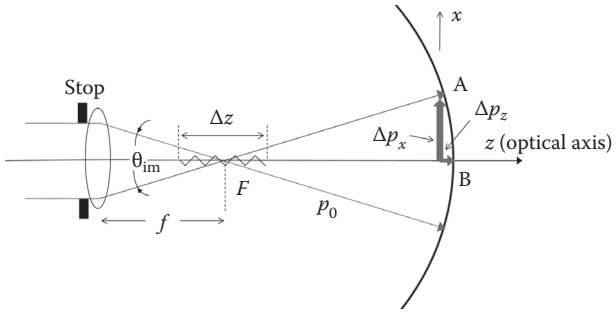

FIGURE 1.9 Uncertainty principle used to find resolution and depth of focus.

1.3.2 Numerical Aperture, Resolution, Depth of Focus, and Depth of Field

The *numerical aperture* NA of a lens is usually defined for an object or image located infinitely far away. It is a measure of the light-gathering ability. Figure 1.9 shows an object located at infinity, which sends rays parallel to the lens. The angle θ_{im} used to define the NA on the image side is

$$NA_i=n_i\sin\left(\frac{\theta_{im}}{2}\right) \tag{1.20}$$

where n_i is refractive index in the image space. Let us now find the lateral resolution, Δx. Since we treat light as particles in geometrical optics, each particle can then be characterized by its momentum, p_0. Quantum mechanics relates the minimum uncertainty in a position of quantum, Δx, to the uncertainty of its momentum, Δp_x, according to the relationship

$$\Delta x\Delta p_x\geq h, \tag{1.21}$$

where

h is *Planck's constant*

Δp_x is the momentum difference between rays FB and FA along the x-direction, i.e., the transverse direction as shown in Figure 1.9

The momentum of the FB ray along the x-axis is zero, while the momentum of the FA ray along the x-axis is $p_0\sin(\theta_{im}/2)$, where $p_0=h/\lambda_0$ with λ_0 being the wavelength in the medium, i.e., in the image space. Hence Δp_x is $\Delta p_x=p_0\sin(\theta_{im}/2)$. By substituting this into Equation 1.21, we have

$$\Delta x\geq\frac{h}{\Delta p_x}=\frac{h}{p_0\sin(\theta_{im}/2)}=\frac{\lambda_0}{\sin(\theta_{im}/2)}. \tag{1.22}$$

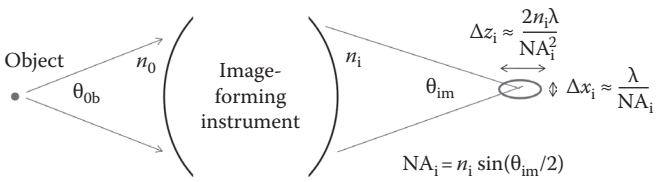

FIGURE 1.10 Image-forming instrument illustrating depth of focus and resolution in the image space.

Since the wavelength in the image space, λ_0 is equal to λ/n_i, where λ is the wavelength in air or in vacuum, Equation 1.22 becomes, using Equation 1.20,

$$\Delta x \geq \frac{\lambda}{n_i \sin(\theta_{im}/2)} = \frac{\lambda}{NA_i}. \tag{1.23}$$

Similarly, we can calculate the *depth-of-focus*, Δz, using $\Delta z \Delta p_z \geq h$, where Δp_z is the momentum difference between rays FB and FA along the z-direction, as shown in Figure 1.9. Details of the derivation have been provided (Poon 2007), and we state the final result as follows:

$$\Delta z \approx \frac{2n_i\lambda}{NA_i^2}. \tag{1.24}$$

To summarize the above results, Figure 1.10 shows an image-forming instrument with θ_{0b} and θ_{im} denoting the ray of maximum divergent angle and maximum convergent angle from the object side and the image side, respectively. If the resolutions in the object space are given by $\Delta x_0 \approx \lambda/NA_0$ and $\Delta z_0 \approx 2n_0\lambda/NA_0^2$, where $NA_0 = n_0 \sin(\theta_{ob}/2)$ is the NA and n_0 is the refractive index in the object space, and if the lateral magnification of the instrument is M, the resolutions on the image space are then $\Delta x_i \approx M\Delta x_0$ and $\Delta z_i \approx M^2\Delta z_0$. Take a 40×, $NA_0 \approx 0.6$ microscope objective as an example, we have $\Delta x_0 \approx 1\,\mu m$ for red light and Δz_0 in the object space is called *depth of field*, which is given by $\Delta z_0 \approx 2n_0\lambda/NA_0^2 \approx 3.3\,\mu m$ for $n_0 = 1$ in air. In the image space, the lateral resolution is $\Delta x_i \approx M\Delta x_0 \approx 40\,\mu m$, and the depth of focus is $\Delta z_i \approx M^2\Delta z_0 \approx 5.2\,mm$.

1.4 Magnifying Lens

The normal eye can form sharp images of objects as close to as about 250 mm away. The distance of 250 mm is known as the *near point* (NP) or *least distance of distinct vision* of the eye. For the eye to view objects closer than the NP, a magnifying lens can be used. The magnifying lens is simply a converging lens. Figure 1.11a shows that the unaided eye forms an image of an object located at the NP, and Figure 1.11b shows that a magnifying lens forms a virtual image, where the object is well within the NP of the eye, and we assume that the eye is close to the lens. Therefore, the magnifying lens makes an erect and magnified virtual image of the object. From Equation 1.16, we can write

$$d_i = \frac{f}{d_0 - f} d_0, \tag{1.25}$$

where

f is the focal length of the magnifying lens
d_i should be negative and greater in magnitude than d_0 as we recall $M = -d_i/d_0$

Note that, if $d_0 = f$, $d_i \to \infty$, the magnification M tends to be infinite. In practice, the paraxial approximation limits M to values about 10 for a single lens. In any case, the *magnifying power* or *angular magnification*, M_θ, is conventionally used for magnifying lenses or microscopes (Nelkon and Parker 1970):

$$M_\theta = \frac{\alpha}{\alpha_{NP}}, \tag{1.26}$$

where α_{NP} and α have been defined in Figure 1.11a and b. Figure 1.11c shows the situation when $d_0 = f$, where the image of the object appears coming from infinity. The eye is most comfortable or relaxed (called *zero accommodation*) to see distant objects, whereas the eye must achieve *maximum accommodation* to see an image located at its NP (see Figure 1.11a and b when $d_i = -250\,mm$].

Now by definition $\alpha_{NP} = h_0/250$, where h_0 is the size of the object as shown in Figure 1.11a for an unaided eye. Also, $\alpha = h_0/d_0$ from Figure 1.11b. Hence according to Equation 1.26, we have

$$M_\theta = \frac{h_0/d_0}{h_0/250} = \frac{250}{d_0}, \tag{1.27}$$

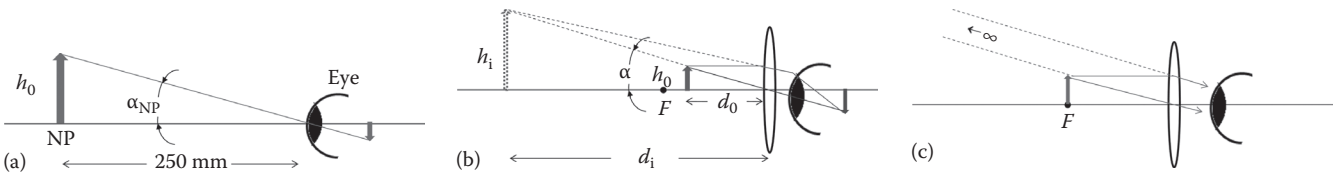

FIGURE 1.11 (a) Object at the NP of the eye, (b) object located within the focal length of the magnifying lens, and (c) object located at the front focal point F of the magnifying lens (eye relaxed as the eye receiving parallel rays).

where in Equation 1.27, d_0 is measured in millimeters. If the image is viewed at infinity, i.e., $d_0 = f$, the eye is relaxed, and we have

$$M_\theta = \frac{250}{f}, \tag{1.28a}$$

where the image is formed at infinity.

At the other extreme, if the image is viewed at the NP, we find $d_0 = fd_i/(d_i - f)$ from Equation 1.16. But $d_i = -250$, so we have $d_0 = -250f/(-250 - f)$. By substituting this quantity into Equation 1.27, we have

$$M_\theta = 1 + \frac{250}{f}, \tag{1.28b}$$

where the image formed at the near point of the eye.

This is the maximum value of the magnifying power that can be achieved with a single magnifying lens for the eye. In Equations 1.28a and b, $M_\theta > 0$ as $f > 0$ for a converging lens, which means that the image is always virtual and erect.

1.5 Compound Microscope

A compound microscope uses two converging lenses. The lens with short focal length f_0 facing the object is called the *objective*. The lens with focal length f_e in front of the eye is called the *eye-piece*. The objective makes a real, inverted, and magnified image of the object within the focal length of the eye-piece. The eye-piece then further magnifies the image from the objective by acting as a magnifying lens to eventually give a magnified, virtual image of the object. The situation is shown in Figure 1.12a. Let us now find the magnifying power of the compound microscope, M_θ^C, if the image is viewed at the NP of the eye. Again, using Equation 1.26, we have

$$M_\theta^C = \frac{\alpha_C}{\alpha_{NP}}, \tag{1.29}$$

where α_C is defined in Figure 1.12a. Let us further work on Equation 1.29. As usual, we assume that the eye is close to the eye-piece. The angle α_C subtended by the image of height h_2 is given by $\alpha_C = h_2/250$. With the unaided eye, the object of height h subtends an angle $\alpha_{NP} = h/250$. Therefore, we can rewrite Equation 1.29 as

$$M_\theta^C = \frac{\alpha_C}{\alpha_{NP}} = \frac{h_2/250}{h/250} = \frac{h_2}{h} = \frac{h_2/h_1}{h/h_1}. \tag{1.30}$$

The factor h_1/h is simply the lateral magnification of the objective, which is given by $M_0 = h_1/h = 1 - d_i/f_0$ (see Equation 1.18). As for h_2/h_1, we see that the angular magnification for the eye-piece by definition is $M_\theta^e = \alpha_C/\alpha_{NP} = \left(h_2/250\right)/\left(h_1/250\right) = h_2/h_1$, which is equal to $1 + (250/f_e)$ according to the results given by Equation 1.28b. Hence Equation 1.30 becomes

$$M_\theta^C = \frac{h_2/h_1}{h/h_1} = M_0 M_\theta^e = \left(1 - \frac{d_i}{f_0}\right)\left(1 + \frac{250}{f_e}\right) \tag{1.31}$$

for the image observed at the NP of the eye. Since $d_i > f_0$, M_θ^C is always negative (as $M_0 < 0$ and $M_\theta^e > 0$) to reflect that we observe a virtual and inverted final magnified image. From Equation 1.31, we recognize that the magnifying power of the compound microscope is conveniently expressed in terms of the product of the lateral magnification produced by the objective and the magnifying power of the eye-piece. Now if the final image is to be formed at infinity, the eye is unaccommodated or relaxed. In this case, we notice that the real image formed by the objective must be located at the focal point of the eyepiece. The situation is shown in Figure 1.12b. We, therefore, have

$$M_\theta^C = \frac{\alpha_\infty}{\alpha_{NP}} = \frac{h_1/f_e}{h/250} = \frac{h_1}{h}\frac{250}{f_e} = \left(1 - \frac{d_i}{f_0}\right)\frac{250}{f_e}. \tag{1.32}$$

Considering the objective only, the image distance must then be equal to $d_i = f_0 + L$, and the lateral magnification of the objective is subsequently given by

$$M_0 = 1 - \frac{d_i}{f_0} = \frac{f_0 - d_i}{f_0} = \frac{-L}{f_0}. \tag{1.33}$$

Incorporating this into Equation 1.32, the magnifying power of the compound microscope becomes

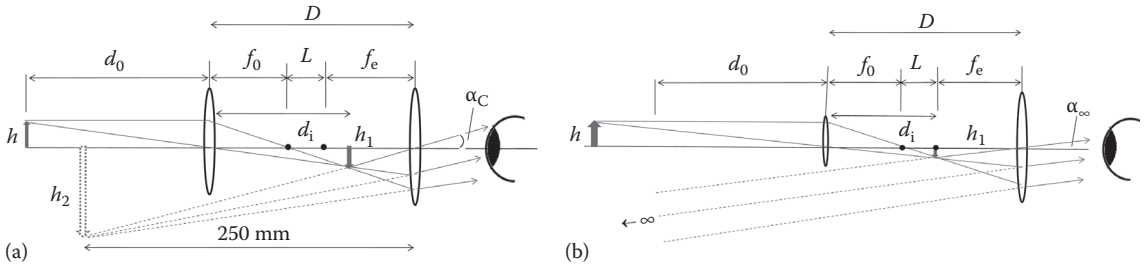

FIGURE 1.12 Compound microscope (a) image formed at the NP of the eye and (b) image formed at infinity (eye relaxed).

$$M_\theta = M_0 \frac{250}{f_e} = \frac{-L}{f_0}\frac{250}{f_e}. \quad (1.34)$$

L is known as the *tube length* and in most US-made microscopes, the standard is that $L = 160$ mm. Therefore, the lateral magnification stated on the barrel of an objective usually assumes 160 mm tube length. Otherwise, L will be stated on the barrel. In practice, the tube length is large compared to either f_0 or f_e.

References

Nelkon, M. and Parker, P. 1970. *Advanced Level Physics*. London: Heinemann Educational Books Ltd.

Poon, T.-C. 2007. *Optical Scanning Holography with MATLAB*. New York: Springer.

Poon, T.-C. and Kim, T. 2006. *Engineering Optics with MATLAB*. Hackensack, NJ: World Scientific.

2

Diffraction Optics

Colin Sheppard
National University of Singapore

2.1 Huygens' Principle

According to Huygens, each point on a wave front serves as the source of a spherical secondary wavelet with the same frequency as the primary wave. The amplitude at any point is the superposition of these wavelets. Note that Huygens' principle considers diffraction as a summation of spherical waves, not as a summation of plane waves, as we will consider in Section 2.7. This theory gives a simple qualitative description of diffraction but needs to be adapted to give good agreement with more exact formulations that will be shown later in Section 2.6.

Consider the propagation of a plane wave. Each point in the wave front can be considered as a source of secondary waves (Figure 2.1). This describes the propagating wave correctly, but suggests the possibility that the wave can equally well propagate backward. This is one reason the model has to be modified to agree with exact theories.

2.2 Fraunhofer and Fresnel Diffraction

Consider an opaque screen illuminated with a plane wave. The light spreads as a result of diffraction. If the observation screen is far enough away from the aperture, the diffraction pattern does not change in structure, but merely changes in size, as the distance is further increased. This situation is called the Fraunhofer diffraction (Figure 2.2). Closer to the aperture the diffraction pattern does change with distance. This is called Fresnel diffraction. Calculation of Fresnel diffraction is based on an approximation, which eventually breaks down: closer to the aperture more advanced theories are required. We shall discuss these regions later. The Fraunhofer diffraction pattern is obtained at a very large distance from the aperture, but using a lens, an image of it can be formed at a finite distance. In this case, the regions either closer or further from the lens give Fresnel diffraction.

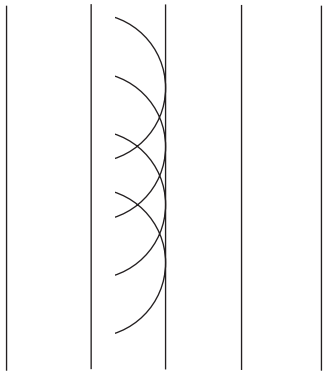

FIGURE 2.1 Huygens' principle for the propagation of a plane wave.

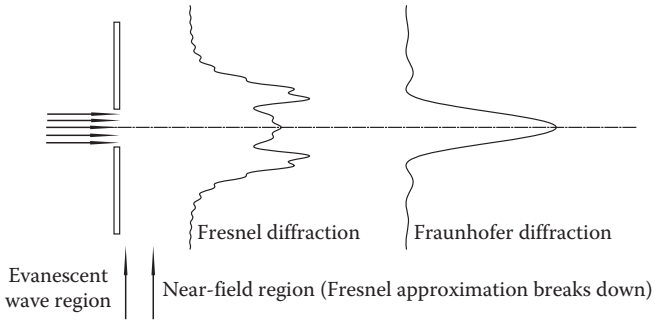

FIGURE 2.2 Regimes of diffraction.

2.3 Huygens' Diffraction Formula

According to Huygens' principle the amplitude at P is obtained by integrating the contribution from the points in the surface S. If the source and observation points are quite close to the axis, for unpolarized light, the field vectors can be represented by scalars because they are almost normal to the axis. We find that

$$U(P) = -\frac{i}{\lambda}\iint_S \frac{e^{ikr}}{r} U(Q)\,\mathrm{d}S. \tag{2.1}$$

This form of the equation is in a slightly different form compared with that of Hecht (1987), for example. $U(Q)$ gives the strength of the illumination at Q, e^{ikr}/r represents a spherical wave emanating from Q, and the factor $-i/\lambda$ results from the fact that Q is a driven dipole. The far-field of a dipole is $\pi/2$ out of phase with the forcing function, similar to the behavior of resonance.

This expression has been developed from a semiqualitative model: it is not strictly correct but gives a good prediction if, first, we are not too close to the aperture, and, second, if the aperture is large compared with the wavelength. In optics these two conditions are usually, but not always, true. The more rigorous Kirchhoff diffraction formula we will derive later, but usually this is not much of an improvement because it still does not take account of the fact that the illumination of the aperture is changed by the presence of the screen, and in addition vector effects are also neglected.

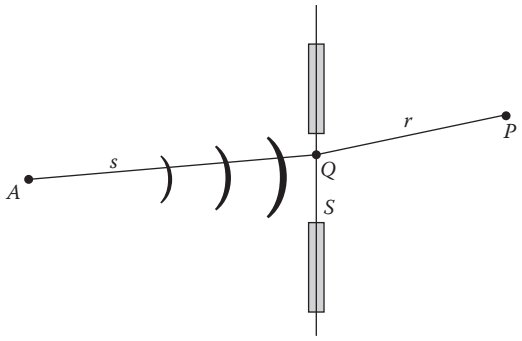

FIGURE 2.3 Geometry of diffraction.

If the aperture is illuminated with a spherical wave from a point a distance s away (Figure 2.3), then

$$U(Q) = A\frac{e^{iks}}{s},$$

and

$$U(P) = -A\frac{i}{\lambda}\iint_S \frac{e^{ik(r+s)}}{rs}\,\mathrm{d}S. \tag{2.2}$$

This is a form of the Huygens–Fresnel diffraction formula.

2.4 Fraunhofer Diffraction

In Equation 2.1, r does not change very much as Q varies over S. Although it is still necessary to take account of the changes in the exponent because this produces a multiplicative factor, in the denominator we can assume r is constant at a value r_0:

$$U_2(P) = -\frac{i}{\lambda r_0}\iint_S e^{ikr} U_1(Q)\,\mathrm{d}S. \tag{2.3}$$

Now, we have (Figure 2.4)

$$\begin{aligned} r^2 &= (x_2 - x_1)^2 + (y_2 - y_1)^2 + z^2 \\ &= r_0^2 + (x_1^2 + y_1^2) - 2(x_1x_2 + y_1y_2). \end{aligned} \tag{2.4}$$

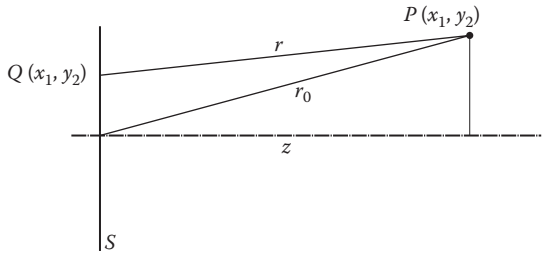

FIGURE 2.4 Geometry of the Fraunhofer diffraction.

Taking the square root of both sides, if $x_1, y_1 \ll r_0$, we can neglect the second term and expand the square root to give

$$r = r_0 - \frac{\left(x_1x_2 + y_1y_2\right)}{r_0}. \tag{2.5}$$

This is one form of the Fraunhofer approximation. Sometimes the expansion is made in terms of z, rather than r_0, but expansion in r_0 is valid for larger values of x_2, y_2. So for our case

$$U_2(P) = -\frac{i}{\lambda r_0}\exp(ikr_0)\iint_S U_1(x_1, y_1)\exp\left[-\frac{ik}{r_0}\left(x_1x_2 + y_1y_2\right)\right]\mathrm{d}x_1\mathrm{d}y_1. \tag{2.6}$$

The condition $x_1, y_1 << r_0$ can be written

$$r_0 >> \frac{k\left(x_1^2 + y_1^2\right)_{\max}}{2}. \tag{2.7}$$

Introducing the Fresnel number N,

$$N = \frac{\left(x_1^2 + y_1^2\right)_{\max}}{\lambda r_0}, \tag{2.8}$$

we have

$$N << \frac{1}{\pi}. \tag{2.9}$$

As an example, for $x_1^2 + y_1^2 = 1\,\mathrm{mm}^2$, that is the width is 2 mm, then for the Fraunhofer condition to be valid, $r_0 \gg 6$ m, which will not be the case in a laboratory experiment. But for $x_1 = 100\,\mu$m, $r_0 \gg 60$ mm, which we can observe easily.

If we define the Fourier transform of $f(x)$ as

$$F\left[f(m)\right] = \int_{-\infty}^{+\infty} f(x)e^{-2\pi imx}\mathrm{d}x, \tag{2.10}$$

Equation 2.6 is recognized as saying that

$$U_2(P) = \text{ const.} \times F\left[U_1(x_1, y_1)\right]. \tag{2.11}$$

In these expressions $U(x_1, y_1)$ is taken as the value of the incident field in the aperture, and zero outside of the aperture. This is called the Kirchhoff boundary condition.

2.4.1 Examples of Fraunhofer Diffraction—Single Slit

For a long slit, length $2l$, width $2a$, in the plane $y_2 = 0$ illuminated uniformly

$$\begin{aligned} U_2(P) &= -\frac{i}{\lambda r_0}\exp(ikr_0)\iint_S \exp\left(-\frac{2\pi ix_1x_2}{\lambda r_0}\right)\mathrm{d}x_1\mathrm{d}y_1 \\ &= -\frac{2\ell i}{\lambda r_0}\exp(ikr_0)\int_{-a}^{a}\exp\left(-\frac{2\pi ix_1x_2}{\lambda r_0}\right)\mathrm{d}x_1 \\ &= -\frac{2\ell i}{\lambda r_0}\exp(ikr_0)2a\left[\frac{\sin\left(2\pi ax_2/\lambda r_0\right)}{2\pi ax_2/\lambda r_0}\right], \end{aligned}$$

so

$$I = \frac{A^2}{\ell^2 r_0^2}\left[\frac{\sin(ka\sin\theta)}{ka\sin\theta}\right]^2, \tag{2.12}$$

where A is the area of the aperture.

2.4.2 Rectangular Aperture

Consider now a rectangular slit, sides $2a$ and $2b$, illuminated with a plane wave. The diffracted amplitude is

$$U(P) = -\frac{i}{\lambda r_0}\exp(ikr_0)\int_{-a}^{a}\exp\left(-\frac{2\pi ix_1x_2}{\lambda r_0}\right)\mathrm{d}x_1\int_{-b}^{b}\exp\left(-\frac{2\pi iy_1y_2}{\lambda r_0}\right)\mathrm{d}y_1. \tag{2.13}$$

so that the intensity is

$$I = \left(\frac{A^2}{\lambda^2 r_0^2}\right)^2\left[\frac{\sin\left(kax_2/r_0\right)}{\left(kax_2/r_2\right)}\right]^2\left[\frac{\sin\left(kby_2/r_0\right)}{\left(kby_2/r_2\right)}\right]^2. \tag{2.14}$$

There are zeros along a series of perpendicular lines in the diffraction pattern (Figure 2.5).

2.4.3 Circular Aperture

Consider diffraction by a circular aperture, radius a. Introducing polar coordinates (Figure 2.6)

$$\begin{aligned} x_1 &= R_1\cos\phi, \quad x_2 = R_2\cos\varphi, \\ y_1 &= R_1\sin\phi, \quad y_2 = R_2\sin\varphi, \end{aligned} \tag{2.15}$$

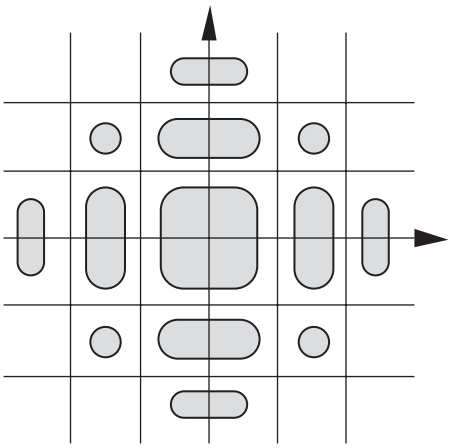

FIGURE 2.5 Fraunhofer diffraction by a rectangular aperture.

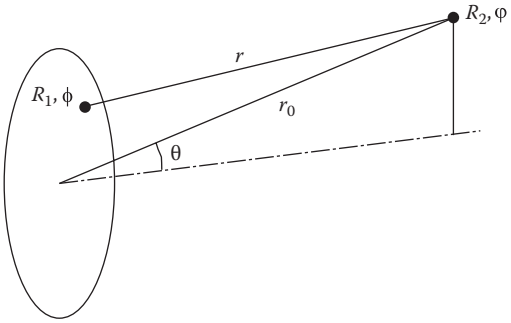

FIGURE 2.6 Diffraction by a circular aperture.

the diffracted field is

$$U_2(P) = -\frac{i}{\lambda r_0}\exp(ikr_0)\int_0^{2\pi}\int_0^{a}\exp\left[-\frac{ikR_1R_2\cos(\phi-\varphi)}{r_0}\right]R_1\mathrm{d}R_1\mathrm{d}\phi. \tag{2.16}$$

As the answer, by symmetry, is independent of φ, we can assume without loss of generality that $\varphi = 0$. The integral in ϕ is thus

$$\int_0^{2\pi}\exp\left(-\frac{ikR_1R_2\cos\phi}{r_0}\right)\mathrm{d}\phi. \tag{2.17}$$

But we know that

$$J_0(z) = \frac{1}{2\pi}\int_0^{2\pi}\exp(iz\cos\phi)\mathrm{d}\phi \tag{2.18}$$

as this is the definition of the Bessel function of the first kind of order zero, so

$$U_2(P) = -\frac{2\pi i}{\lambda r_0}\exp(ikr_0)\int_0^{a}J_0\left(\frac{kR_1R_2}{r_0}\right)R_1\,\mathrm{d}R_1. \tag{2.19}$$

This integral can be solved by use of the recurrence relationship (Abramowitz and Stegun 1965, Eq. 9.1.20)

$$\frac{1}{z}\frac{\mathrm{d}}{\mathrm{d}z}\left[zJ_1(z)\right] = J_0(z). \tag{2.20}$$

We have

$$\int J_0(z)z\,\mathrm{d}z = zJ_1(z), \tag{2.21}$$

so the diffracted amplitude is

$$\begin{aligned}U(R_2) &= -\frac{i}{\lambda r_0}\exp(ikr_0)\frac{r_0^2}{k^2R_2^2}\int_0^{kaR_0/r_0}J_0(\xi)\xi\,\mathrm{d}\xi\\ &= -\frac{i}{\lambda R_2^2}\left[\xi J_1(\xi)\right]_0^{kaR_2r_0}\\ &= \frac{i\pi a^2}{r_0\lambda}\exp(ikr_0)\left[\frac{2J_1(kaR_2/r_0)}{(kaR_2/r_0)}\right]\\ &= -\frac{iA}{\lambda r_0}\exp(ikr_0)\left[\frac{2J_1(ka\sin\theta)}{(ka\sin\theta)}\right],\end{aligned} \tag{2.22}$$

where A is the area of the aperture, or

$$I(\theta) = \left(\frac{A}{\lambda r_0}\right)^2\left[\frac{2J_1(ka\sin\theta)}{(ka\sin\theta)}\right]^2. \tag{2.23}$$

This is illustrated in Figure 2.7. Note that we have written the equation in this form as $2J_1(v)/v \to 1$ for $v \to 0$.

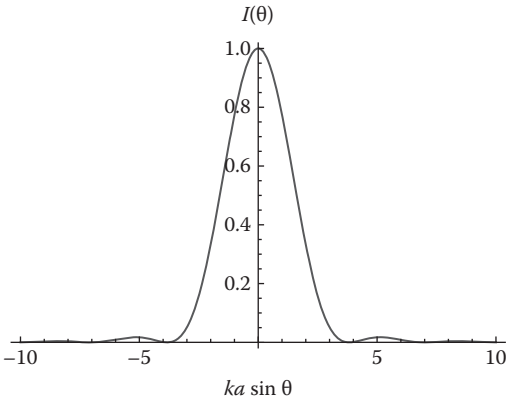

FIGURE 2.7 Fraunhofer diffraction by a circular aperture.

2.4.4 Annular Aperture

For an annular aperture (Figure 2.8), in Equation 2.6 we can put

$$U_1(R_1) = \mathrm{circ}\left(\frac{R_1}{a}\right) - \mathrm{circ}\left(\frac{R_1}{\varepsilon a}\right), \tag{2.24}$$

where $\mathrm{circ}(x) = 1,\ x < 1; = 0,\ x > 1$. Therefore

$$U(R_2) = -\frac{i\pi a^2}{\lambda r_0}\exp(ikr_0)\left\{\left[\frac{2J_1\left(kaR_2/r_0\right)}{\left(kaR_2/r_0\right)}\right] - \varepsilon^2\left[\frac{2J_1\left(k\varepsilon aR_2/r_0\right)}{\left(k\varepsilon aR_2/r_0\right)}\right]\right\}. \tag{2.25}$$

As the value of ε is increased (Figure 2.9), the central peak becomes narrower, but the side lobes become stronger. For the limiting case when $\varepsilon \to 1$,

$$U(R_2) = -\frac{i}{\lambda r_0}\exp(ikr_0)\int_0^\infty \delta(R_1 - a)J_0\left(\frac{2\pi R_1 R_2}{\lambda r_0}\right)2\pi R_1\,\mathrm{d}R_1$$
$$= -\frac{i}{\lambda r_0}\exp(ikr_0)J_0\left(\frac{2\pi R_2 a}{\lambda r_0}\right). \tag{2.26}$$

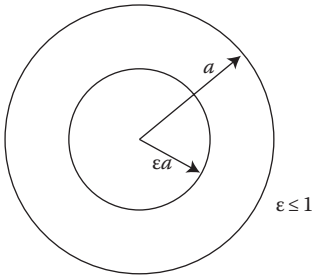

FIGURE 2.8 An annular aperture.

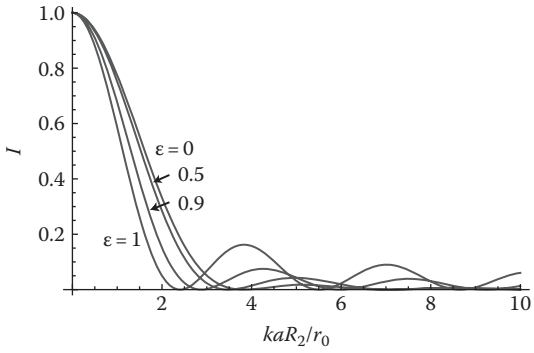

FIGURE 2.9 Normalized intensity in the Fraunhofer diffraction pattern of an annular aperture.

2.5 Fresnel Diffraction

2.5.1 Introduction

Knowing the field in a plane x_1, y_1 we can calculate the field in the plane x_2, y_2 using Equation 2.3.

Using the binomial theorem to expand the square root, Equation 2.5 now becomes (without neglecting the second term)

$$r = r_0 + \frac{x_1^2 + y_1^2}{2r_0} - \frac{x_1^2 x_1^2 + y_1 y_2}{r_0}. \tag{2.27}$$

So

$$U_2(P) = -\frac{i}{\lambda r_0}\exp(ikr_0)\iint_S U_1(x_1, y_1)\exp\left[-\frac{ik}{r_0}(x_1 x_2 + y_1 y_2)\right]$$
$$\times \exp\left[\frac{ik}{2r_0}\left(x_1^2 + y_1^2\right)\right]\mathrm{d}x_1\,\mathrm{d}y_1. \tag{2.28}$$

2.5.2 Circular Aperture

Consider a circular aperture illuminated with a plane wave. Using Equation 2.18 in Equation 2.28, we have

$$U_2(R_2, r_0) = -\frac{ik}{r_0}\exp(ikr_0)\int_0^a J_0\left(\frac{kR_1R_2}{r_0}\right)\exp\left(\frac{ikR_1^2}{2r_0}\right)R_1\,\mathrm{d}R_1. \tag{2.29}$$

Unfortunately, this integral cannot be solved in terms of elementary functions, though it can be solved in terms of Lommel functions (Born and Wolf 1975). This approach is not particularly useful, and it is usually easier to solve it numerically.

But, along the axis it reduces to a simple form:

$$U_2(0, r_0) = -\frac{ik}{r_0}\exp(ikr_0)\int_0^a \exp\left(\frac{ikR_1^2}{2r_0}\right)R_1\,\mathrm{d}R_1$$
$$= -\frac{ik}{r_0}\exp(ikr_0)\left[\exp\left(-\frac{ikR_1^2}{2r_0}\right)\right]_0^a$$
$$= -2i\exp(ikr_0)\exp\left(-\frac{ika^2}{4r_0}\right)\sin\left(\frac{ka^2}{4r_0}\right). \tag{2.30}$$

or

$$I_2(0, r_0) = 4\sin^2\left(\frac{ka^2}{4r_0}\right). \tag{2.31}$$

So we obtain a series of maxima and minima (zeros) in intensity along the axis (Figure 2.10). This can be explained in terms of

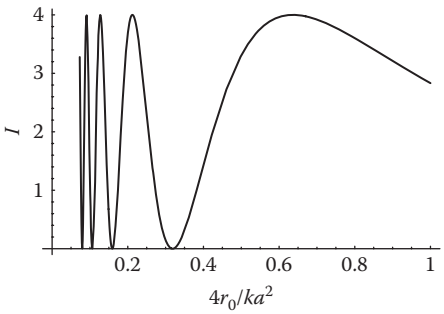

FIGURE 2.10 Intensity along the axis of a circular aperture according to the Fresnel diffraction theory.

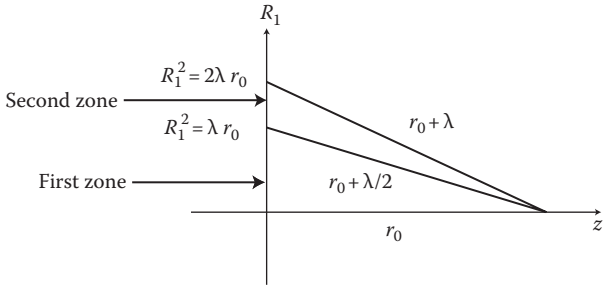

FIGURE 2.11 The Fresnel half-period zone construction.

Fresnel's zones in the following way. Contributions from successive zones (Figure 2.11) tend to cancel as the phases are different by 180°. If the number of zones is N, then $a^2 = n\lambda r$ and

$$N = \frac{a^2}{\lambda r}, \tag{2.32}$$

that is, it is equal to the Fresnel number, defined earlier (Equation 2.8).

Note that for a general axially symmetric U_1, along the axis, the amplitude is

$$U_2\left(0, r_0\right) = -\frac{ik}{r_0}\exp\left(ikr_0\right)\int_0^a U_1\left(R_1\right)\exp\left\{\frac{ikR_1^2}{2r_0}\right\}R_1\mathrm{d}R_1. \tag{2.33}$$

Let $R_1^2/a^2 = t$, and put $U_1(R_1) = V_1(t)$. We have

$$2R_1\mathrm{d}R_1 = a^2\ \mathrm{d}t,$$

and thus

$$U_2\left(0, r_0\right) = -i\pi N\exp\left(ikr_0\right)\int_0^1 V_1\left(t\right)\exp\left(i\pi Nt\right)\mathrm{d}t. \tag{2.34}$$

So the intensity is given by

$$I_2\left(N\right) = \left(\pi N\right)^2\left|FT\left[V_1\left(t\right)\right]\right|^2. \tag{2.35}$$

2.5.3 Rectangular Aperture

If a rectangular aperture, sides $2a$, $2b$, is illuminated by a plane wave, from Equation 2.28,

$$U_2\left(x_2, y_2\right) = -\frac{ie^{ikr_0}}{\lambda r_0}\int_{-a}^{+a}\exp\left\{\frac{ikx_1^2}{2r_0}\right\}\exp\left[-\frac{ik\left(x_1x_2\right)}{r_0}\right]\mathrm{d}x_1$$
$$\times\int_{-b}^{+b}\exp\left\{\frac{iky_1^2}{2r_0}\right\}\exp\left[-\frac{ik\left(y_1y_2\right)}{r_0}\right]\mathrm{d}y_1. \tag{2.36}$$

Now, consider the integral

$$I = \int_{-a}^{+a}\exp\left\{\frac{ikx_1^2}{2r_0}\right\}\exp\left[-\frac{ik\left(x_1x_2\right)}{r_0}\right]\mathrm{d}x_1$$
$$= \int_{-a}^{+a}\exp\left[\frac{ik\left(x_1^2 - 2x_1x_2\right)}{2r_0}\right]\mathrm{d}x_1. \tag{2.37}$$

By completing the square

$$I = \exp\left\{-\frac{ikx_2^2}{2r_0}\right\}\int_{-a}^{+a}\exp\left[\frac{ik\left(x_2 - x_1\right)^2}{2r_0}\right]\mathrm{d}x_1$$
$$= \sqrt{\frac{r_0\lambda}{2}}\exp\left\{-\frac{ikx_2^2}{2r_0}\right\}\int_{-w_1}^{+w_1}\exp\left\{\frac{i\pi w^2}{2}\right\}\mathrm{d}w, \tag{2.38}$$

where

$$w = \sqrt{\frac{2}{r_0\lambda}}\left(x_2 - x_1\right). \tag{2.39}$$

We now introduce the Fresnel integrals, defined as

$$C\left(w\right) = \int_0^w \cos\left\{\frac{\pi w'^2}{2}\right\}\mathrm{d}w'.$$

$$S\left(w\right) = \int_0^w \sin\left\{\frac{\pi w'^2}{2}\right\}\mathrm{d}w'.$$

$$F\left(w\right) = C + iS \tag{2.40}$$

So

$$I = 2\exp\left\{-\frac{ikx_2^2}{2r_0}\right\}\frac{F(w_2) - F(w_1)}{\sqrt{2/r_0\lambda}} \tag{2.41}$$

$$= 2\exp\left\{-\frac{ikx_2^2}{2r_0}\right\}\frac{F\left[\sqrt{\frac{2}{r_0\lambda}}\left(x_2 + a\right)\right] - F\left[\sqrt{\frac{2}{r_0\lambda}}\left(x_2 - a\right)\right]}{\sqrt{2/r_0\lambda}} \tag{2.42}$$

and thus

$$U_2(x_2, y_2) = -2i\exp(ikr_0)\exp\left\{\frac{ikR_2^2}{2r_0}\right\}\left[F(w_{x_2}) - F(w_{x_1})\right] \times \left[F(w_{y_2}) - F(w_{y_1})\right]. \quad (2.43)$$

The Fresnel integrals are tabulated. They have the following important properties:

$$\left.\begin{aligned} C(w) &= -C(-w) \\ S(w) &= -S(-w) \end{aligned}\right\} \text{odd function}, \quad (2.44)$$

$$C(0) = S(0) = 0, \quad F(0) = 0, \quad (2.45)$$

$$C(\infty) = S(\infty) = \frac{1}{2}, \quad F(-\infty) = -\frac{1}{2}(1-i), \quad (2.46)$$

$$C(-\infty) = S(-\infty) = -\frac{1}{2}, \quad F(-\infty) = -\frac{1}{2}(1-i). \quad (2.47)$$

Let us consider first what happens as the aperture becomes very large. Then from Equation 2.43, at $x_2 = y_2 = 0$

$$U_2(0, 0) = -\frac{i}{2}\exp(ikr_0)(1+i)^2 = \exp(ikr_0). \quad (2.48)$$

So it reproduces exactly what is expected for a plane wave! This is remarkable, because the assumptions we have made such as the Fresnel approximation and that $r \approx r_0$ break down for points on the aperture that are far from the axis. It works because the contributions from these off-axis points are weak because they are far away. So now it is established that the method can be applied even for an infinitely large aperture, giving us some confidence to look at some other cases.

2.5.4 Single Slit

Here $b \to \infty$ and we can take $y_2 = 0$ without loss of generality, so that

$$U_2(x_2) = \frac{(1-i)}{2}\exp(ikr_0)\exp\left\{-\frac{ikx_2^2}{2r_0}\right\} \times \left\{F\left[\sqrt{\frac{2}{r_0\lambda}}(x_2 + a)\right] - F\left[\sqrt{\frac{2}{r_0\lambda}}(x_2 - a)\right]\right\}. \quad (2.49)$$

For slits we get intensity distributions such as those shown in Figure 2.12.

Along the axis, $x_2 = 0$, and

$$U_2(0) = -(1+i)\exp(ikr_0)F\left\{\sqrt{\frac{2}{r_0\lambda}}a\right\}. \quad (2.50)$$

The intensity along the axis (Figure 2.13) should be compared with that for a circular aperture. It should be noted that the minima are not zeros this time. The maxima correspond approximately to the case when the Fresnel number $N = a^2/\lambda r_0$ is

$$N = 2n + \frac{3}{4}, \quad (2.51)$$

and the minima approximately to the case when

$$N = 2n + \frac{5}{4}, \quad (2.52)$$

where n is zero or a positive integer.

2.5.5 Half-Plane

For diffraction by an edge, we can put in Equation 2.42

$$w_2 = \sqrt{\frac{2}{r_0\lambda}}x_2, \quad w_1 \to \infty, \quad (2.53)$$

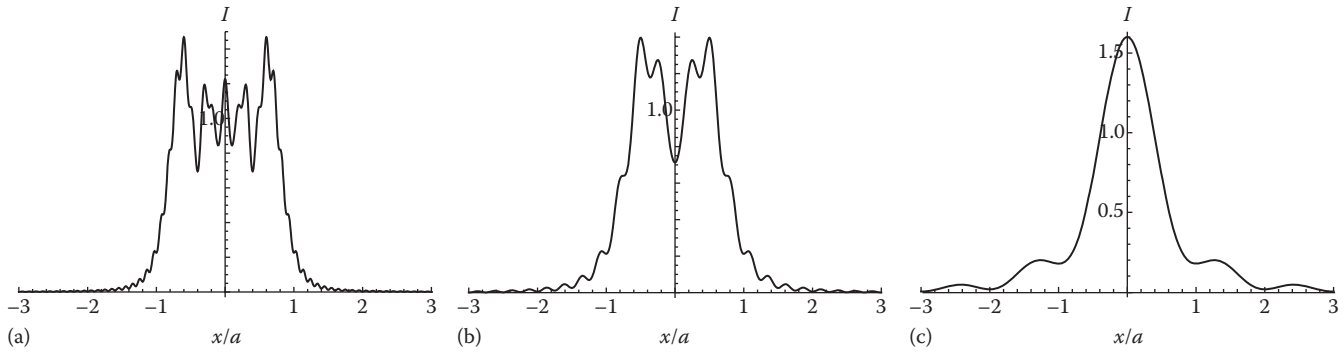

FIGURE 2.12 Fresnel diffraction by a slit: (a) $N = a^2/\lambda r_0 = 5$, (b) $N = 2$, and (c) $N = 0.5$.

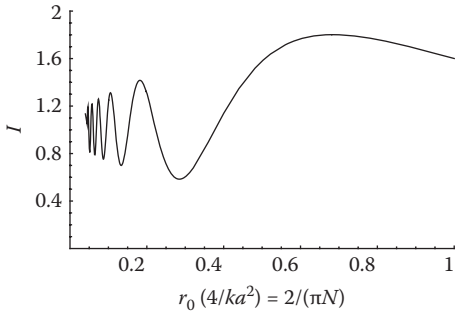

FIGURE 2.13 Intensity along the axis for a slit aperture according to Fresnel diffraction.

so that the intensity is

$$I = \frac{1}{2}\left\{\left[\frac{1}{2} - C\left(\sqrt{\frac{2}{r_0\lambda}}x_2\right)\right]^2 + \left[\frac{1}{2} - S\left(\sqrt{\frac{2}{r_0\lambda}}x_2\right)\right]^2\right\}. \tag{2.54}$$

Note that the diffraction pattern is the same at any distance, but of course it scales with distance (Figure 2.14). This is in contrast to the diffraction pattern for a slit, which changes with distance.

2.5.6 Circular Obstruction

Consider a circular obstruction, radius a. Then

$$\begin{aligned} U_2(0, r_0) &= \frac{ik}{r_0}\exp(ikr_0)\int_a^\infty \exp\left(-\frac{ikR_1^2}{2r_0}\right)R_1\,dR_1 \\ &= \frac{ik}{r_0}\exp(ikr_0)\int_0^\infty \exp\left(-\frac{ikR_1^2}{2r_0}\right)R_1\,dR_1 \\ &\quad - \frac{ik}{r_0}\exp(ikr_0)\int_0^a \exp\left(-\frac{ikR_1^2}{2r_0}\right)R_1\,dR_1. \end{aligned} \tag{2.55}$$

The first term represents a plane wave and the second the diffracted field for a circular aperture, rather than an obstruction.

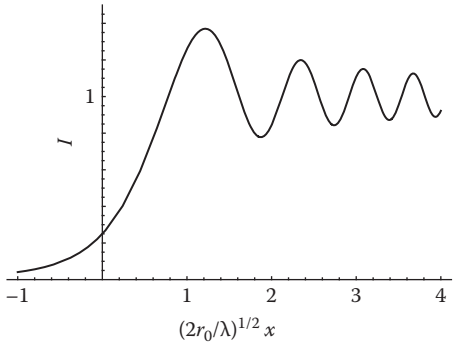

FIGURE 2.14 Fresnel diffraction by an edge.

This is a particular case of Babinet's principle, which states that the sum of the fields for two complementary screens is equal to the unobstructed disturbance.

The first term we evaluated in Equation 2.48: it gave just $\exp(ikr_0)$. So

$$\begin{aligned} U_2 &= -\exp(ikr_0) + \left\{\exp(ikr_0)\left[\exp\left(\frac{ika^2}{2r_0}\right) + 1\right]\right\} \\ &= \exp(ikr_0)\exp\left(\frac{ika^2}{2r_0}\right). \end{aligned} \tag{2.56}$$

The intensity is thus constant along the axis (the approximations break down when we get too close to the obstruction). This is the Poisson (or Arago) spot. It can be regarded as being caused by light scattered from the edge of the disc. Comparing with the case of the circular aperture, Equation 2.30 represents interference between the edge-diffracted wave and the undiffracted wave ($R_1 = 0$). This is the principle of the boundary diffraction wave concept introduction by Young.

2.6 Kirchhoff Diffraction Integral

We now return to the problem of deriving the diffraction integral starting from the wave equation

$$\left(\nabla^2 + k^2\right)U = 0. \tag{2.57}$$

Consider a closed surface S with inward normal $\mathbf{n}$ (Figure 2.15). Then we can show using Green's theorem that at any point inside S

$$\iint_S \left(U\frac{\partial U'}{\partial n} - U'\frac{\partial U}{\partial n}\right)dS = 0, \tag{2.58}$$

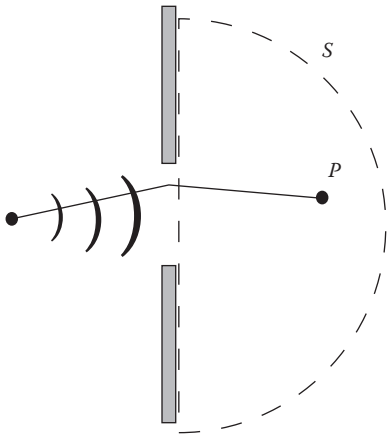

FIGURE 2.15 Diffraction according to Kirchhoff diffraction theory.

for any U' which also satisfies the wave equation. We can get different solutions by appropriate choice of the so-called Green function U' that satisfies the wave equation.

If we choose $U' = \exp(ikr)/r$, which satisfies the wave equation, and excluding the region around its singularity at $r = 0$, we obtain the Kirchhoff diffraction integral:

$$U_p = \frac{1}{4\pi}\iint\left[U\frac{\partial}{\partial n}\left(\frac{e^{ikr}}{r}\right) - \left(\frac{e^{ikr}}{r}\right)\frac{\partial U}{\partial n}\right]\mathrm{d}S. \tag{2.59}$$

It must be stressed that this expression is rigorously correct (in the scalar approximation).

Consider now diffraction by a planar screen with an aperture illuminated with a wave emanating from point S. We assume that the field in the aperture is the same as in the screen were absent. So in the aperture

$$U = \frac{U_0 e^{iks}}{s}, \quad \frac{\partial U}{\partial n} = \frac{U_0 e^{iks}}{s}\left(ik - \frac{1}{s}\right)\cos(n, s), \tag{2.60}$$

where $\cos(n,s)$ is the cosine of the angle between the n and s directions, whereas on the screen

$$U = 0, \quad \frac{\partial U}{\partial n} = 0. \tag{2.61}$$

These are called the Kirchhoff boundary conditions.

So, neglecting $1/s$ and $1/r$ in comparison to k,

$$U(P) = -\frac{iU_0}{\lambda}\iint_{\text{Aperture}}\frac{e^{ik(r+s)}}{rs}\left[\frac{\cos(n, \mathrm{s}) - \cos(n, r)}{2}\right]\mathrm{d}S. \tag{2.62}$$

Here we have assumed that the contribution from the spherical part of the surface S vanishes as its radius is made increasingly large. The factor in square brackets is called the obliquity factor. Apart from this, the expression is identical to Equation 2.22. Equation 2.62 is the Kirchhoff diffraction formula. Note that the surface of integration is arbitrary.

Although the Kirchhoff diffraction integral is exact, the assumed boundary conditions are not. So the Kirchhoff diffraction formula is also not exact. In particular it fails to reproduce the assumed field in the aperture. This is because the choice of U and $\partial\, U/\partial n$ in the aperture is inconsistent.

It can be shown that in fact, for the special case of integration over a planar region,

$$U_{\mathrm{p}} = \frac{1}{2\pi}\iint U\frac{\partial}{\partial n}\left(\frac{e^{ikr}}{r}\right)\mathrm{d}S \tag{2.63}$$

and

$$U_{\mathrm{p}} = -\frac{1}{2\pi}\iint\left(\frac{e^{ikr}}{r}\right)\frac{\partial U}{\partial n}\,\mathrm{d}S \tag{2.64}$$

These are called the Rayleigh–Sommerfeld I and II diffraction formulae. They have the advantage of being consistent as we need only assume U or $\partial U/\partial n$ in the aperture, but are not in practice any more accurate as we do not usually know either U or $\partial U/\partial n$ accurately. In fact, Equation 2.59 can be thought of as the average of Equations 2.63 and 2.64 and has been claimed to give a better prediction of the field at P.

2.7 Angular Spectrum of Plane Waves

Consider a single plane wave propagating at angle θ, for which

$$E = E_0\exp(-ikx\sin\theta)\exp(-ikz\cos\theta). \tag{2.65}$$

In the plane $z = 0$

$$E = E_0\exp(-ikx\sin\theta),$$

which is a harmonic variation with spatial wavelength $\lambda/\sin\theta$. So if we alter θ, we alter the spatial wavelength in the x direction. According to Fourier synthesis, *any* field in the plane $z = 0$ can be represented by a sum of Fourier components:

$$E_y(x) = \int_{-\infty}^{+\infty} G(\theta)\exp(-ikx\sin\theta)d(\sin\theta). \tag{2.66}$$

$G(\theta)$ is the strength of a plane wave component traveling in direction θ. Here $G(\theta)$ is *complex* to account for the relative phase of the components. As a simple example, for the field

$$\begin{aligned} E_y(x) &= A\cos\left(\frac{2\pi x}{\Lambda}\right) \\ &= \frac{A}{2}\exp\left(\frac{2\pi ix}{\Lambda}\right) + \frac{A}{2}\exp\left(-\frac{2\pi ix}{\Lambda}\right), \end{aligned} \tag{2.67}$$

we have immediately $\sin\theta = \pm\lambda/\Lambda$, representing two plane waves. In this case, the diffraction pattern of the cosine grating in the far field is two bright spots.

2.8 Evanescent Waves

Note that

$$E(x) = \int_{-\infty}^{+\infty} G(\theta)\exp(-ikx\sin\theta)d(\sin\theta) \tag{2.68}$$

has limits $\pm\infty$. But $|\sin\theta| \le 1$ for θ to be real. So $|\sin\theta| > 1$ corresponds to waves traveling at a complex angle. We have

$$k_x^2 + k_z^2 = k^2, \tag{2.69}$$

so that

$$k_z^2 = k^2 - k_x^2. \tag{2.70}$$

If $k_x > k$, we can put

$$k_z = \pm i\sqrt{k_x^2 - k^2},$$

which is imaginary. Therefore the electric field is

$$E = E_0 \exp(-ik_x x)\exp\left[\pm\left(z\sqrt{k_x^2 - k^2}\right)\right], \tag{2.71}$$

representing a wave traveling in the x direction, but with an exponential decay in the z direction. Note that we take the positive root to give a physical solution for $z \geq 0$.

This is an *evanescent* wave. The integral integrates over all propagating *and evanescent waves*. In the far field, the evanescent waves have decayed, so they make no contribution: only the propagating waves remain.

2.9 Diffraction by a Phase Screen

Suppose we have a screen that has an amplitude transmittance $t(x,y)$. This is complex to account for amplitude and phase effects. Then if it is illuminated by a plane wave, amplitude U_0 the field immediately after the screen is $U_0 t$.

A thin lens can be thought of as such a screen. It consists of an amplitude term $P(x_1, y_1)$ that is called the pupil function of the lens, which is unity in the aperture and zero outside, and a phase term $e^{i\Phi(x,y)}$.

$$t(x_1, y_1) = P(x_1, y_1)e^{i\Phi(x,y)}, \tag{2.72}$$

or in polar coordinates

$$t(r_1, \theta_1) = P(r_1, \theta_1)e^{i\Phi(r_1,\theta_1)}. \tag{2.73}$$

In general, we can expand Φ as a power series in r and also as a series in cos $n\theta$:

$$\begin{aligned}\Phi &= \sum_{n,m} a_{mn} r^m \cos n\theta \\ &= a_{00} + a_{20}r^2 + a_{40}r^4 + \cdots + a_{11}r\cos\theta \\ &\quad + a_{21}r^2\cos\theta + \cdots + a_{22}r^2\cos 2\theta + \cdots,\end{aligned} \tag{2.74}$$

or, collecting all the terms except the squared term,

$$\Phi = a_{20}r^2 + \Phi'(r, \theta). \tag{2.75}$$

So the amplitude transmittance of the lens can be written as

$$t(r_1, \theta_1) = P(r_1)\exp(ia_{20}r_1^2)e^{i\Phi'}. \tag{2.76}$$

After the lens, neglecting diffraction for the present,

$$\begin{aligned}U(r) &= P(r)\exp(ia_{20}r^2)e^{i\Phi'}e^{ikz} \\ &= P(r_1)e^{i\Phi'}\exp\left[ik\left(\frac{a_{20}r^2}{k} + z\right)\right].\end{aligned} \tag{2.77}$$

Neglecting Φ', which represents aberrations, the phase front through the origin is the paraboloid with equation (Figure 2.16)

$$\frac{a_{20}r^2}{k} + z = 0. \tag{2.78}$$

Consider a sphere, radius f, centered on $z = f$. Then

$$r^2 + (z - f)^2 = f^2,$$

and using the binomial theorem and assuming $r \ll f$

$$(z - f) = \sqrt{f^2 - r^2} = \pm f\left(1 - \frac{r^2}{2f}\right). \tag{2.79}$$

Taking the negative root and comparing Equations 2.78 and 2.79

$$\frac{a_{20}}{k} = -\frac{1}{2f}. \tag{2.80}$$

So the lens, including the aberration term, can be taken as

$$t(r_1, \theta_1) = P(r_1)\exp\left(-\frac{ikr_1^2}{2f}\right)e^{i\Phi'}, \tag{2.81}$$

where f is the focal length of the lens.

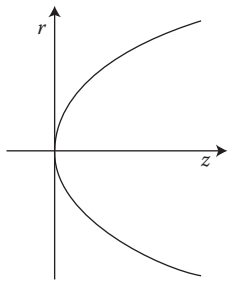

FIGURE 2.16 Phase front of a lens.

2.10 Thin Lens

For a thin lens, the field after the lens is (Goodman 1968)

$$U_2(x,y) = U_1(x,y)P(x,y)\exp\left[ik(n-1)\Delta(x,y)\right] \tag{2.82}$$

where

Δ is the thickness of the glass
n is its refractive index (Figure 2.17)

In the paraxial approximation, $x/R_1 \ll 1$ and higher powers can be neglected, so

$$\Delta(x,y) = \Delta_0 - \frac{x^2+y^2}{2}\left(\frac{1}{R_1} - \frac{1}{R_2}\right). \tag{2.83}$$

Thus,

$$U_2(x,y) = U_1(x,y)P(x,y)\exp\left[ik(n-1)\Delta_0\right] \times \exp\left[-ik(n-1)\left(\frac{x^2+y^2}{2}\right)\left(\frac{1}{R_1} - \frac{1}{R_2}\right)\right]. \tag{2.84}$$

Putting

$$\frac{1}{f} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right), \tag{2.85}$$

we then have

$$U_2(x,y) = U_1(x,y)P(x,y)\exp\left[ik(n-1)\Delta_0\right]\exp\left[-\frac{ik}{2f}\left(x^2+y^2\right)\right]. \tag{2.86}$$

The term $\exp[ik(n-1)\Delta_0]$ is a constant phase term, which we neglect.

2.11 Focus of a Lens

We assume the lens is illuminated by a plane wave so that

$$U_2(x,y) = P(x,y)\exp\left[-\frac{ik\left(x^2+y^2\right)}{2f}\right]. \tag{2.87}$$

This represents a spherical wave convergent on the point F (Figure 2.18). But

$$U_3(x_3,y_3) = -\frac{ie^{ikf}}{\lambda f}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} U_2(x,y) \times \exp\left\{\frac{ik}{2f}\left[(x_3-x)^2 + (y_3-y)^2\right]\right\}dx\,dy \tag{2.88}$$

as

$$r^2 = f^2 + (x_3-x)^2 + (y_3-y)^2,$$

so that if $(x_3 - x) \ll f$.

$$r \approx f + \frac{(x_3-x)^2}{2f} + \frac{(y_3-y)^2}{2f}.$$

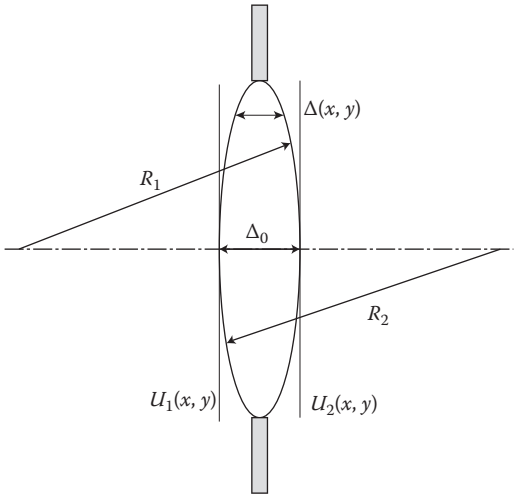

FIGURE 2.17 The thin lens.

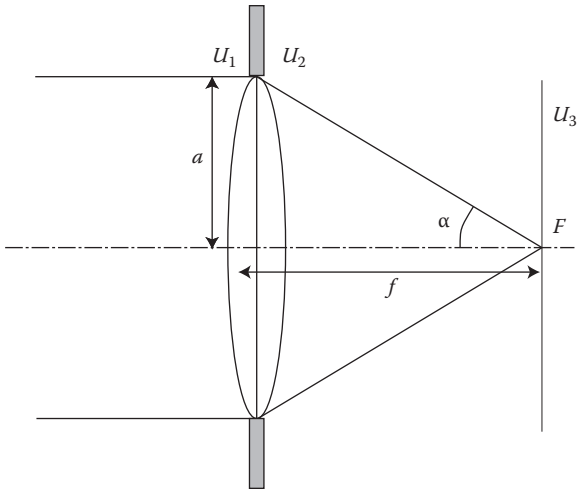

FIGURE 2.18 Focusing by a lens.

From now on we shall not write explicitly the limits $\pm\infty$ on the integrals. The field after the lens is

$$U_3(x_3,y_3) = -\frac{ie^{ikf}}{\lambda f}\iint P(x,y)\exp\left[-\frac{ik(x^2+y^2)}{2f}\right]\exp\left[\frac{ik(x^2+y^2)}{2f}\right] \times \exp\left[\frac{ik(x_3^2+y_3^2)}{2f}\right]\exp\left[-\frac{ik}{f}(xx_3+yy_3)\right]dx\,dy$$
$$= -\frac{ie^{ikf}}{\lambda f}\exp\left[\frac{ik(x_3^2+y_3^2)}{2f}\right]\iint P(x,y) \times \exp\left[-\frac{ik}{f}(xx_3+yy_3)\right]dx\,dy. \quad (2.89)$$

We now consider some important special cases for the pupil function.

2.12 Circularly Symmetric Aperture

Most real optical systems have circular symmetry, so we can put

$$P(x,y) = P(r)$$

and

$$U_3(r_3) = -\frac{ie^{ikf}}{\lambda f}\exp\left(\frac{i\pi r_3^2}{\lambda f}\right)\int_0^\infty P(r)J_0\left(\frac{2\pi r r_3}{\lambda f}\right)2\pi r\,dr. \quad (2.90)$$

In particular, for a circular aperture of radius a

$$U_3(r_3) = -\frac{ie^{ikf}}{\lambda f}\exp\left(\frac{i\pi r_3^2}{\lambda f}\right)\pi a^2\left[\frac{2J_1(2\pi r_3 a/\lambda f)}{2\pi r_3 a/\lambda f}\right]. \quad (2.91)$$

Let us define the *numerical aperture* of the lens:

$$NA = \sin\alpha = \frac{a}{f}. \quad (2.92)$$

We define a normalized optical coordinate

$$v = kr_3\sin\alpha \approx \frac{2\pi r_3 a}{\lambda f}. \quad (2.93)$$

and also put

$$\rho = \frac{r}{a}. \quad (2.94)$$

Then in general

$$U_3(v) = -\frac{ie^{ikf}}{\lambda f}\exp\left(\frac{iv^2\lambda f}{4\pi a^2}\right)2\pi a^2\int_0^1 P(\rho)J_0(v\rho)\rho\,d\rho. \quad (2.95)$$

or introducing the Fresnel number

$$N = \frac{a^2}{\lambda f}, \quad (2.96)$$

and the field is

$$U_3(v) = -i\pi N e^{ikf}\exp\left(\frac{iv^2}{4\pi N}\right)\int_0^1 2P(\rho)J_0(v\rho)\rho\,d\rho. \quad (2.97)$$

For a plain circular aperture

$$U_3(v) = -i\pi N e^{ikf}\exp\left(\frac{iv^2}{4\pi N}\right)\left[\frac{2J_1(v)}{v}\right]. \quad (2.98)$$

This is called the amplitude point spread function or impulse response. Note that

$$N = \frac{a^2}{\lambda f} = \frac{a\sin\alpha}{\lambda} = \frac{f\sin^2\alpha}{\lambda},$$

that is, for a big lens (compared with the wavelength), N is very large and the exponential term is close to unity. The intensity is the modulus squared of the amplitude, giving the Airy disc:

$$I(v) = \pi^2 N^2\left[\frac{2J_1(v)}{v}\right]^2. \quad (2.99)$$

The intensity variation is as shown in Figure 2.7. In Figure 2.9, the effect of a central obstruction is also shown. As the obstruction ratio ε increases, the point spread function becomes narrower, but with larger side-lobes.

2.13 Effect of Defocus

We now consider the field on a defocused plane, $z=f+\delta z$ (Figure 2.19). The field is

$$U_3(x_3,r_3) = -\frac{ie^{ikz}}{\lambda z}\iint P(x,y)\exp\left[-\frac{ik}{2f}(x^2+y^2)\right] \times \exp\left\{\frac{ik}{2z}\left[(x_3-x)^2+(y_3-y)^2\right]\right\}dx\,dy$$
$$= -\frac{ie^{ikz}}{\lambda z}\exp\left[\frac{ik}{2z}(x_3^2+y_3^2)\right]\iint P(x,y)\exp\left[-\frac{ik}{2f}(x^2+y^2)\right] \times \exp\left[\frac{ik}{2z}(x^2+y^2)\right]\exp\left[-\frac{2\pi i}{\lambda z}(xx_3+yy_3)\right]dx\,dy.$$

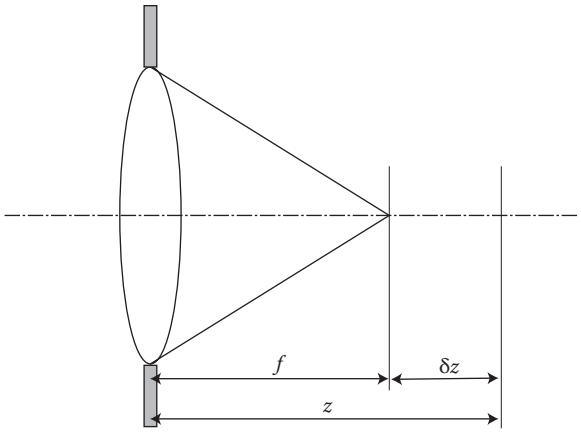

FIGURE 2.19 Geometry of defocus.

For a radially symmetric system

$$U_3(r_3) = -\frac{ie^{ikz}}{\lambda z}\exp\left(\frac{i\pi r_3^2}{\lambda z}\right)\int P(r) \times \exp\left[-\frac{ikr^2}{2}\left(\frac{1}{f}-\frac{1}{z}\right)\right] J_0\left(\frac{2\pi r r_3}{\lambda z}\right) 2\pi r\,dr. \quad (2.100)$$

For a circular aperture, again of radius a, put

$$v = \frac{2\pi a r_3}{\lambda z} = \frac{2\pi a r_3}{\lambda f} \quad (2.101)$$

$$u = \frac{2\pi a^2}{\lambda}\left(\frac{1}{f}-\frac{1}{z}\right) \approx \frac{2\pi a^2}{\lambda f^2}\delta z \quad (2.102)$$

if $\delta z \ll f$. So

$$U_3(v,u) = -i\pi N e^{ikz}\exp\left(\frac{iv^2}{4\pi N}\right)\int_0^1 2J_0(v\rho)\exp\left(-\frac{iu\rho^2}{2}\right)\rho\,d\rho \quad (2.103)$$

Along the axis, $v = 0$, the field is

$$\begin{aligned} U_3(v,u) &= -i\pi N e^{ikz}\exp\left(\frac{iv^2}{4\pi N}\right)\int_0^1 2\exp\left(-\frac{iu\rho^2}{2}\right)\rho\,d\rho \\ &= -i\pi N e^{ikz}\exp\left(\frac{iv^2}{4\pi N}\right)\left[-\frac{2}{iu}\exp\left(-\frac{iu\rho^2}{2}\right)\right]_0^1 \\ &= -i\pi N e^{ikz}\exp\left(-\frac{iu}{4}\right)\left[\frac{\sin(u/4)}{u/4}\right]. \end{aligned} \quad (2.104)$$

So the intensity along the axis is

$$I(0,u) = \pi^2 N^2\left[\frac{\sin(u/4)}{u/4}\right]^2, \quad (2.105)$$

The field at a general point can be calculated from Lommel functions or by numerical integration.

It is interesting to consider also the *annular* aperture $P(\rho) = \delta(\rho - 1)$. Then from Equation 2.100

$$U_3(v,u) = -2i\pi N e^{ikz}\exp\left(\frac{iv^2}{4\pi N}\right)\exp\left(-\frac{iu}{2}\right)J_0(v) \quad (2.106)$$

or

$$I(v,u) = 4\pi^2 N^2 J_0^2(v). \quad (2.107)$$

Note that the intensity does not change with u (within the range of validity of the equation). This represents a Bessel beam (a so-called diffraction-free beam), which is the subject of much research at present. Actually, it is very well known as a mode of free space in cylindrical coordinates, for example, of a circular waveguide. Power diffracts outward, but also *inward* from the large side lobes, to achieve a dynamic equilibrium.

2.14 Image Formation

Before the lens (Figure 2.20)

$$U_2(x_2,y_2) = -\frac{ie^{ikd_1}}{\lambda d_1}\iint U_1(x_1,y_1) \times \exp\left\{\frac{ik}{2d_1}\left[(x_2-x_1)^2+(y_2-y_1)^2\right]\right\}dx_1dy_1.$$

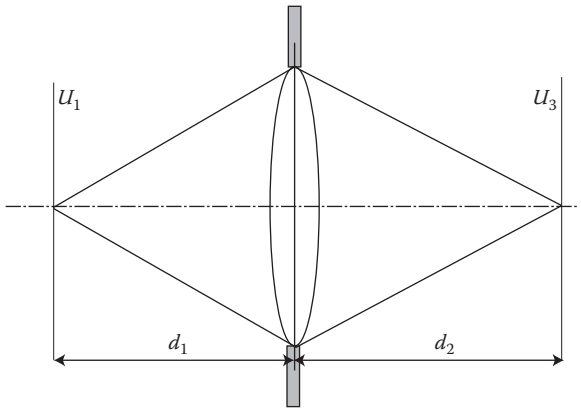

FIGURE 2.20 Imaging by a lens.

Multiplying by the pupil function, the amplitude after the lens is

$$U_2(x_2,y_2) = -\frac{ie^{ikd_1}}{\lambda d_1} P(x_2,y_2)\exp\left[-\frac{ik}{2f}\left(x_2^2+y_2^2\right)\right]\iint U_1(x_1,y_1) \times \exp\left\{\frac{ik}{2d_1}\left[(x_2-x_1)^2+(y_2-y_1)^2\right]\right\}\mathrm{d}x_1\,\mathrm{d}y_1.$$

So, finally the image amplitude is

$$\begin{aligned} U_3(x_3,y_3) &= -\frac{1}{\lambda^2 d_1 d_2}\exp\left[ik(d_1+d_2)\right]\iiiint P(x_2,y_2)U_1(x_1,y_1) \\ &\quad\times\exp\left\{\frac{ik}{2d_1}\left[(x_2-x_1)^2+(y_2-y_1)^2\right]\right\} \\ &\quad\times\exp\left\{\frac{ik}{2d_2}\left[(x_3-x_2)^2+(y_3-y_2)^2\right]\right\} \\ &\quad\times\exp\left[-\frac{ik}{2f}\left(x_2^2+y_2^2\right)\right]\mathrm{d}x_1\,\mathrm{d}y_1\,\mathrm{d}x_2\,\mathrm{d}y_2. \\ &= -\frac{1}{\lambda^2 d_1 d_2}\exp\left[ik(d_1+d_2)\right]\iiiint P(x_2,y_2)U_1(x_1,y_1) \\ &\quad\times\exp\left[\frac{ik}{2d_1}\left(x_1^2+y_1^2\right)\right]\exp\left[\frac{ik}{2d_2}\left(x_3^2+y_3^2\right)\right] \\ &\quad\times\exp\left[\frac{ik}{2}\left(\frac{1}{d_1}+\frac{1}{d_2}-\frac{1}{f}\right)\left(x_2^2+y_2^2\right)\right] \\ &\quad\times\exp\left\{-ik\left[x_2\left(\frac{x_1}{d_1}+\frac{x_3}{d_2}\right)+y_2\left(\frac{y_1}{d_1}+\frac{y_3}{d_2}\right)\right]\right\}\mathrm{d}x_1\,\mathrm{d}y_1\,\mathrm{d}x_2\,\mathrm{d}y_2. \end{aligned} \tag{2.108}$$

We now look at some special cases.

2.14.1 Special Cases

The first case that we consider is when the lens law is satisfied, $1/d_1+1/d_2=1/f$. Then, $d_2=Md_1$, where M is magnification. The field is

$$\begin{aligned} U_3(x_3,y_3) &= -\frac{1}{\lambda^2 M d_1}\exp\left[ikd_1(1+M)\right]\iiiint P(x_2,y_2)U_1(x_1,y_1) \\ &\quad\times\exp\left[\frac{ik}{2d_1}\left(x_1^2+y_1^2\right)\right]\exp\left[\frac{ik}{2Md_1}\left(x_3^2+y_3^2\right)\right] \\ &\quad\times\exp\left\{-\frac{ik}{d_1}\left[x_2\left(x_1+\frac{x_3}{M}\right)+y_2\left(y_1+\frac{y_3}{M}\right)\right]\right\}\mathrm{d}x_1\,\mathrm{d}y_1\,\mathrm{d}x_2\,\mathrm{d}y_2. \end{aligned} \tag{2.109}$$

Performing the integrals in terms of x_2,y_2, we have

$$h(x,y) = \iint P(x_2,y_2)\exp\left[\frac{ik}{d_1}\left(x_2x+y_2y\right)\right]\mathrm{d}x_2\,\mathrm{d}y_2 \tag{2.110}$$

which is the point spread function, given by the Fourier transform of the pupil function. Then

$$\begin{aligned} U_3(x_3,y_3) &= -\frac{1}{\lambda^2 M d_1^2}\exp\left[ikd_1(1+M)\right] \\ &\quad\times\exp\left[\frac{ik}{2Md_1}\left(x_3^2+y_3^2\right)\right]\iint U_1(x_1,y_1) \\ &\quad\times\exp\left[\frac{ik}{d_1}\left(x_1^2+y_1^2\right)\right]h\left(x_1+\frac{x_3}{M},y_1+\frac{y_3}{M}\right)\mathrm{d}x_1\,\mathrm{d}y_1. \end{aligned} \tag{2.111}$$

Now, for good imaging, h falls off quickly, that is, x_1+x_3/M is small, or $x_1\approx -x_3/M$, so that (Goodman 1968)

$$\begin{aligned} U_3(x_3,y_3) &= -\frac{1}{\lambda^2 M d_1^2}\exp\left[ikd_1(1+M)\right] \\ &\quad\times\exp\left[-\frac{ik}{2Md_1}\left(x_3^2+y_3^2\right)\left(1+\frac{1}{M}\right)\right] \\ &\quad\times\iint U_1(x_1,y_1)h\left(x_1+\frac{x_3}{M},y_1+\frac{y_3}{M}\right)\mathrm{d}x_1\,\mathrm{d}y_1. \end{aligned} \tag{2.112}$$

The intensity can thus be written as

$$I_3(x_3,y_3) = \frac{1}{\left(\lambda^2 M d_1^2\right)^2}\left|(U_1\otimes h)\right|^2, \tag{2.113}$$

where $\otimes$ represents the convolution operation. We can show that this is also valid with defocus: then h is then the defocused point spread function and

$$u = \frac{2\pi a^2}{\lambda}\left[\frac{1}{f}-\left(\frac{1}{d_1}+\frac{1}{d_2}\right)\right]. \tag{2.114}$$

So, for coherent imaging, for an object $t(x, y)$, the intensity in the image is

$$I_3(x_3,y_3) = \frac{1}{\left(\lambda^2 M d_1^2\right)^2}\left|(t\otimes h)\right|^2. \tag{2.115}$$

Imaging is *linear* in amplitude, and space invariant, that is the convolution means that each point of the object results in a distribution of *amplitude* in the image given by the *amplitude* point spread function. Finally, the *intensity* in the image is given by finding the modulus squared of the amplitude.

Note that Equation 2.112 shows that the image of a point (x_1, y_1) in the object occurs at a point $(x_1+x_3/M=0, y_1+y_3/M=0)$, the center of the point spread function h. That is, at $x_3=-Mx_1$, $y_3=-My_1$: thus the image is *inverted* and magnified by a factor M.

Next, we consider the defocused case, $1/d_1 + 1/d_2 \neq 1/f$. We take

$$\frac{1}{d_1} + \frac{1}{d_2} - \frac{1}{f} = \frac{1}{d_0}. \tag{2.116}$$

From Equation 2.108 we can see that Equation 2.109 is still valid if we put $P_{\text{eff}}(x_2, y_2)$, given by

$$P_{\text{eff}}\left(x_2, y_2\right) = P\left(x_2, y_2\right)\exp\left[-\frac{ik}{2d_0}\left(x_2^2 + y_2^2\right)\right]. \tag{2.117}$$

This effective pupil function is called the defocused pupil function. It is a complex quantity, given by multiplying the ordinary pupil function by a quadratic phase variation. Equation 2.118 is only true for small defocus, as the approximation in Equation 2.112 is otherwise not valid. This is because the point spread function becomes broader with defocus, that is, the *intensity* point spread function behaves as shown in Figure 2.21.

As the system is defocused, the peak intensity decreases, and the pattern spreads out. By conservation of energy the total energy in the pattern, $\iint \left|h\left(x, y\right)\right|^2 \mathrm{d}x\mathrm{d}y$, must be constant. The zeros in the pattern also disappear with defocus. Note that the *amplitude* in the point spread function is complex for the defocused case.

Next, we now look at an example when defocus is not small. We consider the special case when $d_1 = d_2 = f$. So,

$$\frac{1}{d_0} = \frac{1}{d_1} + \frac{1}{d_2} - \frac{1}{f} = \frac{1}{f}, \tag{2.118}$$

and $M = 1$. To solve this case, we return to Equation 2.108. As it stands, the expression does not simplify much! The reason is that there is the Fresnel diffraction by the object, which is then truncated by the pupil (Figure 2.22). So it is quite a complicated problem. However, it can be solved if we consider the pupil to be very big, so there is negligible truncation. Then Equation 2.108 becomes

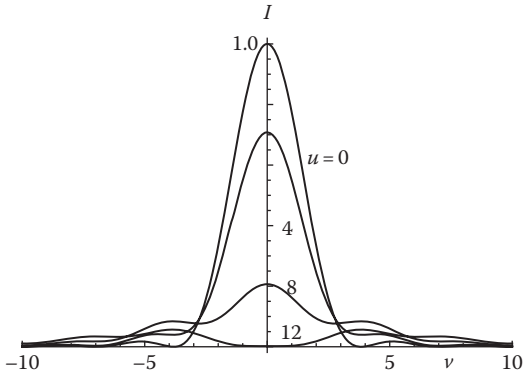

FIGURE 2.21 Defocused image of a point.

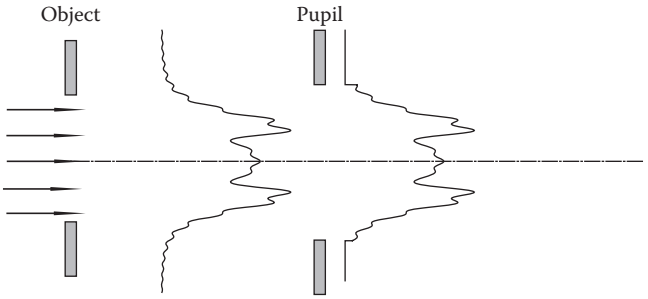

FIGURE 2.22 Truncation of the beam by the pupil.

$$\begin{aligned} U_3\left(x_3, y_3\right) = -\frac{1}{\lambda^2 f^2}\exp\left(ikf\right)\iiiint U_1\left(x_1, y_1\right) \\ \times \exp\left[\frac{ik}{2f}\left(x_1^2 + y_1^2\right)\right]\exp\left[\frac{ik}{2f}\left(x_3^2 + y_3^2\right)\right] \\ \times \exp\left[-\frac{ik}{2f}\left(x_2^2 + y_2^2\right)\right] \\ \times \exp\left\{-\frac{ik}{f}\left[x_2\left(x_1 + x_3\right) + y_2\left(y_1 + y_3\right)\right]\right\}\mathrm{d}x_1\,\mathrm{d}y_1\,\mathrm{d}x_2\,\mathrm{d}y_2 \end{aligned} \tag{2.119}$$

So far, this is not much simpler! But we can do the integrals in x_2 and y_2 now. We have

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\left[-\frac{ik}{2f}\left(x_2^2 + y_2^2\right)\right]\exp\left\{-\frac{ik}{f}\left[x_2\left(x_1 + x_3\right) + y_2\left(y_1 + y_3\right)\right]\right\}\mathrm{d}x_2\,\mathrm{d}y_2. \tag{2.120}$$

This is just the Fourier transform of a Gaussian (albeit of imaginary argument). You can get this from tables of Fourier transforms, or it can be evaluated using the properties of the Fresnel integrals. After putting $x = x_1 + x_3$, $y = y_1 + y_3$, it is

$$-\frac{2\pi f}{ik}\exp\left[-\frac{ik}{2f}\left(x^2 + y^2\right)\right]. \tag{2.121}$$

The important features to notice are that it is independent of x_2, y_2, and when you put it back in Equation 2.119, the quadratic terms in x_1, x_3, y_1 and y_3, all cancel to give

$$U_3\left(x_3, y_3\right) = -\frac{ie^{2ikf}}{\lambda f}\iint U_1\left(x_1, y_1\right)\exp\left[-\frac{ik}{f}\left(x_1x_3 + y_1y_3\right)\right]\mathrm{d}x_1\,\mathrm{d}y_1. \tag{2.122}$$

Compare this with Equation 2.89. Again we have the 2D Fourier transform, but now the parabolic phase factor of Equation 2.89 is no longer present (Figure 2.23). This is a very important result, which is the basis of most Fourier optics systems. It allows us

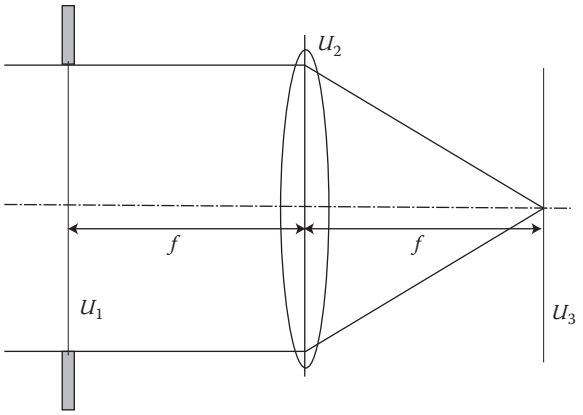

FIGURE 2.23 Imaging of an object in the front focal plane.

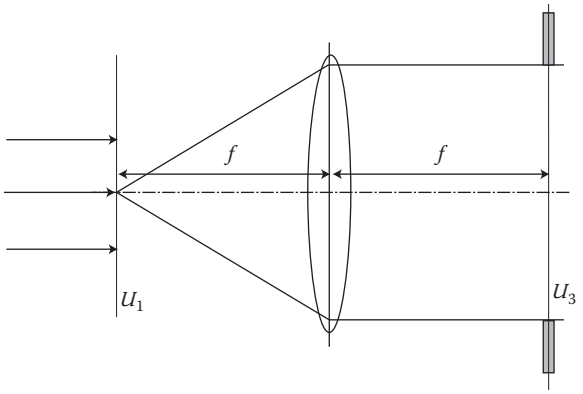

FIGURE 2.24 Fourier transformation by a lens.

to perform a 2D Fourier transform almost instantaneously (Figure 2.24), in the time for light to travel a distance 2*f*.

Note that if U_1 is a constant, assuming the pupil P is very big, U_3 is a delta function. If the pupil is finite, we get the point spread function. If U_1 is a circular function, a constant value within a circle, smaller than the pupil P, then the final amplitude is its Hankel function, that is the Airy disk $2J_1(v)/v$.

We have seen that if U_1 is a delta function, then U_3 is a constant. Two of these units may be coupled together, as in Figure 2.25. But the Fourier transform of the Fourier transform of $U_1(x)$ is $U_1(-x)$. So we just form an inverted, unity magnification image of U_1. This is called a 4*f* optical system.

2.15 Coherent Transfer Function

We have looked at two imaging systems, shown in Figures 2.20 and 2.25. For the system in Figure 2.20, Equation 2.89 shows that the image amplitude is multiplied by a parabolic phase factor, but for the system in Figure 2.25 there is no phase factor, and the field U_3 in the output plane can be

$$U_3(x_3, y_3) = (U_1 \otimes h)(x_3, y_3). \tag{2.123}$$

In each case the amplitude point spread function h is given by the 2D Fourier transform of the pupil function.

Considering the system in Figure 2.25, we see that if the pupils are very big, then U_3 is a perfect image of U_1. If we think of U_1 as a grating, it produces various diffraction orders, and these are combined by the second lens to produce an image. However, the pupil P cuts off some of the diffraction orders and hence a perfect image is not formed in practice. P can therefore be thought of as having the effect of a coherent transfer function, a low pass filter. We resolve U_1 into gratings, and some of these orders get through the system. The strength of the spatial frequency components is multiplied by P to give their strength in the image. This is the principle of the Abbe theory of image formation in a microscope.

Mathematically, we introduce the Fourier transform of the object amplitude U_1, given by $\tilde{U}_1$

$$\tilde{U}_1(m, n) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} U_1(x_1, y_1)\exp\left[-2\pi i(mx_1 + ny_1)\right] \mathrm{d}x_1\,\mathrm{d}y_1. \tag{2.124}$$

Inverting, we get

$$U_1(x_1, y_1) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \tilde{U}_1(m,n)\exp\left[2\pi i(mx_1 + ny_1)\right] \mathrm{d}m\,\mathrm{d}n. \tag{2.125}$$

But from Equation 2.123, and neglecting the multiplying constants we had previously

$$U_3(x_3, y_3) = \iint U_1(x_1, y_1) h\left(x_1 + \frac{x_3}{M}, y_1 + \frac{y_3}{M}\right) \mathrm{d}x_1\,\mathrm{d}y_1. \tag{2.126}$$

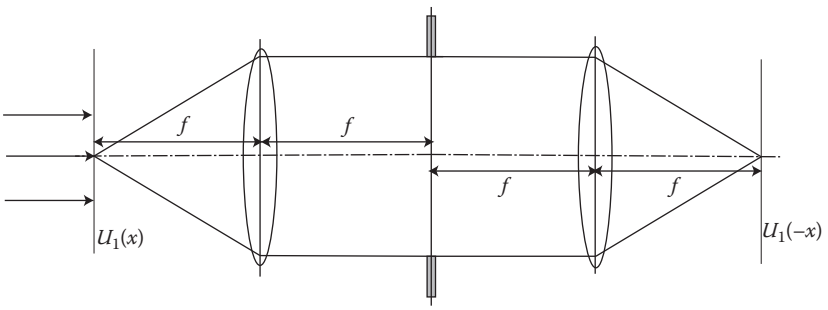

FIGURE 2.25 A 4*f* system.

Substituting Equation 2.125 in Equation 2.136:

$$U_3(x_3, y_3) = \iiiint \tilde{U}_1(m, n) h\left(x_1 + \frac{x_3}{M}, y_1 + \frac{y_3}{M}\right) \times \exp\left[2\pi i(mx_1 + ny_1)\right] dm\, dn\, dx_1\, dy_1. \quad (2.127)$$

But we also have

$$h(x, y) = \iint P(x_2, y_2) \exp\left[-\frac{ik}{f_1}(x_2 x + y_2 y)\right] dx_2\, dy_2, \quad (2.128)$$

$$P(x_2, y_2) = \iint h(x, y) \exp\left[\frac{ik}{f_1}(x_2 x + y_2 y)\right] dx\, dy. \quad (2.129)$$

So, doing the integrals in x_1, y_1 in Equation 2.127, and putting $x = x_1 + x_3/M$ and so on,

$$\iint h_1(x, y) \exp\left\{2\pi i\left[m\left(x - \frac{x_3}{M}\right) + n\left(y - \frac{y_3}{M}\right)\right]\right\} dx\, dy$$

$$= P(m\lambda f_1, n\lambda f_1) \exp\left[-2\pi i\left(\frac{mx_3}{M} + \frac{ny_3}{M}\right)\right].$$

So Equation 2.127 can be written as

$$U_3(x_3, y_3) = \iint \tilde{U}_1(m, n) P(m\lambda f_1, n\lambda f_1) \times \exp\left[-2\pi i\left(\frac{mx_3}{M} + \frac{ny_3}{M}\right)\right] dm\, dn. \quad (2.130)$$

So if pupil function is as shown in Figure 2.26a, the coherent transfer function is as shown in Figure 2.26b. For an object that is only a function of x, $n = 0$ and the corresponding coherent transfer function is given by a section through the two-dimensional transfer function (Figure 2.27). Note that in general

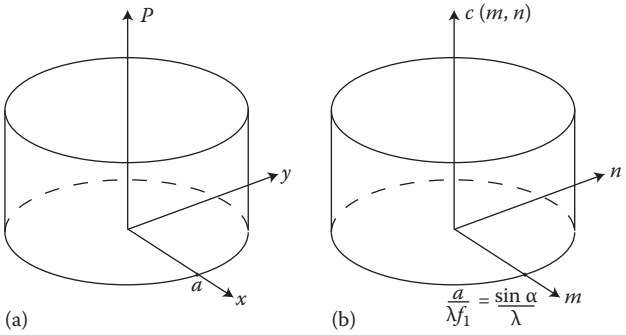

FIGURE 2.26 (a) The pupil function and (b) the coherent transfer function.

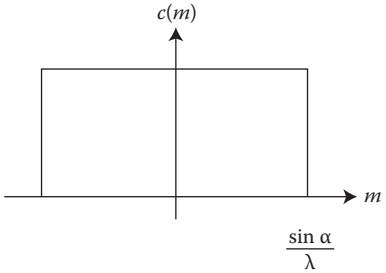

FIGURE 2.27 The coherent transfer function for imaging of a line structure.

there will be both positive and negative spatial frequencies. The imaging system behaves as a low-pass filter, that is, it transmits spatial frequencies less than (sin α)/λ. Note the sudden cutoff, corresponding to the edge of the pupil, as compared with transfer functions of electrical systems, which roll off smoothly.

2.15.1 A Grating Object

Consider as an example an amplitude transmittance that is a cosinusoidal variation on a constant background (Figure 2.28a):

$$U_1 = 1 + a\cos(2\pi\nu x). \quad (2.131)$$

The period of grating is $\Lambda = 1/\nu$. The spectrum of the object is the Fourier transform of U_1. As

$$U_1 = 1 + \frac{a}{2}e^{2\pi i\nu x} + \frac{a}{2}e^{-2\pi i\nu x}$$

we have for the object spectrum (Figure 2.28b)

$$T(m) = \tilde{U}_1 = \left[\delta(m) + \frac{a}{2}\delta(m - \nu) + \frac{a}{2}\delta(m - \nu)\right]\delta(n). \quad (2.132)$$

So, in this case the image amplitude is

$$U_3(x_3, y_3) = \iint \left[\delta(m) + \frac{a}{2}\delta(m - \nu) + \frac{a}{2}\delta(m + \nu)\right] c(m, n)\delta(n) \times \exp\left[-\frac{2\pi i}{M}(mx_3 + ny_3)\right] dm\, dn$$

$$= c(0) + \frac{a}{2}c(\nu)\exp\left(-\frac{2\pi i\nu x_3}{M}\right) + \frac{a}{2}c(-\nu)\exp\left(\frac{2\pi i\nu x_3}{M}\right).$$

For a circular pupil, c is even; so $c(-\nu) = c(+\nu)$, and we obtain for the image amplitude

$$U_3(x_3, y_3) = c(0) + ac(\nu)\cos\left(\frac{2\pi\nu x_3}{M}\right).$$

In this case, if $\nu < a/\lambda f_1$ and $c(\nu) = 1$, the image is the same as the object but magnified by M.

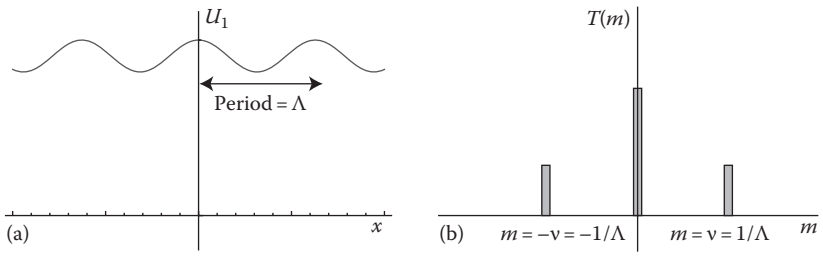

FIGURE 2.28 (a) A grating object and (b) its spatial frequency content.

The image *intensity* is thus

$$I_3(x_3, y_3) = \left| c(0) + ac(v)\cos\left(\frac{2\pi v x_3}{M}\right)\right|^2.$$

For the moment, take a as *real*, and also c as real; then

$$I_3(x_3, y_3) = c^2(0) + 2c(0)c(v)a\cos\left(\frac{2\pi v x_3}{M}\right) + a^2c^2(v)\cos^2\left(\frac{2\pi v x_3}{M}\right). \tag{2.133}$$

Using the identity $\cos(2\theta) = \cos^2\theta - 1$, we obtain

$$I_3(x_3, y_3) = \left[c^2(0) + \frac{a^2}{2}c^2(v)\right] + 2c(0)c(v)a\cos\left(\frac{2\pi v x_3}{M}\right) + \frac{a^2}{2}c^2(v)\cos\left(\frac{4\pi v x_3}{M}\right). \tag{2.134}$$

Note that a second harmonic term is introduced by the squaring operation, but this can be neglected if a is small.

Now let us consider what happens if a and c are complex, and $|a|$ is small. Now Equation 2.133 becomes

$$I_3(x_3, y_3) = |c(0)|^2 + 2c(0)\mathrm{Re}\left[ac(v)\right]\cos\left(\frac{2\pi v x_3}{M}\right). \tag{2.135}$$

If we consider an object whose *thickness* l varies in a cosinusoidal fashion

$$l = l_0 + l_1\cos(2\pi v x),$$

Then if its refractive index is n, the phase change on passing through it is nl, and so the amplitude on the far side if it is illuminated with a plane wave is

$$t = \exp\left\{ik\left[l_0 + l_1\cos(2\pi v x)\right]\right\} = \exp(ikl_0)\exp\left[ikl_1\cos(2\pi v x)\right]. \tag{2.136}$$

The first part of this is just a constant phase term and hence can be ignored. We then expand the second part into its Fourier components. Actually, this is identical to frequency modulation (FM) in communication theory, and the strengths of the various harmonics are given by Bessel functions. Let us just look at the much simpler case when $kl_1 \ll 1$ however. Then

$$t = 1 + ikl_1\cos(2\pi v x). \tag{2.137}$$

We can see that this is the same as Equation 2.131 with a given by an imaginary quantity. Thus a phase object is represented by imaginary a. We see from Equation 2.135, that if $c(\nu)$ is *real* then there is no image, and we just see a constant intensity. Note that this is not in general true for a strong phase object where terms of strength a^2 cannot be ignored.

Previously, we introduced the concept of the defocused pupil function. From Equations 2.100 or 2.117 we have

$$P_{\mathrm{eff}}(\rho) = P(\rho)\exp\left(-\frac{iu}{2}\rho^2\right), \tag{2.138}$$

so that now we introduce the defocused transfer function $c(m,u)$, as shown in Figure 2.29. For $u = 0$, that is the focused case, $c(m)$ is purely real. As u is increased, the imaginary part increases in strength. For $u > \pi$, the real part starts to go negative, which is

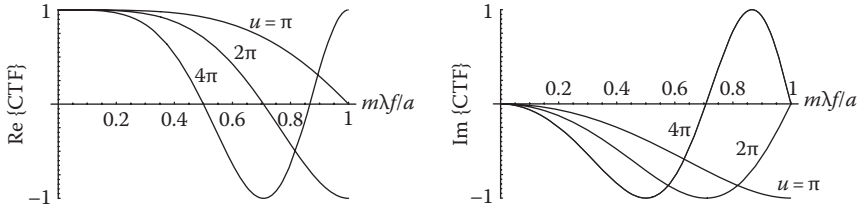

FIGURE 2.29 Defocused coherent transfer function.

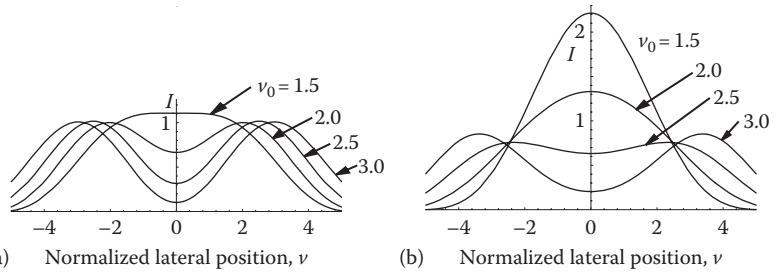

FIGURE 2.30 Images of a two-point object: (a) incoherent imaging and (b) coherent imaging.

not good for imaging of amplitude information. The imaginary part images phase information, but this also starts to go negative for $u > 2\pi$. So we only get good imaging of phase information if $u \le 2\pi$. If u is negative the imaginary part of $c(m, u)$ also becomes positive, so contrast is reversed. For a phase object, there is no contrast at the focus, and positive or negative contrast on either side of focus. Defocusing of the microscope was often used to image weak phase structures, as for example in biological samples, before more modern methods of phase imaging were invented.

2.15.2 Square Wave Object

Consider a square wave object, with transmittance as shown in Figure 2.30. Then

$$U_1(x) = \frac{1}{2} + \frac{2}{\pi}\left[\cos(2\pi\nu x) - \frac{1}{3}\cos(6\pi\nu x) + \frac{1}{5}\cos(10\pi\nu x) - \cdots\right]. \tag{2.139}$$

The harmonics are transmitted up to the cutoff frequency. By making $1/\nu$ very large, we can calculate the image of a single edge. Interestingly, although the number of terms then is very large we still get an overshoot as in the image of a square wave object. The image of an edge can be expressed in terms of Fresnel integrals. The fact that the wiggles do not disappear for a large number of terms is a consequence of the Gibbs phenomenon.

2.16 Spatial Filtering

The principle of spatial filtering is to alter the strength in the image of the Fourier components of the object by putting in an appropriate mask: For example, suppose we change the phase everywhere except at $m = 0$ by $\pi/2$. Then the transfer function is

$$c(\nu) = \begin{cases} -i & \nu \neq 0 \\ 1 & \nu = 0 \end{cases}. \tag{2.140}$$

Then from Equation 2.143

$$I_3 = 1 + 2\,\mathrm{Re}(-ai)\cos\left(\frac{2\pi\nu x_3}{M}\right) = 1 + 2\,\mathrm{Im}(a)\cos\left(\frac{2\pi\nu x_3}{M}\right). \tag{2.141}$$

So again we have managed to image the phase information. This method is called Zernike phase contrast, for the invention of which Zernike received the Nobel Prize. In practice, it is easier to change the phase just of the direct beam $m = 0$, rather than vice versa, but the result is exactly the same. Note that changing the phase by $-\pi/2$ rather than by $+\pi/2$ reverses the contrast. Also, if we make the filter

$$c(\nu) = \begin{cases} -bi, & \nu \neq 0, \\ 1, & \nu = 0, \end{cases} \tag{2.142}$$

we then have

$$I_3 = 1 + 2b\,\mathrm{Im}(a)\cos\left(\frac{2\pi\nu x_3}{M}\right). \tag{2.143}$$

If $b > 1$, we also enhance the contrast by a factor b, making very weak phase objects visible.

2.17 Incoherent Imaging

Most of what we have said so far is applicable to *coherent* optical systems. This requires that the point spread function of the imaging system is small in extent compared with the lateral spatial coherence of the illumination. The opposite condition, where the point spread function is large compared with the spatial coherence, results in *incoherent* imaging. Incoherent imaging is in everyday life a much more common phenomenon. For example photography, or just seeing, is normally an incoherent process. Fluorescence also results in incoherent image formation. The in-between case, where the point spread function is about the size of the spatial coherence of the illumination, is *partially coherent*. This case is much more complicated, and arises for example in the theory of microscope imaging. In incoherent imaging, because there is no coherent interference between neighboring points, we have to sum the intensities for

these points, rather than the complex amplitudes. We return to the geometry of Section 2.14.

Each point in the object results in a diffraction blur in the image. To obtain the total image we add the contributions from the *intensities* of the individual points.

For *coherent* imaging, we had in Equation 2.108

$$I_3(x_3, y_3) = \frac{1}{\left(\lambda^2 M d_1^2\right)^2} |t \otimes h|^2,$$

but now for *incoherent* imaging, we have

$$I_3(x_3, y_3) = \frac{1}{\left(\lambda^2 M d_1^2\right)^2} \left(|t|^2 \otimes |h|^2\right). \tag{2.144}$$

That is we must convolve the object *intensity* $|t|^2$ with the *intensity* point spread function $|h|^2$. Note that the term point spread function can refer to either the amplitude or the intensity point spread function in the literature.

We can derive Equation 2.144 properly by considering a single point in the object U_1 at (x,y). Its image from Equation 2.111 is

$$\begin{aligned} U_3(x_3, y_3) &= \frac{1}{\lambda^2 M d_1^2} \exp\left[ikd_1(1+M)\right] \exp\left[\frac{ik}{2Md_1}\left(x_3^2 + y_3^2\right)\right] \\ &\quad \times \iint \delta(x_1 - x)\delta(y_1 - y) \\ &\quad \times \exp\left[\frac{ik}{d_1}\left(x_1^2 + y_1^2\right)\right] h\left(x_1 + \frac{x_3}{M}, y_1 + \frac{y_3}{M}\right) \mathrm{d}x_1 \mathrm{d}y_1 \\ &= \frac{1}{\lambda^2 M d_1^2} \exp\left[ikd_1(1+M)\right] \exp\left[\frac{ik}{2Md_1}\left(x_3^2 + y_3^2\right)\right] \\ &\quad \times \exp\left[\frac{ik}{d_1}\left(x^2 + y^2\right)\right] h\left(x + \frac{x_3}{M}, y + \frac{y_3}{M}\right). \end{aligned}$$

So the intensity in the image of a single point is simply

$$I_3(x_3, y_3) = \frac{1}{\left(\lambda^2 M d_1^2\right)^2} \left| h\left(x + \frac{x_3}{M}, y + \frac{y_3}{M}\right) \right|^2. \tag{2.145}$$

Finally, adding up for the points of the object

$$I_3(x_3, y_3) = \frac{1}{\left(\lambda^2 M d_1^2\right)^2} \iint \left| h\left(x + \frac{x_3}{M}, y + \frac{y_3}{M}\right) \right|^2 |t(x,y)|^2 \, \mathrm{d}x \, \mathrm{d}y \tag{2.146}$$

which represents Equation 2.144.

2.17.1 Two-Point Object

One of the most important theoretical objects is two bright points in a dark background (e.g., two pinholes in an opaque screen, or two stars). The image is then, for different normalized separations $2v_0$, as shown in Figure 2.31a. For $v_0 = 2.5$ the points are well resolved. For $v_0 = 1.5$ they are not: they just almost look like one point. It is traditional to say that the points are just resolved when the maximum of one point spread function is placed over the zero of the other. This is called the Rayleigh criterion, and occurs when $v_0 = 1.92$. It is found that for a circular pupil aperture, the intensity in the middle is then 0.735 times the intensity at the points themselves. Note that the Rayleigh criterion applies for incoherent imaging. In general, though, we may introduce the *generalized Rayleigh criterion*, which states that the points are just resolved if the intensity at the center is 0.735 times that at the points. The intensity at the points need not be the same as the intensity at the maximum: some published papers have got this wrong! By putting in the values for the Bessel function we obtain, for the incoherent case

$$(\Delta x)_{\min} = 0.61 \frac{\lambda}{\sin\alpha}. \tag{2.147}$$

This is often written as $1.22f\lambda/D$ in terms of the diameter D of the pupil, rather than the radius. A typical value for a high power microscope is about 0.5 μm. For the coherent case, similar plots are shown in Figure 2.31b. We find

$$(\Delta x)_{\min} = 0.82 \frac{\lambda}{\sin\alpha}, \tag{2.148}$$

that is, the resolution is not as good.

2.17.2 Optical Transfer Function

Equation 2.146 is linear in intensity. This means we can introduce a transfer function, usually called the optical transfer function (OTF). Note that this operates on *intensities*, rather than the amplitudes for the coherent transfer function described earlier. So they are not strictly comparable.

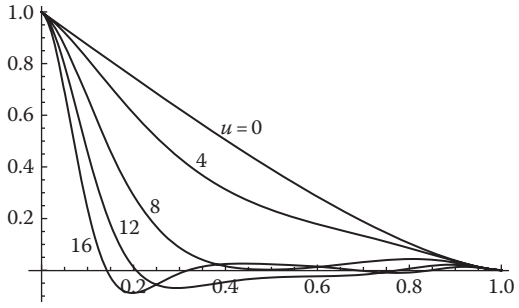

FIGURE 2.31 Defocused optical transfer function for a circular pupil.

We introduce the object *intensity* spectrum, given by the Fourier transform of its intensity, which is in contrast with Equation 2.124,

$$V_1(m,n) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \left|U_1(x_1, y_1)\right|^2 \exp\left[-2\pi i\left(mx_1 + ny_1\right)\right] \mathrm{d}x_1\, \mathrm{d}y_1. \tag{2.149}$$

Then we can show by exactly the same method as in Section 2.16 that

$$I(x_3, y_3) = \iint V_1(m, n) C(m, n) \exp\left[-2\pi i\left(\frac{mx_3}{M} + \frac{ny_3}{M}\right)\right] \mathrm{d}m\, \mathrm{d}n \tag{2.150}$$

where the OTF, $C(m,n)$, is given by the Fourier transform of the intensity point spread function $|h|^2$. As for the coherent case we had $P(m\lambda f, n\lambda f) = F(h)$, where $F(.)$ represents the Fourier transform, now we have

$$\begin{aligned} C(m, n) &= F\left(|h|^2\right) \\ &= P(m\lambda f, n\lambda f) \otimes P^*(m\lambda f, n\lambda f), \end{aligned} \tag{2.151}$$

This 2D convolution represents the area of overlap of the two pupils. We can show that the area of overlap for two circles is

$$\frac{2}{\pi}\left[\cos^{-1}(\tilde{m}) - \tilde{m}\sqrt{1-\tilde{m}^2}\right], \tag{2.152}$$

which is shown in Figure 2.30. The cutoff frequency is twice that for a coherent system.

If the system is defocused, we must integrate over the area of overlap taking into account the phase of the pupil. This cannot be done analytically, but can be expressed in terms of a single integral (Hopkins 1955). The result is also shown in Figure 2.31. The response drops off with defocus, that is, imaging of these particular spatial frequency components is worse. It is the mid-spatial frequencies that are most strongly affected, resulting in poorer imaging. Note that the OTF must always be purely real. An important feature is that the OTF can go negative with defocus. This means that some spatial frequency components have their contrast reversed. This results in optical artifacts, which means that you can see something that is not really in the object.

References

Abramowitz, M. and Stegun, I. A. 1965. *Handbook of Mathematical Functions*. New York: Dover.

Born, M. and Wolf, E. 1975. *Principles of Optics*. Oxford: Pergamon Press.

Goodman, J. W. 1968. *Introduction to Fourier Optics*. New York: McGraw-Hill.

Hecht, E. 1987. *Optics*. Reading, MA: Addison-Wesley.

Hopkins, H. H. 1955. The frequency response of a defocused optical system. *Proc. R. Soc. Lond. Ser. A* 231: 91–103.

3

Optics: Basic Physics

Raghuveer Parthasarathy
University of Oregon

3.1 Introduction

Optics is both a very old and a very contemporary field of research. Mirrors made millennia ago as well as the advanced imaging methods that decorate recent years' lists of scientific breakthroughs (Betzig et al. 2006, Bates et al. 2007, Hell 2007, Abbott 2009) can all be understood with the same physical framework. We will explore the basic physics of optics in this chapter, intended to serve as a review of elementary principles, or as an introduction for readers new to optics. Our treatment is necessarily brief and minimal—the reader interested in further elaboration should consult a textbook devoted to optics, such as Hecht (2002) or Born and Wolf (1997).

3.2 Electromagnetic Waves and Wave Motion

3.2.1 Light Is an Electromagnetic Wave

Insightful experiments by Hans Christian Ørsted, Michael Faraday, and others in the first half of the nineteenth century revealed the principle of electromagnetic induction: a changing magnetic field gives rise to an electric field, and, conversely, a changing electric field creates a magnetic field. Later, James Clerk Maxwell synthesized these and other observations into a set of succinct mathematical expressions, known as Maxwell's equations, which encapsulate the core of classical electromagnetism. It follows simply from these that electric ($\vec{E}$) and magnetic ($\vec{B}$) fields in vacuum can be connected by the relations

$$\nabla^2\vec{E} = \int \varepsilon_0\mu_0 \frac{\partial^2\vec{E}}{\partial t^2},$$

$$\nabla^2\vec{B} = \varepsilon_0\mu_0 \frac{\partial^2\vec{B}}{\partial t^2},$$

where

ε_0 is the permittivity of free space (a constant)
μ_0 is the permeability of free space (another constant)
t is the time

Both of these expressions have form of a wave equation

$$\nabla^2\psi = \frac{1}{v^2}\frac{\partial^2\psi}{\partial t^2},$$

which admits solutions of the form $\psi(\vec{r}, t) = f(\vec{r} - \vec{v}t)$, i.e., waves traveling through space ($\vec{r}$) and time (t) with velocity $\vec{v}$. Maxwell therefore realized that electric and magnetic fields can propagate as traveling waves, with a speed that is a simple function of electrostatic and magnetostatic constants: $c = (\varepsilon_0\mu_0)^{-1/2}$. Inserting values for ε_0 and μ_0 yields $c = 3.0 \times 10^8$ m/s, in striking correspondence

to the speed of light, which had been measured with few-percent accuracy by the mid-nineteenth century. Especially following the experiments of Heinrich Rudolph Hertz, in which electromagnetic waves were generated and detected, it became clear that light is an electromagnetic wave and that visible light is but one part of a broader electromagnetic spectrum.

3.2.2 Waves—Notation

As discussed above, electric and magnetic fields in space obey wave equations. To define the terms and symbols related to wave motion, we will first consider a one-dimensional wave equation, the simplest solution to which is a sinusoidal traveling wave of *amplitude A*, *wavenumber k*, and *angular frequency* ω: $\psi(x,t) = A\cos(kx-\omega t-\delta) = \mathrm{Re}\left\{A\exp\left[j\left(kx-\omega t-\delta\right)\right]\right\}$, where $j=\sqrt{-1}$ and δ is a phase offset. (We generally will not bother explicitly writing that the real part of the complex exponential is to be considered.) The *wavelength* is given by $\lambda = 2\pi/k$, and the *frequency* by $f = \omega/2\pi$; if we consider a particular position in space, ψ oscillates with *period* $T = f^{-1}$. The wave speed is related to the other variables by $v = \omega k^{-1} = \lambda f$, as the reader may wish to illustrate by drawing the wave for various values of *t*. The argument of the oscillatory function is often referred to as the phase: $\phi(x,\ t) = kx - \omega t - \delta$. Considering a particular moment in time, the phase advances by 2π over a distance is given by λ; over an arbitrary distance Δ*x*, the phase shift is $\Delta\phi = 2\pi\Delta x\lambda^{-1}$.

For the one-dimensional traveling wave noted above, each point in space corresponds to a particular phase. In two- or three-dimensions, more complex structures arise. It is useful to consider points of equal phase, which we will refer to as *wavefronts*.

Plane waves. A simple and very useful construction is the *plane wave*. Let us illustrate this for a two-dimensional wave (Figure 3.1), in which we can plot the value of ψ along the third dimension.

Note that ψ only varies along one spatial dimension (in this case, *x*). Contours of equal phase (i.e., wavefronts) are lines in the *xy* plane. As the wave travels, for the example shown in Figure 3.1, it moves in the $\hat{x}$-direction—i.e., parallel to a wavevector, $\vec{k}$,

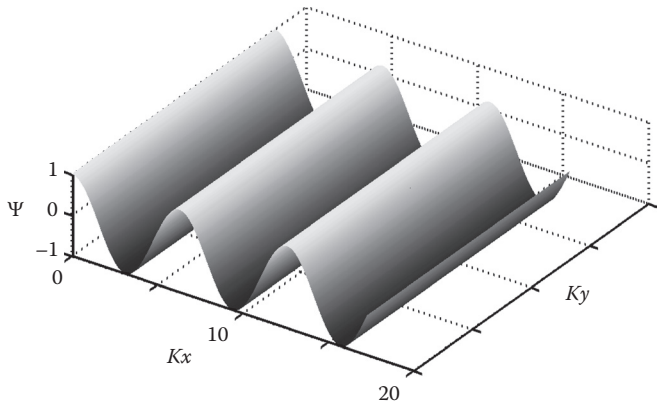

FIGURE 3.1 A two-dimensional plane wave: $\psi(x,y) = \cos(kx-\omega t)$, plotted at time $t = 0$.

that is perpendicular to these lines of constant phase and parallel to $\hat{x}$. We can write $\psi(x,\ y) = A\ \exp[j(kx-\omega-\delta)]$, or $\psi(\vec{r}) = A\ \exp[j(\vec{k}\cdot\vec{r}-\omega t-\delta)]$, where $\vec{r}$ is a vector in the *xy* plane—note that the dot product selects the *x*-component of $\vec{r}$.

For a three-dimensional plane wave, positions of constant phase (i.e., wavefronts) form a set of parallel planes. This is a good description of many sorts of light beams. Furthermore, any three-dimensional wave can be expressed as a combination of plane waves by Fourier analysis. The three-dimensional plane wave is described by $\psi(\vec{r}) = A\ \exp[j(\vec{k}\cdot\vec{r}-\omega t-\delta)]$, where $\vec{r}$ is any vector in three-dimensional space. We will show this explicitly: consider a position vector $\vec{r} = x\hat{x}+y\hat{y}+z\hat{z}$, where ^ indicates a unit vector, and some particular vector $\vec{r}_0$. Their difference:

$$\vec{r}-\vec{r}_0 = (x-x_0)\hat{x}+(y-y_0)\,\hat{y}+(z-z_0)\,\hat{z}\,.$$

Consider the set of points $\{\vec{r}\}$ described by $(\vec{r}-\vec{r}_0)\cdot\vec{k}=0$. As $\vec{r}$ varies, this sweeps out a plane perpendicular to $\vec{k}$. Expanding this: $\left(\vec{r}-\vec{r}_0\right)\cdot\vec{k} = k_x\left(x-x_0\right)+k_y\left(y-y_0\right)+k_z\left(z-z_0\right)=0$, or $k_x x + k_y y + k_z z = a$, where $a = k_x x_0 + k_y y_0 + k_z z_0$ is a constant. Therefore, the equation of a plane perpendicular to $\vec{k}$ is $\vec{k}\cdot\vec{r} = \text{constant} = a$. The set of planes over which $\psi(\vec{r})$ (at $t=0$) varies sinusoidally is $\psi(\vec{r}) = A\ \cos(\vec{k}\cdot\vec{r})$ or $\psi(\vec{r}) = A\ \exp\ (j\ \vec{k}\cdot\vec{r})$. This function is periodic if $\vec{k}\cdot\vec{r}$ changes by 2π, i.e., $|\vec{k}|\lambda = 2\pi$, or $k = |\vec{k}| = 2\pi/\lambda$, as expected. The traveling plane wave is described by $\psi(\vec{r}) = A\ \exp[j\ (\vec{k}\cdot\vec{r}-\omega t-\delta)]$. To reiterate, the wavefronts of a three-dimensional plane wave are planes. Typically, we will only draw wavefronts that are separated in phase by $\Delta\phi = 2\pi$, which are therefore spatially separated by distance λ. The wavevector $\vec{k}$ points perpendicular to these planes. Often, one describes the wave by a *ray* that points along $\vec{k}$.

Spherical waves. A point-source of light emits *spherical waves*—the wavefronts are concentric spheres that travel away from the point. The wave function is

$$\psi(\vec{r},\ t) = \frac{A}{r}\exp\left[j\left(kr-\omega t-\delta\right)\right],$$

where

A is a constant
r is the distance from the source

The amplitude decreases with *r*, for reasons that will become clear shortly.

Cylindrical waves. A line-source of light, for example, a slit, emits *cylindrical waves*—the wavefronts are concentric cylinders that travel away from the line. The wave function is

$$\psi(\vec{r},\ t) = \frac{A}{\sqrt{r}}\exp\left[j\left(kr-\omega t-\delta\right)\right],$$

where

A is a constant
r is the distance from the line

Again, the amplitude decreases with *r*.

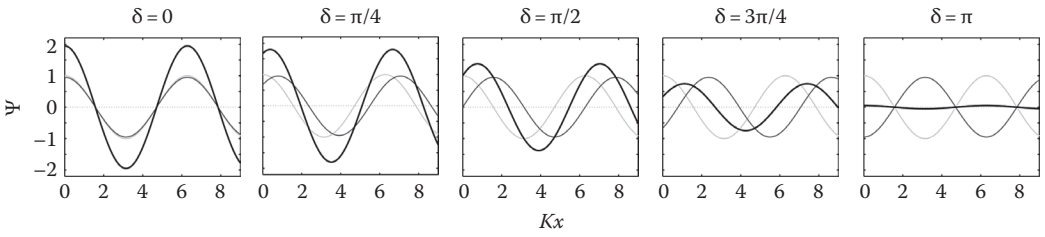

FIGURE 3.2 The sum $\psi = \psi_1 + \psi_2$ (black) of the waves $\psi_1 = 1.0\cos(kx)$ (light gray) and $\psi_2 = 0.95\cos(kx - \delta)$ (medium gray), plotted for various values of δ.

3.2.3 Superposition

The wave equation is linear in ψ; therefore, its solutions obey the principle of *superposition*: If ψ_1 and ψ_2 each satisfy the wave equation, then $\psi = \psi_1 + \psi_2$ is also a solution. The relative phase difference between ψ_1 and ψ_2 is important in determining their *interference*:

Figure 3.2 shows an illustration of the superposition of two sine waves. I have plotted $\psi_1 = 1.0\cos(kx)$, $\psi_2 = 0.95\cos(kx - \delta)$, and $\psi = \psi_1 + \psi_2$ for various values of δ. (I have chosen slightly different amplitudes for these two waves to make the illustrations clearer.)

Note that a phase difference $\delta = 0$ leads to *constructive interference*, and a phase difference $\delta = \pi$ leads to *destructive interference*.

3.2.4 Further Properties of Electromagnetic Waves

3.2.4.1 Electromagnetic Waves in Matter

We noted above that electric and magnetic fields can propagate in free space as waves, with speed $c = 3.0 \times 10^8$ m/s. In a transparent material of index of refraction n (related to the polarizability of the material), fields also propagate as waves, but more slowly, with speed $v = c/n$. For air at 20°C and atmospheric pressure, $n = 1.0003$. For water at 20°C, $n = 1.33$. For typical glass, $n = 1.46$. The *frequency* of the wave is unchanged from its value in vacuum—the rate of oscillation of the atoms excited by the electric field is constant. The *wavelength* of the light is different from its value in vacuum and obeys the general relation encountered earlier: $v = \lambda f$. Therefore, waves in matter are shorter than in free space: $\lambda = v/f = c/nf$. The wavelength in matter $\lambda = \lambda_0/n$, where λ_0 is the free space wavelength, and so the phase shift corresponding to a change in position Δx along the wave is $\Delta\phi = 2\pi(\Delta x/\lambda) = 2\pi n(\Delta x/\lambda_0)$.

Generally, when one states a wavelength for light, it is the *free space wavelength*, λ_0, that is being referred to—we say that orange light has a wavelength of ≈600 nm, even though when it enters your eye ($n = 1.3$), its wavelength shortens to ≈450 nm.

3.2.4.2 Polarization

Another consequence of electrodynamics is that the electric and magnetic field vectors at any point are perpendicular to one another and to the wave's propagation direction (Figure 3.3). The magnitudes of the field amplitudes are related by $|\vec{E}| = v|\vec{B}|$, where v is the speed. The direction of $\vec{E}$ specifies the *polarization* of the wave. If this direction is constant, as in Figure 3.4, we say that the wave is *linearly polarized* (or *plane polarized*). In Figure 3.4, for example, note that $\vec{E}$ is always parallel to the x-axis (in other words, $\vec{E} = E(z,t)\hat{z}$). Waves do not have to be plane polarized and can do a variety of interesting things. If the direction of $\vec{E}$ rotates as the wave propagates, then we have *circular* or elliptical polarization. (We will not go into the difference between the two—the reader can explore this as well as other constructions, such as radial polarization.)

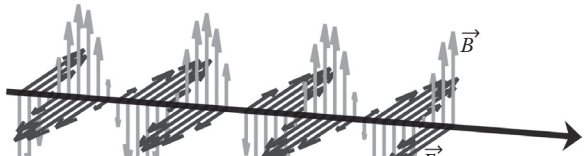

FIGURE 3.3 Electric (dark gray) and magnetic (light gray) fields of a plane-polarized electromagnetic wave. The black arrow indicates the propagation direction, perpendicular to $\vec{E}$ and $\vec{B}$.

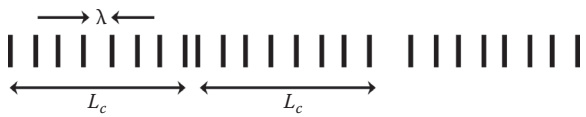

FIGURE 3.4 The coherence length, L_C, describes the spatial extent over which wavefronts (planes that differ by a phase shift of 2π) are separated by integer multiples of the wavelength. Over distances larger than $\approx L_C$, the coherence of the wave with itself—the ability to translate by an integer number of wavelengths and "match up"—is lost.

3.2.4.3 Energy and Momentum

Electromagnetic waves carry energy and momentum. The power per unit area crossing a surface is $\vec{S} = c^2\epsilon_0\vec{E} \times \vec{B}$, known as the Poynting vector. Note that it points along the propagation direction (i.e., parallel to $\vec{k}$), not surprisingly. Because $|\vec{E}| = v|\vec{B}|$, $\vec{S}$ is proportional to $|\vec{E}|^2$.

The *intensity* (or *irradiance*), I, of the wave is the average energy carried per unit area per unit time, i.e., the power per unit area. It is the intensity, not the electric field directly, that we "see" as brightness. "Average" means that we consider the average power over a period. (Note that the intensity is a *number*, not a vector.) Since $\vec{S}$ is proportional to $\left|\vec{E}\right|^2$, the intensity of an electromagnetic wave is proportional to $\left|\vec{E}\right|^2$ as well. This is in general true for vibrations and waves: the energy of a wave is proportional to the square of its amplitude, and therefore

$$I \propto \left|\vec{E}\right|^2 = \vec{E}\cdot\vec{E}^{*}.$$

The principle of conservation of energy and the proportionality of $\vec{S}$ and I is on $\left|\vec{E}\right|^2$ explain the decaying amplitude of the spherical and cylindrical waves discussed in Section 3.2.2. For a spherical wave, integrating the power carried by the wave over a shell of radius r surrounding the source must give a result that is independent of r—all the power must cross the shell, regardless of the size of the shell. The shell area scales as r^2 and so $\vec{S}$ must scale as r^{-2} for the product to be independent of r, from which we conclude that the amplitude scales as $\sqrt{r^{-2}} = r^{-1}$ for a three-dimensional spherical wave.

We often can ignore the constant of proportionality, being concerned with relative intensities. However, for completeness: $I = (1/2)\varepsilon_0 c E_0^2$ in vacuum, where E_0 is the electric field amplitude. In matter, $I = (1/2)\varepsilon_0 v E_0^2$, where v is the speed of the wave and ε is the permittivity of the medium. The ability of light to carry energy and momentum has become especially important in recent years with the development of optical trapping techniques, in which light itself is used to grab, pull, push, and twist microscopic objects.

Though it travels as a wave, light carries energy in discrete, quantized packets. This realization, primarily by Max Planck and Albert Einstein in the early twentieth century, marked the birth of quantum mechanics. The energy of a *photon*, the quantized "unit" of light, is proportional to its frequency and hence inversely proportional to its wavelength. More precisely, the photon energy $E = hf = hc/\lambda$, where $h = 6.626\times10^{-34}$ m^2 kg/s is Planck's constant. Photons of lower wavelength have more energy. This explains, for example, why the emission of light from fluorescent molecules necessarily occurs at higher wavelengths than does absorption: a photon is absorbed, and some energy is converted into nonradiative (e.g., vibrational) modes, leaving a smaller quantum of energy for emission.

3.2.5 The Electromagnetic Spectrum

The range of wavelengths of electromagnetic waves that are relevant to science and technology is enormous, ranging from very high-energy gamma rays spouting from astrophysical sources ($\lambda \approx 10^{-13}$ m) to x-rays used to probe molecular structure ($\lambda \approx 10^{-10}$ m) to microwaves ($\lambda \approx 10^{-2}$ m) to radio waves ($\lambda > 1$ m). "Visible light" spans the tiny range of wavelengths from about 400 (violet/blue) to 800 nm (red), yet its correspondence to the energetics of electronic transitions in molecules and, relatedly, our ability to see it, makes it an immensely useful part of the electromagnetic spectrum.

3.2.6 Coherence

We have been considering waves as ideal sinusoidal forms that oscillate at a unique frequency and extend infinitely through space. For such a perfectly *coherent* wave, the wavefronts are always separated by distance λ, and knowing the phase at one point specifies it at all points. For real waves, this is not exactly the case. The wavefronts of light from a real, imperfect wave, are separated by λ if we consider some finite span of approximate size L_C, but if we look at larger lengths, the phase relations appear "randomized"—see Figure 3.4.

This lack of perfect coherence arises from the emission of any real source not being perfectly monochromatic, but rather consisting of a range of output frequencies, Δf. (This is due to factors such as the finite linewidth of electronic transitions and the thermal velocities of atoms and molecules.) Roughly, $L_C = c/n\,\Delta f$. Furthermore, the light from extended sources such as an incandescent light bulb or the sun is emitted by many independent sources throughout the object, and each emitted wave has a random phase relative to any other. Such sources are referred to as *incoherent* light sources. The length L_C referred to above is called the *coherence length*—it is about 10 μm (around 20 λ) for a light bulb. (There is a more precise way to define the coherence length that we will not go into here.)

A *laser* is a *coherent* light source—all the waves emitted by the device have the same phase. Moreover, L_C is typically around 1 m (>10^6 λ) and can even be kilometers in length—a good approximation to our ideal infinite wave. This remarkable property of lasers contributes to their tremendous utility, as will be evident later in this book.

3.3 Diffraction

For centuries, debate raged over whether light is a wave or a particle—an interesting history that we will not go into. Whether or not it is important to consider the wave-nature of light in describing its propagation, rather than simply imagining rays of light that travel in simple geometric paths, depends on the spatial scale of the phenomena being considered. For features that are not large compared to the wavelength (λ), for example, visible light passing through micron-sized slits or kilometer-sized radio waves detected by an array of dishes, the wave nature of light is inescapable. Light's interference with itself determines its intensity profile, and diffraction—this interference being induced by barriers or obstacles—is paramount. This regime in which the wave nature of light is important is called *physical optics*. The regime in which the system size is much greater than the wavelength of light, and hence wave properties are relatively unimportant, is called *geometric optics* or *ray optics*.

Diffraction is a general property of waves, and the phenomena we will explore in this section also apply to water waves, sound waves, etc.

3.3.1 Two-Slit Interference

Consider a plane wave incident on a barrier with two slits, separated by a distance D (Figure 3.5). (Imagine the slits themselves to have negligible width—we will return to this later.) Each slit acts as a point-source for waves, which continue propagating to the right in the figure. Far to the right is a screen. We want to know the *intensity*, I, of the light hitting the screen as a function of θ, the angle relative to a line perpendicular to the barrier (see Figure 3.5).

The electric field of the incident wave is

$$\vec{E} = \vec{E}_0 \exp\left[j\left(kx - \omega t\right)\right],$$

with $k = 2\pi/\lambda$, as usual (see Section 3.2). We could add any phase offset to this—it does not matter, as we will see shortly. We are concerned with the light hitting a far-off screen, at angle θ. If the screen were close by, a ray would have to leave slit #1 at some angle θ_1 and slit #2 at some angle θ_2, where θ_1 and θ_2 may be different, to both reach the screen at angle θ. However, as the screen moves farther and farther away, both θ_1 and θ_2 approach θ—try drawing this if it's not evident. So, to consider $I(\theta)$, we need to consider rays leaving each slit at angle θ. Let us define our coordinates so that the barrier is at $x = 0$.

The two rays that travel at angle θ are indicated in Figure 3.6; their fields are

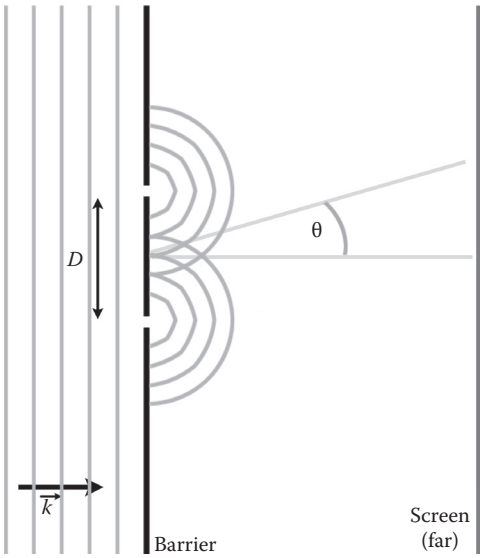

FIGURE 3.5 Two-slit interference: A plane wave is incident from the left on two slits of negligible width separated by distance D. Each slit acts like a point source for waves continuing to the right; the two resulting waves interfere with one another. This interference is manifested in the pattern of light intensity observed on a distant screen, and is a function of the wavelength, D, and the angle θ.

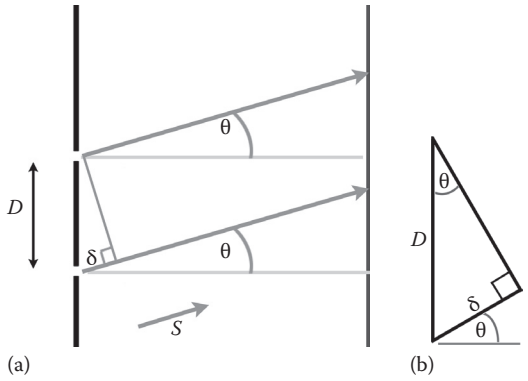

FIGURE 3.6 The geometry of light propagation for two-slit interference. (a) For the angle θ illustrated, light traveling from the lower slit (slit #2) travels a greater distance than light from slit #1. The extra path length is denoted δ and is the reason for a phase difference between the two waves. (b) A "zoomed in" view of the geometry relating D, δ, and θ.

$$\vec{E}_1 = \vec{E}_0 \exp\left[j\left(ks - \omega t\right)\right],$$
$$\vec{E}_2 = \vec{E}_0 \exp\left[j\left(k\left(s + \delta\right) - \omega t\right)\right],$$

where we have defined s as the coordinate in the "tilted" θ-direction, and we have indicated the extra distance that ray 2 has to travel by δ. Note that $\vec{E}_1(s=0)$ and $\vec{E}_2(s=-\delta)$ have the same phase, as they should, since they come from the same incident wave.

Graphically, we can see that if δ is an integer multiple of λ, the two waves will add constructively (Figure 3.7). If δ is a half-integer multiple of λ, the two waves will add destructively and give zero light intensity.

Let us examine this mathematically. The superposition of the two electric fields:

$$\vec{E} = \vec{E}_1 + \vec{E}_2 = \vec{E}_0 \exp\left[j\left(ks - \omega t\right)\right]\left\{1 + \exp\left[jk\delta\right]\right\}.$$

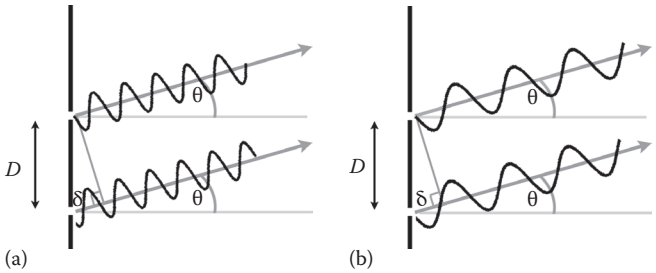

FIGURE 3.7 (a) If the extra path length, δ, between the two paths is an integer multiple of the wavelength, the two waves will constructively interfere, leading to high intensity at the screen. (b) If the extra path length δ between the two paths is a half-integer multiple of the wavelength, the two waves will destructively interfere, leading to zero intensity at the screen—note that when wave #1 is "up," wave #2 is "down" and vice versa.

From geometry, $\delta = D \sin\theta$ (Figure 3.6b), so

$$\vec{E} = \vec{E}_0 \exp\left[j\left(ks - \omega t\right)\right]\left\{1 + \exp\left[jkD\sin\theta\right]\right\},$$
$$\vec{E} = \vec{E}_0 \exp\left[j\left(ks - \omega t\right)\right]\left\{1 + \exp\left[2\pi jD\sin\theta/\lambda\right]\right\}.$$

The intensity (see Section 3.2.4.3) is given by $I \propto \left|\vec{E}\right|^2 = \vec{E}\cdot\vec{E}^{\star}$. Therefore,

$$I \propto \left|\vec{E}_0\right|^2 (1)\left\{1 + \exp\left[2\pi jD\sin\frac{\theta}{\lambda}\right]\right\}\left\{1 + \exp\left[-2\pi jD\sin\frac{\theta}{\lambda}\right]\right\},$$
$$I \propto \left|\vec{E}_0\right|^2 \left[2 + 2\cos\left(\frac{2\pi}{\lambda}D\sin\theta\right)\right],$$

making use of the Euler relation $\cos(x) = (1/2)(\exp(jx) + \exp(-jx))$. Via the identity $2[1 + \cos(2x)] = \cos^2(x)$, the intensity becomes

$$I \propto 4\left|\vec{E}_0\right|^2 \cos^2\left(\pi\frac{D\sin\theta}{\lambda}\right)$$

Note that *without* interference—just considering the incident plane wave, for example, $I \propto \left|\vec{E}_0\right|^2$, with the same constant of proportionality (*c*'s etc.)—we will define this intensity as I_0. Therefore,

$$I = 4I_0 \cos^2\left(\pi\frac{D\sin\theta}{\lambda}\right).$$

As we saw graphically, if $D \sin\theta = m\lambda$, where m is an integer, the $\cos^2$ factor is maximal, and we have *constructive* interference. If $D \sin\theta = (m/2)\lambda$, where m is an odd integer (i.e., $m/2 = 1/2, 3/2, 5/2, \ldots$), the $\cos^2$ factor is zero, and we have *destructive* interference. The intensity pattern, we see on the screen, therefore, is *not uniform* but rather has a sequence of maxima and minima. This is plotted in Figure 3.8.

Note that the maximal value of the intensity is four times that of a single wave. If interference "did not exist," we would have light from the two slits combining to simply give twice the single-wave intensity. With interference, we have bright peaks with four times the intensity and dark minima with zero intensity.

3.3.2 *N*-Slit Interference

Now consider N slits, *each* separated by distance D (drawn in Figure 3.9a for $N = 5$).

Building on our $N = 2$ analysis above, we can write the total electric field as

$$\vec{E} = \vec{E}_0 \exp\left[j\left(ks - \omega t\right)\right]\left\{1 + \exp\left[j\alpha\right] + \exp\left[j2\alpha\right] + \exp\left[j3\alpha\right] + \cdots + \exp\left[j(N-1)\alpha\right]\right\},$$

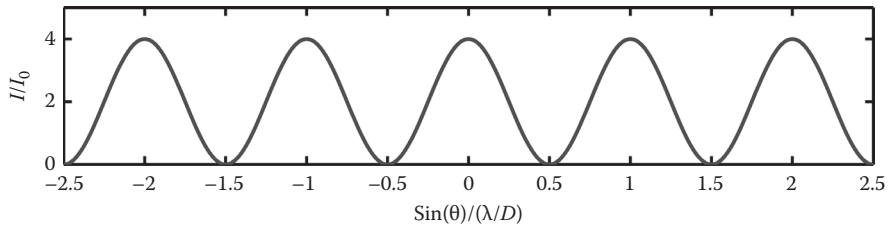

FIGURE 3.8 The two-slit intensity function: $I = 4I_0 \cos^2(\pi D \sin\theta\lambda^{-1})$.

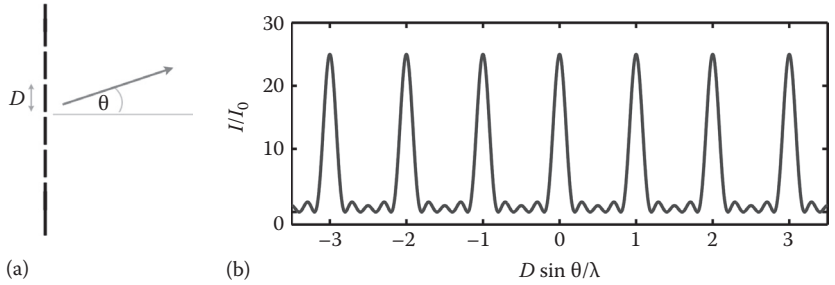

FIGURE 3.9 *N*-slit interference. (a) Geometry. Each slit is of negligible width and is separated from its neighbor by distance D. In the example drawn, $N = 5$. (b) The N-slit intensity function, $I = I_0 \dfrac{\sin^2\left(N\pi D\sin\theta/\lambda\right)}{\sin^2\left(\pi D\sin\theta/\lambda\right)}$, plotted for $N = 5$. Note that there are infinitely many large maxima located at angles for which $\sin(\theta) = m\lambda/D$, where m is any integer. Between each pair of these peaks are $N-2$ smaller local maxima and $N-1$ zeros. In this example, the zeros are located at $\sin(\theta) = \lambda/5D, 2\lambda/5D, 3\lambda/5D, 4/5D$.

where, for convenience, we have defined $\alpha \equiv (2\pi/\lambda)\, D \sin\theta$. Note that this is

$$\vec{E} = \vec{E}_0 \exp\left[j(ks - \omega t)\right]\left\{1 + \left(\exp\left[j\alpha\right]\right) + \left(\exp\left[j\alpha\right]\right)^2 + \left(\exp\left[j\alpha\right]\right)^3 + \cdots + \left(\exp\left[j\alpha\right]\right)^{(N-1)}\right\}.$$

The terms in the braces form a finite geometric series, since each term is equal to the preceding one times $e^{j\alpha}$. Therefore,

$$\vec{E} = \vec{E}_0 \exp\left[j(ks - \omega t)\right]\left\{\frac{1 - \exp(jN\alpha)}{1 - \exp(j\alpha)}\right\}.$$

We can simplify the expression in the parentheses by factoring out exponentials from the numerator and denominator:

$$\frac{1 - \exp(jN\alpha)}{1 - \exp(j\alpha)} = \frac{\exp\left(-jN\alpha/2\right)\left[\exp\left(jN\alpha/2\right) - \exp\left(-jN\alpha/2\right)\right]}{\exp\left(-j\alpha/2\right)\left[\exp\left(j\alpha/2\right) - \exp\left(-j\alpha/2\right)\right]} = \exp\left\{\frac{-j(N-1)\alpha}{2}\right\}\frac{\sin\left(N\alpha/2\right)}{\sin\left(\alpha/2\right)},$$

using the Euler relation, $\sin(x) = (1/2j)(\exp(jx) - \exp(-jx))$.

Therefore, $\vec{E} = \vec{E}_0 \exp\left[j(ks - \omega t)\right]\exp\left\{-\frac{j(N-1)\alpha}{2}\right\}\frac{\sin\left(N\alpha/2\right)}{\sin\left(\alpha/2\right)}$.

The intensity $I \propto \left|\vec{E}\right|^2$:

$$I = I_0 \frac{\sin^2\left(N\alpha/2\right)}{\sin^2\left(\alpha/2\right)}.$$

Explicitly writing the α's:

$$I = I_0 \frac{\sin^2\left(N\pi D \sin\theta/\lambda\right)}{\sin^2\left(\pi D \sin\theta/\lambda\right)}.$$

This is plotted in Figure 3.9b for $N = 5$.

Maxima and minima. We see that the *numerator* of $I(\theta)$ is zero when $N\pi D \sin\theta/\lambda = m\pi$, i.e., $D \sin\theta/\lambda = m/N$, where m is an integer—but note that *both* numerator and denominator are zero if m is an integer multiple of N. We see that the *denominator* is zero when $\pi D \sin\theta/\lambda = m'\pi$, i.e., $D \sin\theta/\lambda = m'$, where m' is an integer—in this case, however, the numerator must also be zero, since $N\pi D \sin\theta/\lambda = Nm'\pi$ and N is an integer. If both the numerator and denominator are zero $I \to I_0 N^2$. A more detailed summary of the locations of maxima and minima is a useful exercise for the reader.

As illustrated in the plot of $I(\theta)$ for $N = 5$ slits (Figure 3.9), there are large maxima separated in angle by $\sin\theta = \lambda/D$. The form of $I(\theta)$ reveals that this *angular spacing* between the peaks is independent of the number of slits. The *angular width* of the large peaks is approximately $\Delta \sin\theta \approx \lambda/ND$—half the distance in angle to the first local minimum—which gets sharper as we increase the number of slits. This is a very useful feature, as we will see shortly.

3.3.3 The Diffraction Grating as a Monochromator

Suppose we have a telescope that collects light from a star, and we want to measure the star's spectrum—i.e., the intensity as a function of wavelength, $I(\lambda)$. How can we do this? Our detector (like most good detectors, at least over some range of wavelengths) simply measures intensity, regardless of the wavelength of the light hitting it.

We can pass the light though an N-slit grating, or, equivalently, reflect it off a surface with N mirrors—a *diffraction grating*. How does this help? Light of wavelength λ_1 is deflected to angle λ_1/D. By this, we mean that the maximal intensity peak for light of this "color" is at the angle given by $\sin\theta_1 = \lambda_1/D$, and integer multiples, as in Section 3.3.2; typically, the angles involved are small, so $\sin\theta \approx \theta$. Light of wavelength λ_2 is deflected to angle λ_2/D, etc. Moving our detector to various positions on the screen and measuring the intensity as a function of *angle* on the screen reveals the intensity as a function of wavelength! (In other words $I(\lambda_1) = I(\theta_1)$, $I(\lambda_2) = I(\theta_2)$, etc.).

The *sharper* the diffraction peaks (high N), the *finer* the resolution in λ—see the end of the preceding section. The discovery (within the past ≈ 10 years) of planets outside our solar system—one of the most remarkable discoveries of recent history—used the approach outlined above to measure tiny shifts in stellar spectra due to the influence of the orbiting planets. The typical N of the diffraction gratings was around 100,000!

3.3.4 Single-Slit Interference

In our initial discussion of two-slit interference, we neglected the finite width of the diffraction grating. This finite width is important—just as waves from each slit interfere with one another, waves traversing various paths through a *single* slit will interfere with one another, and lead to diffraction. Fortunately, it is easy to analyze single-slit interference—it is simply the limit of the N-slit case discussed in Section 3.2.2 as $N \to \infty$, $D \to 0$, and the product $ND \to a$, where a is the width of the slit. The reader can verify that

$$I(\theta) = I_0\left(\frac{\sin\beta}{\beta}\right)^2,$$

where $\beta = \pi a \sin\theta\, \lambda^{-1}$, as plotted in Figure 3.10.

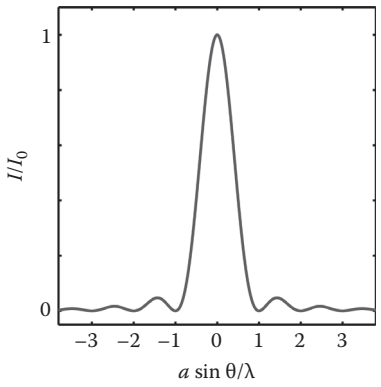

FIGURE 3.10 The intensity function of a single slit of width a. Note that the angular width of the peak is approximately λ/a.

3.3.5 Diffraction and Resolution

This single-slit diffraction pattern is exceptionally important. Any optical element—the pupil of your eye, a telescope mirror, a microscope lens, etc—is an aperture, and the $I(\theta)$ above describes how light travels through it. Why?

We have been considering light leaving an aperture, i.e., being "transmitted," and reaching a screen, where it is "received." But look carefully at Figure 3.5, 3.7, or 3.9—our wave interference scenarios. Our analysis did not invoke at all the *direction* the waves were traveling, only the path length difference between various paths. So we would get the same interference effects if light were *transmitted* from a point source at angle θ on the screen, passed through aperture(s), and were detected at the left.

Consider light from a point source (e.g., a star) located at the far-off "screen." We observe the point source by detecting the intensity passing through a single-slit aperture of width a (e.g., a telescope lens plus an intensity detector). We tilt the barrier containing our slit (e.g., our telescope), so that the angular position of the star of interest is $\theta_1 = 0$; this angle gives the maximum of the single slit intensity function, and we happily detect light from the star. We tilt the telescope; at the new $\theta_2 = 0$ there is no star; we see no light. We tilt further; at this third $\theta_3 = 0$ there is light again. "Aha!," we say, "We have seen two stars!"

Now suppose there were two stars very close to one another in angular position—let us say the difference in sin θ is just $0.1\lambda/a$. (We typically deal with small angles, by the way, so sin $\theta \approx \theta$.) Since the width of our interference function is $\approx \lambda/a$, no matter how precisely we point at one star, we will be detecting a sizeable fraction of the intensity of the other—*there is no way we can tell that we are looking at two stars rather than only one*!

The *angular limit of resolution*, often just referred to as the *resolution*, of our single-slit aperture—the minimum angular separation that two objects must have in order to be able to distinguish them—is $\theta_{res} \approx \lambda/a$, where a is the aperture size. (It is an "approximately equals" sign, because there are slightly different ways of defining criteria for distinguishability that will not concern us here; most commonly, one uses the "Rayleigh criterion" $\theta_{res} = 1.22\ \lambda/a$.) Note that smaller θ_{res} means that we can more finely distinguish objects—we can "see" better—and that this can be achieved by increasing the size of our aperture. This is why one builds big telescopes. (Big telescopes have another, unrelated, advantage: they collect more light.)

This issue of diffraction sets the fundamental limit on the performance of telescopes, microscopes, and other optical devices. Regarding microscopy, and all the diverse applications of it described, for example, in this book, the above angular description of resolution together with expressions for the focusing ability of lenses (Section 3.5 and Chapter 1) set a spatial limit for optical resolution. Roughly, objects separated by distance less than $\Delta x \approx \lambda$ cannot be resolved as separate entities. We will revisit the diffraction limit on resolution in Section 3.5.8.

The diffraction of light as it passes through an aperture maps the emission of an ideal pointlike source onto the observed pattern of intensity. For a one-dimensional slit, this mapping is given by the profile shown in Figure 3.10. In microscopy, one is interested in the analogous pattern caused by diffraction through the two-dimensional, typically circular, objective lens. One refers to the resulting intensity profile of a point source as the point spread function (PSF). In other words, the image of a point source (e.g., a fluorescent molecule) will not look like a point, but will look like the PSF. For an ideal aberration-free circular lens of radius a and focal length f (defined in Section 3.5), the PSF is given (Gu 1999) by

$$I(v) = I_0 \left[\frac{2J_1(v)}{v} \right]^2,$$

where

$v = \frac{2\pi}{\lambda} \frac{a}{f} r$, r is the radial coordinate in the imaging plane

I_0 is the central intensity

J_1 is a Bessel function of the first kind

This function is illustrated in Figure 3.11; note that the width of the intensity peak is roughly $\lambda/2$.

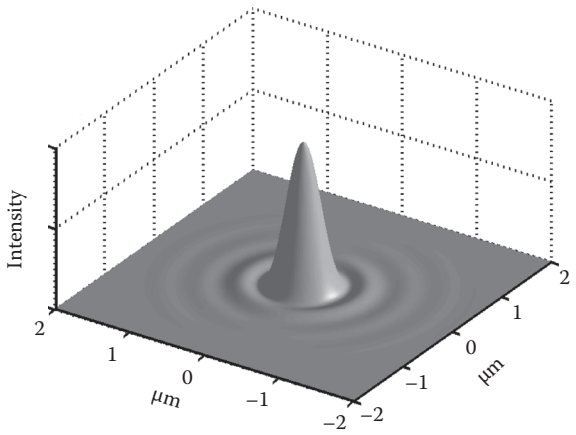

FIGURE 3.11 The PSF for an ideal circular aperture, plotted for $a/f = 0.7$ and $\lambda = 0.6\,\mu m$.

3.4 Refraction

When light travels from one medium to another, it may change direction. This phenomenon—familiar whenever we see the "bent" shape of a straw poking out of a glass of water—is known as *refraction*. (The light may also change its intensity at a boundary between media, which we will discuss in Section 3.6.) Refraction, like diffraction, is inherently a consequence of the wave nature of light, and our analysis below applies also to waves in water, sound waves, etc.

3.4.1 Snell's Law

The basic setup for issues of refraction is shown in Figure 3.12a: a ray of light crosses the boundary between two media, with indices of refraction n_1 and n_2, making angles θ_1 and θ_2 with respect to the normal in each medium, respectively. The question is: *How are* θ_1 *and* θ_2 *related?* The answer is crucial to the propagation of light and to the design of lenses and other optical elements.

The answer, as we will show, is that light obeys *Snell's law*:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2.$$

There are several ways to derive this result. We could directly examine the wave equations for electromagnetic fields and look for solutions consistent with the presence of a boundary between two media, but this would be both painful and unilluminating. There are, fortunately, simpler ways of thinking about light propagation.

3.4.2 Fermat's Principle

3.4.2.1 Minimal Time Paths and Snell's Law

A general principle describing wave motion was put forth by Pierre de Fermat in the seventeenth century. It is sometimes stated as "light travels from one point to another along the path that takes the minimal amount of time." This is not quite correct—we will fix it in a few paragraphs—but it is a good place to start. We will also return to justifying Fermat's principle shortly. First, let us use it to derive Snell's law.

Imagine you are on a beach, and someone in the ocean is drowning. You rush out to help, which requires both running on land and swimming in the water. You can run faster than you can swim. What path should you take? With a bit of thought, you will realize that a straight line between you and the drowning person is not the best idea—rather, you should reduce the length of the swim to minimize the overall time to your target. How much should you run and how much should you swim?

The same dilemma is encountered by our light beam, traveling from position A in a medium of index of refraction n_1 (your position on the beach, in the above analogy) to position B in a medium of n_2(the swimmer's position, in the water) in Figure 3.12b. The speed of light in medium 1 is $v_1 = c/n_1$ and in medium 2 is $v_2 = c/n_2$. Within each medium, the light travels in a straight line—itself a consequence of Fermat's principle, as you can convince yourself. There are many possible paths between A and B, as illustrated in the figure, that we can label based on the position x at which they cross the interface. One of these—let us call it the one that goes through position x_0—minimizes the total travel time. What is this path? (What is x_0?)

The travel time in medium 1, t_1, is the distance traveled in medium 1 divided by the speed in medium 1: $t_1 = \left(n_1/c\right)\sqrt{y_1^2 + x^2}$; similarly, the travel time in medium 2 is $t_2 = \left(n_2/c\right)\sqrt{y_2^2 + (L-x)^2}$. The total travel time is $t = t_1 = t_2$. To find the minimal time, we determine the x for which $\mathrm{d}t/\mathrm{d}x = 0$; call this x_0:

$$\frac{\mathrm{d}t}{\mathrm{d}x} = \frac{n_1}{c}\frac{x}{\sqrt{y_1^2 + x^2}} - \frac{n_2}{c}\frac{(L-x)}{\sqrt{y_2^2 + (L-x)^2}},$$

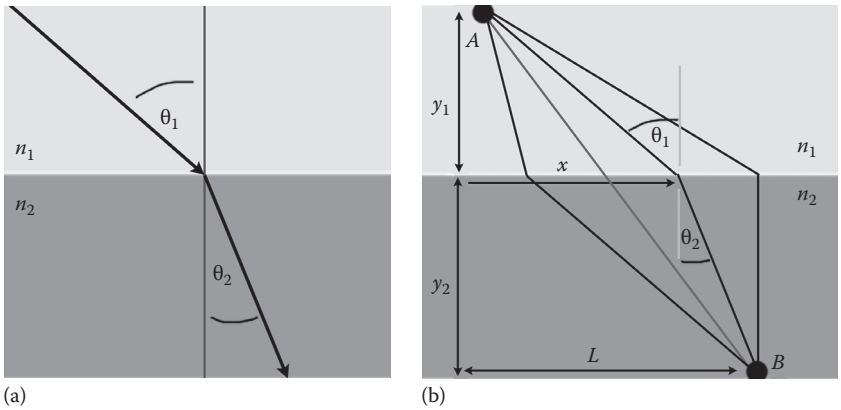

FIGURE 3.12 Refraction. (a) The path taken by light traveling between two media bends at the interface; the relation between the angles θ_1 and θ_2 depends on the indices of refraction of the two materials, and is given by Snell's law. (b) Light traveling from point A in medium 1 to point B in medium 2 can follow infinitely many possible paths, four of which are illustrated here. Fermat's principle states that the actual path the light takes is that for which the total travel time is minimal (actually extremal, as discussed in the text), from which Snell's law follows.

$$\left.\frac{dt}{dx}\right|_{x=x_0} = 0 \rightarrow n_1 \frac{x_0}{\sqrt{y_1^2 + x_0^2}} = n_2 \frac{(L - x_0)}{\sqrt{y_2^2 + (L - x_0)^2}}.$$

(To show that x_0 is a minimum, we should also examine the second derivative of t, but as we will see later, it does not actually matter if x_0 is the site of a minimum or a maximum. Furthermore, we can intuit from the form of $t(x)$ that this extremum is, in fact, a minimum.)

Note from geometry that

$$\sin\theta_1 = \frac{x_0}{\sqrt{y_1^2 + x_0^2}},$$

$$\sin\theta_2 = \frac{(L - x_0)}{\sqrt{y_2^2 + (L - x_0)^2}}.$$

Therefore, the above condition becomes

$$n_1 \sin\theta_1 = n_2 \sin\theta_2.$$

We have shown that when the above condition is met, the travel time for light propagation is minimized. This is Snell's law.

We can also use Fermat's principle to derive Snell's *law of reflection*, which states that the *reflected* ray makes the same angle with the interface as the incident ray.

3.4.2.2 Explaining Fermat's Principle

Now let us explain Fermat's principle. Suppose light travels along many paths, all of which interfere with one another. Paths for which the phase difference is near zero will constructively interfere. Consider the minimal time path. As we saw in our derivation of Snell's law above, this is the path for which $\sum n_i d_i$ is minimal, where the sum runs over all the segments being considered (two segments in the above example), and d_i is the length of segment i. From Section 3.2, minimizing $\sum n_i d_i$ is the equivalent to minimizing the *phase* traversed by the wave along the path. Therefore, the minimal time path is also the path of minimal phase and is also the path of minimal $\sum n_i d_i$—all these statements are equivalent. This sum $\sum n_i d_i$ is more properly written as an integral and is called the optical path length (OPL): $\mathrm{OPL} = \int_A^B n(x)\,dx$.

Why should the path of minimal OPL be the path light takes? Let us call this path P. By construction, $\left.\frac{d(\mathrm{OPL})}{ds}\right|_P$, is zero, where s indicates any variable that characterizes the paths. Therefore, nearby paths are similar in phase and so constructively interfere. Consider a path for which $\frac{d(\mathrm{OPL})}{ds}$ is *not* zero—moving to a slightly different path, the OPL can change appreciably, perhaps higher in one direction, lower in another, etc., and so we would not expect constructive interference.

You may be thinking: the minimal OPL path is not the only one for which we can guarantee constructive interference. What about the *maximal* OPL path? This too provides constructive interference. And so the proper formulation of Fermat's principle is that *light travels along paths of extremal optical path length*. Typically, these are minimal OPL paths; but, in certain geometries, they can be maximal OPL paths as well.

3.4.3 Total Internal Reflection

Let us look more carefully at Snell's law. What if $n_1 > n_2$, and θ_1 is large, so that $(n_1/n_2)\sin\theta_1 > 1$? What θ_2 can satisfy Snell's law: $n_1 \sin\theta_1 = n_2 \sin\theta_2$? None. This means that *there can be no wave transmitted to medium 2*; the light from medium 1 is totally reflected at the interface, a condition known as *total internal reflection*.

Fiber optics, which underpin much of modern communication, work because of total internal reflection. Consider a glass fiber ($n = 1.5$) surrounded by air ($n = 1.0$). We want the light to travel along the fiber and not leak out into the air. This is automatically enforced by total internal reflection due to the higher index of refraction of the glass and the large incident angles of light traveling along the fiber (as long as the fiber is not severely bent). In a fiber optic cable, light can propagate for kilometers with losses of a fraction of a percent!

A careful treatment of electromagnetic fields would show that the light intensity in medium 2 is not exactly zero. Rather, an "evanescent wave" decays to zero over a short distance, comparable to λ, in medium 2. This attribute finds applications in biophysical imaging. In total internal reflection fluorescence microscopy, excitation of fluorescent probes by the evanescent wave allows discrimination of molecules near interface despite the presence of large concentrations of probes in the bulk (Axelrod et al. 1984, Axelrod 2001, Groves et al. 2009).

3.5 Lenses

We often wish to collect and reshape electromagnetic wavefronts to create images of objects. *Lenses* are powerful tools for achieving these goals and are obviously very useful, forming the essential imaging elements of telescopes, microscopes, cameras, your eyes, and many other devices. The "ideal" shape of a lens surface is generally some non-spherical conic section (hyperbola, parabola, etc.), but, in practice, spherical lenses are typically used, since they are much easier to make than aspheric (non-spherical) lenses. Typically, one uses spherical lenses and then corrects for their aberrations (nonideal behavior), e.g., by using combinations of lenses. We will briefly explore lenses.

3.5.1 A Spherical Interface

Consider a point source emitting spherical waves from point S, in a medium of index of refraction n_1 (see Figure 3.13). Can

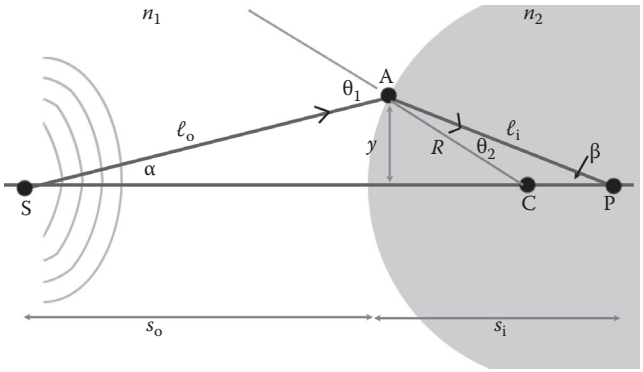

FIGURE 3.13 A spherical interface. C, S, A, and P refer to particular points—the center of the spherical interface, the object point, the point at which the ray drawn hits the interface, and the image point, respectively. Italicized letters refer to distances. Greek letters refer to angles—note that $\alpha = \angle ASC$ and $\beta = \angle CPA$.

we construct a spherical interface of radius R that focuses the emitted light to point P, regardless of where it hits the interface? What should R be? Point P is embedded in a medium of index of refraction n_2; we are considering the shape of the interface between media 1 and 2. Consider $n_2 > n_1$, so that the rays from S will be refracted "inwards."

In Figure 3.13, Point C is the center of the sphere of radius R. The distance between the "object" point, S, and the interface is s_o, and the distance between the "image" point, P, and the interface is s_i. The angles that the incident and reflected rays make with respect to the normal to the interface are θ_1 and θ_2. As usual θ_1 and θ_2 are related by Snell's law: $n_1 \sin\theta_1 = n_2 \sin\theta_2$. We can relate θ_2 to β via the law of sines: $(\sin\beta/R) = (\sin\theta_2/(s_i - R))$. Relating θ_1 to α is not quite as transparent; first note that $\angle SAC = \pi - \theta_1$, so $\sin(\angle SAC) = \sin(\pi - \theta_1) = \sin\theta_1$, and then apply the law of sines to ΔSAC to get $(\sin\alpha/R) = (\sin\theta_1/(R + s_o))$. Inserting all this into Snell's law:

$$n_1 \frac{R + s_o}{R} \sin\alpha = n_2 \frac{s_i - R}{R} \sin\beta,$$

$$\text{i.e., } n_1 (R + s_o) \sin\alpha = n_2 (s_i - R) \sin\beta.$$

$$\text{More geometry: } \sin\alpha = \frac{y}{l_o} = \frac{y}{\sqrt{s_o^2 + y^2}} \text{ and } \sin\beta = \frac{y}{l_i} = \frac{y}{\sqrt{s_i^2 + y^2}}.$$

$$\text{From which } n_1 (R + s_o) s_i \sqrt{1 + \left(\frac{y}{s_i}\right)^2} = n_2 (s_i - R) s_o \sqrt{1 + \left(\frac{y}{s_o}\right)^2}$$

We have derived a relation that must hold for focusing at P to occur. In other words, we know what R we need—the R that satisfies the above expression. Unfortunately, it depends on y, the position at which our ray hits the interface! Therefore, different rays will not focus to the same image spot.

3.5.2 A Spherical Interface—The Paraxial Regime

What we have shown is that a truly spherical interface will not serve as an ideal lens. There is a way out of this, however, which is to limit ourselves to the paraxial regime, meaning that we consider only light that is nearly parallel with the optical axis, *SP*. In other words, we consider small α and β. Therefore, y/s_o and y/s_i are small, allowing us to neglect them in the above equation: $n_1(R + s_o)\, s_i \approx n_2\, (s_i - R)\, s_o$, from which $(n_1/s_o) + (n_2/s_i) = (n_2 - n_1)/R$. A simple, useful relation! (By the way, we could also have derived this directly from Fermat's principle, by determining the R for which SAP is an extremal path for any A.)

Should we be bothered by limiting ourselves to the paraxial case? Yes and no. In practice one *does* try to design optical systems such that beams are close to the center of spherical lens elements or, equivalently, to have one's image and object distances be large compared to the size of the lens. If one does this, the above relation works very well. In practice, one works in the paraxial regime and applies additional corrections if necessary. We will continue limiting ourselves to the paraxial regime.

3.5.3 Focal Points

If R, n_1, and n_2 are fixed, decreasing s_o means that s_i increases (and vice versa), from the above boxed relation. Let us increase s_i until $s_i \to \infty$; in other words, parallel rays emerge from the interface; what is s_o? From above: $(n_1/s_o) + (n_2/\infty) = (n_2 - n_1)/R$, therefore $s_o = (n_1/(n_2 - n_1))R$—an object at this distance focuses "to infinity." We will call this distance the *object focal length*, $f_o \equiv (n_1/(n_2 - n_1))R$. The spherical waves from the point source turn into plane waves.

The same holds if we do not consider a "semi-infinite" medium on the right, but rather a finite lens with a spherical surface at the left and a flat surface at the right—a *planoconvex lens* (see Figure 3.14). Note that since the right edge is flat, all rays are normal to it, and there is no "bending" of the rays due to refraction.

We can of course consider the opposite situation, in which plane waves (parallel rays from $s_o = \infty$) are focused to an image at some s_i. This particular s_i is denoted f_i, the *image focal length*.

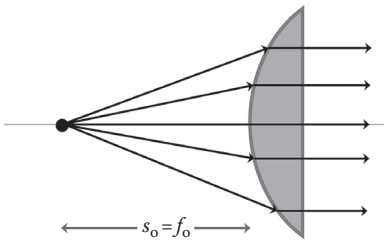

FIGURE 3.14 A planoconvex lens. Light emanating from a source located at the object focal length is focused to an image distance of infinity (i.e., the rays become parallel).

3.5.4 Real and Virtual Images

Solving the lens equation above for s_i, we have

$$s_i = n_2\left(\frac{n_2 - n_1}{R} - \frac{n_1}{s_o}\right)^{-1} = \frac{n_2}{n_1}\left(\frac{1}{f_o} - \frac{1}{s_o}\right)^{-1}.$$

If $s_o > f_o$, then $s_i > 0$, and point P is to the right of the interface. The rays from S converge at P. To an observer at the right, *it looks as if light is emanating from point* P. We have what is called a *real image* at P (see Figure 3.15a). If, for example, we put a power meter at P, we detect a high degree of power due to the focused light.

If $s_o < f_o$, then $s_i < 0$, and point P is to the left of the interface. The rays do not actually hit point P, but they *appear* to an observer at the right as if they are emanating from P (see Figure 3.15b). We have what is called a *virtual image* at P. If, for example, we put a power meter at P, we *do not* detect a high-intensity focused spot, since there is no "spot" there.

3.5.5 Concave Lenses

The same analysis works for concave lenses, but we treat R as *negative* ($R < 0$). Since

$$\frac{n_1}{s_o} + \frac{n_2}{s_i} = \frac{n_2 - n_1}{R}, \text{ if } n_2 > n_1 \text{ then } s_i < 0\text{—we have a virtual image.}$$

3.5.6 Thin Lenses

Let us glue one lens of radius of curvature R_1 onto another of R_2 (see Figure 3.16). We will consider thin lenses and so neglect the lens thickness d (i.e., we are assuming d is smaller than other lengths involved).

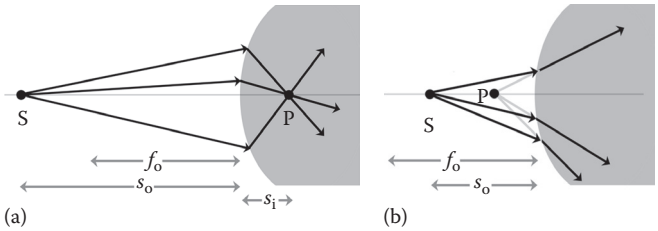

FIGURE 3.15 (a) Real and (b) virtual images. (a) Light emanates from P. (b) Light looks to an observer like it is emanating from point P located to the left of the interface.

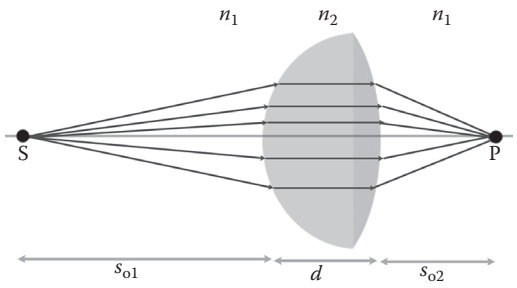

FIGURE 3.16 Focusing light with a thin lens (imagine d is small).

The object and image lengths for "lens 1" (the left half of the lens) are related by

$$\frac{n_1}{s_{o1}} + \frac{n_2}{s_{i1}} = \frac{n_2 - n_1}{R_1}.$$

The image of lens 1 provides the "object" for lens 2. Therefore, $s_{o2} = -s_{i1} + d \approx -s_{i1}$, where the negative sign arises, because, as defined above, a positive image length and a positive object length lie in opposite directions. Considering lens 2:

$$\frac{n_2}{-s_{i1}} + \frac{n_1}{s_{i2}} = \frac{n_1 - n_2}{R_2},$$

where we keep track of which index of refraction is which.

We need to adopt a consistent set of sign conventions for the radii. As noted above, a convex "left" lens has $R > 0$, and a concave "left" lens has $R < 0$. For the right side lens, these are switched. Returning to our thin lens, adding the two expressions above:

$$\frac{n_1}{s_{o1}} + \frac{n_1}{s_{i2}} = (n_2 - n_1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right).$$

For a thin lens in air, $n_1 \approx 1$; $n_2 = n_{\text{lens}}$, giving us the *thin lens equation*, or *Lensmaker's formula*:

$$\frac{1}{s_o} + \frac{1}{s_i} = (n_{\text{lens}} - 1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right).$$

The *focal length*, f, is given either by s_o or $s_i \to \infty$ (it does not matter which):

$$\frac{1}{f} = (n_{\text{lens}} - 1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right).$$

We can then write the thin lens equation as $(1/s_o) + (1/s_i) = 1/f$, also known as the Gaussian Lens Formula. This is one the most important relations for the design of optical systems.

For example: Consider parallel rays incident on the flat side of a glass ($n = 1.5$) planoconvex lens with a radius of curvature of 50 mm. Where will these rays be focused to? Answer: $R_1 = \infty$, $R_2 = -50$ mm. $1/f = (1.5 - 1)\,(1/50\text{ mm})$, so $f = 100$ mm, $s_i = 100$ mm. The rays will focus to a point 100 mm beyond the curved side of the lens.

3.5.7 Magnification

Lenses magnify objects. The magnification can be >1 or <1. See Figure 3.17 depicting a thin lens, which is magnifying an extended object (i.e., not a point source)—in this case, a pear.

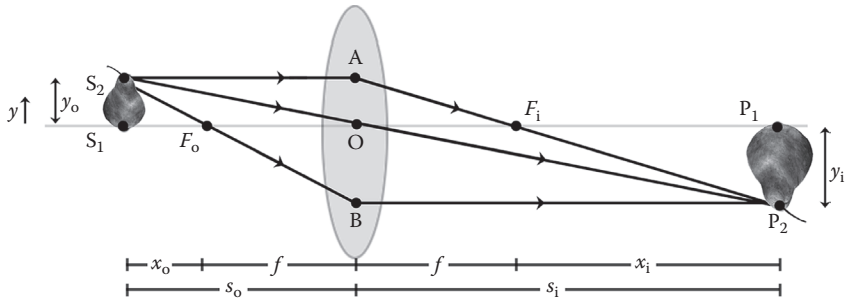

FIGURE 3.17 Magnification of an image by a lens. Rays emanating from point S_1, on the optical axis, are focused to point P_1 (not drawn). Rays from S_2 are focused to P_2.

Points F_o and F_i are each a distance f, the focal length, from the lens. Consider light emanating from the top of the pear. The ray that goes through F_o will emerge from the lens parallel to the axis (think about why this is). The ray the leaves the pear parallel to the axis will go through F_i. The ray that goes through the center of the lens will be undeflected in the thin lens limit (Hecht 2002).

The *magnification*, M_T, is defined to be the height of the image relative to the height of the object—i.e., $M_T \equiv y_i/y_o$. Triangle S_1S_2O is similar to triangle P_1P_2O, so $y_o/s_o = -y_i/s_i$, so $M_T \equiv -s_i/s_o$; the negative sign shows that the image is *inverted*.

Triangle AOF_i is similar to triangle $P_1P_2F_i$, so $y_o/f = |y_i|/(s_i - f)$.

Triangle BOF_o is similar to triangle $S_1S_2F_o$, so $y_o/(s_o - f) = |y_i|/f$. Combining these, $(f/(s_o - f)) = ((s_i - f)/f)$. Note that $s_i - f = x_i$ (see figure). Using the last similar triangle relation again, $(f/(s_o - f)) = (-y_i/y_o)$. And so, $M_T = (y_i/y_o) = -(x_i/f)$. We could also have written: $M_T = -(f/x_o)$.

As the object distance x_o is lowered, the magnification increases. The reader may think about what happens if $x_o < 0$, i.e., the object is closer than the focal point. Drawing rays, convince yourself that the lens cannot form an image of the object.

Another important aspect of lenses that follows from the diagrams above, and for which it is useful to develop an intuition: Parallel rays that are also parallel to the optical axis are focused to the focal point on the optical axis. Parallel rays that are *tilted* with respect to the optical axis are also focused to a point at distance f, but *off* the optical axis; such points define the *focal plane*. The reader may wish to draw such rays.

3.5.8 Resolution

When considering single-slit diffraction in Section 3.3, we realized that the angular resolution of a device is given by $\theta_{min} \approx \lambda/a$, where λ is the wavelength of light and a is the diameter of the imaging aperture. Two objects must have an angular separation of at least θ_{min} if they are to be resolved as separate objects. Using lenses to magnify objects, this angular resolution criterion still holds. Moreover, the fact that the object distance cannot be closer than the focal length turns our resolution relation into a distance criterion. We will briefly sketch this:

Consider two objects separated in position by Δy at a distance s from a lens (see Figure 3.18). For the two to be resolvable, we need $\theta > \lambda/a$, where $\theta \approx \Delta y/s$. Therefore, we need $\Delta y > s\lambda/a$. Since $s > f$, $f = (n_1/(n_2 - n_1))R$, and $R > a$, we can write $s > (n_1/(n_2 - n_1))a$. Combining the two inequalities: $\Delta y > (n_1/(n_2 - n_1))\lambda$. The numerical factor $n_1/(n_2 - n_1) \approx 1$ in our rough treatment. Therefore, our minimum resolvable spatial separation is $\Delta y_{min} \approx \lambda$. We cannot resolve objects smaller than (approximately) the wavelength of light.

More precise statements of optical resolvability can be constructed. Typically, one invokes the Abbe criterion that the minimal $\Delta y = \lambda/(2n\sin\theta)$, where n is the index of refraction of the medium and θ is the maximum angle over which light is collected. The value of $\sin\theta$ is bounded by 1, so at best $\Delta y = \lambda/(2n)$. For $\lambda = 400$ nm (blue) light in water ($n = 1.3$), the theoretical resolution limit is about 150 nm. (In practice, any aberrations or imperfections further reduce the resolution.) Hence, the wavelengths of visible light set a limit of roughly a few hundred nanometers as the minimal size of resolvable structures—smaller, for example, than cells but far larger than the characteristic sizes of proteins or small molecules.

It is important to keep in mind that the issues of resolution discussed above govern the discrimination of two (or more) objects. If one knows that only a *single* point source contributes to an image, giving an intensity profile like that of Figure 3.11, for example, the center of this profile can be determined to arbitrarily high precision (in practice, a few nanometers typically). A few of the many applications of this principle are illustrated in (Crocker

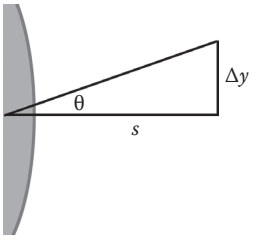

FIGURE 3.18 Schematic, for considering spatial resolution.

and Grier 1996, Weihs et al. 2006, Crocker and Hoffman 2007, Roichman et al. 2008, Kong and Parthasarathy 2009).

Despite the fundamental nature of the diffraction-limited resolution, the past decade or so has seen the birth of several very clever techniques for surmounting it, using interferometry, nonlinear optical processes, or single-molecule imaging (e.g., Betzig et al. 2006, Bates et al. 2007, Hell 2007, Abbott 2009) to yield optical information at scales an order of magnitude smaller than what was traditionally thought possible.

3.6 Reflection and Transmission (Fresnel's Equations)

The law of reflection ($\theta_r = \theta_i$, where *r* and *i* refer to reflected and incident rays; Figure 3.19a) and Snell's law ($n_i \sin\theta_i = n_t \sin\theta_t$, where *t* refers to the transmitted ray) give the *directions* of reflected and transmitted rays at boundaries. What are the *amplitudes* of the electromagnetic waves? In other words, *how much* light is reflected and transmitted? Similar questions arise when considering other sorts of waves hitting boundaries—for example, waves on strings, incident at an interface between two media with different propagation speeds. In all these situations, transmission and reflection are analyzed by considering the boundary conditions imposed by the junction.

To consider the general case of a plane electromagnetic wave hitting a surface at some angle θ_i (with respect to the normal), we will have to separately consider the components with electric field perpendicular and parallel to the *plane of incidence*. (The incident, reflected, and transmitted rays all lie in and define the plane of incidence, "POI," which also includes the normal to the surface.)

Recall from Section 3.2.4.2 some properties of electromagnetic waves:

- The electric and magnetic field vectors of an electromagnetic wave are perpendicular to each other.
- $\vec{E} \times \vec{B}$ points along $\vec{k}$, the wavevector, i.e., along the direction of propagation.
- The field amplitudes are related by $|\vec{E}| = v|\vec{B}|$, where $v = c/n$ is the wave speed.

The boundary conditions that govern electric and magnetic fields at the interface between media are

i. The tangential (i.e., parallel to the interface) components of the electric field, $\vec{E}$ are continuous across the boundary.
ii. The tangential (i.e., parallel to the interface) components of $\vec{B}/\mu$, where μ is the magnetic permeability of the medium, are continuous across the boundary.

Let us consider the two cases.

3.6.1 Case I: $\vec{E}$ Perpendicular to the Plane of Incidence

Note that a circle with a dot in it indicates a vector that points out of the page towards you (see Figure 3.19b). The electric field vectors are completely tangential to the interface. The magnetic field vectors are not. Applying boundary condition (i) to the amplitudes (E_0) of the electric fields,

$$E_{0i} + E_{0r} = E_{0t}$$

Applying boundary condition (ii) to the amplitudes (B_0) of the magnetic fields, $-\frac{B_{0i}}{\mu_i}\cos\theta_i + \frac{B_{0r}}{\mu_i}\cos\theta_r = -\frac{B_{0t}}{\mu_t}\cos\theta_t$ (see Figure 3.19 to understand the signs).

Using $B_0 = E_0/v$ (from above), $v_i = v_r$ (since they are in the same media), $\theta_i = \theta_r$ (law of reflection), and $v_i = c/n_i$, we can write the above relation as

$$\frac{n_i}{\mu_i}\left(E_{0i} - E_{0r}\right)\cos\theta_i = \frac{n_t}{\mu_t} E_{0t}\cos\theta_t.$$

Combining this with the boundary condition (i) equation above, substituting to eliminate E_{0t}, we can solve for the ratio of the reflected wave amplitude to the incident wave amplitude:

$$\left(\frac{E_{0r}}{E_{0i}}\right)_\perp = \frac{n_i\mu_i^{-1}\cos\theta_i - n_t\mu_t^{-1}\cos\theta_t}{n_i\mu_i^{-1}\cos\theta_i + n_i\mu_i^{-1}\cos\theta_t}.$$

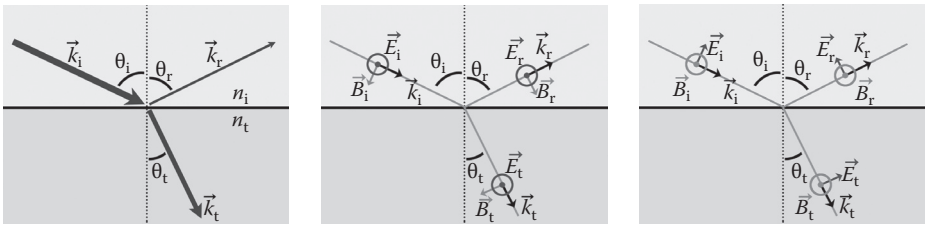

FIGURE 3.19 Reflection and refraction at an interface. The incident wave (wavevector $\vec{k}_i$) is reflected (wavevector $\vec{k}_r$) and transmitted (wavevector $\vec{k}_t$). Both the angles of the reflected and transmitted waves and their amplitudes are determined by the dielectric properties of the materials that comprise the interface. (a) Electric and magnetic field vectors for light polarized with $\vec{E}$ perpendicular to the plane of incidence. (b) Electric and magnetic field vectors for light polarized with $\vec{E}$ parallel to the plane of incidence.

Similarly solving instead for the ratio of the transmitted wave amplitude to the incident wave amplitude,

$$\left(\frac{E_{0t}}{E_{0i}}\right)_{\perp} = \frac{2n_i\mu_i^{-1}\cos\theta_i}{n_i\mu_i^{-1}\cos\theta_i + n_t\mu_t^{-1}\cos\theta_t}.$$

Typically, one deals with nonmagnetic materials: $\mu \approx \mu_0$, the permeability of free space. The above equations simplify, yielding two of the four *Fresnel equations*, for the amplitude reflection coefficient, $r_\perp$ and the amplitude transmission coefficient, $t_\perp$.

$$r_\perp = \left(\frac{E_{0r}}{E_{0i}}\right)_{\perp} = \frac{n_i\cos\theta_i - n_t\cos\theta_t}{n_i\cos\theta_i + n_t\cos\theta_t},$$

$$t_\perp = \left(\frac{E_{0t}}{E_{0i}}\right)_{\perp} = \frac{2n_i\cos\theta_i}{n_i\cos\theta_i + n_t\cos\theta_t}.$$

3.6.2 Case II: $\vec{E}$ Parallel to the Plane of Incidence

Applying the boundary conditions to this geometry leads to (see Figure 3.19c):

$$\left(\frac{E_{0r}}{E_{0i}}\right)_{\parallel} = \frac{n_t\mu_t^{-1}\cos\theta_i - n_i\mu_i^{-1}\cos\theta_t}{n_i\mu_i^{-1}\cos\theta_i + n_t\mu_t^{-1}\cos\theta_t},$$

and

$$\left(\frac{E_{0t}}{E_{0i}}\right)_{\perp} = \frac{2n_i\mu_i^{-1}\cos\theta_i}{n_i\mu_i^{-1}\cos\theta_t + n_t\mu_t^{-1}\cos\theta_i}.$$

For typical nonmagnetic media, we get the other two *Fresnel equations*:

$$r_\parallel = \left(\frac{E_{0r}}{E_{0i}}\right)_{\parallel} = \frac{n_t\cos\theta_i - n_i\cos\theta_t}{n_i\cos\theta_t + n_t\cos\theta_i},$$

$$t_\parallel = \left(\frac{E_{0t}}{E_{0i}}\right)_{\parallel} = \frac{2n_i\cos\theta_i}{n_i\cos\theta_t + n_t\cos\theta_i}.$$

3.6.3 Brewster's Angle

Let us plot $r_\perp$ and $r_\parallel$ as a function of θ_i for light incident from air ($n_i = 1$) to water ($n_t = 1.33$)—see Figure 3.20. We notice something very interesting: a particular θ_i for which the reflection coefficient is *zero* for light with its electric field parallel the plane of incidence. There is no such angle for the perpendicular polarization.

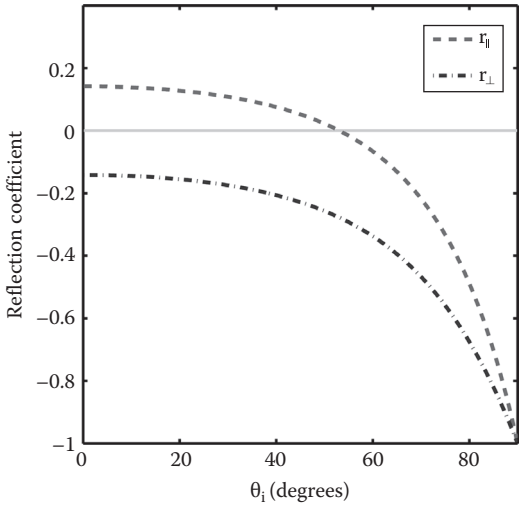

FIGURE 3.20 Fresnel coefficients $r_\perp$ and $r_\parallel$ for light incident from air to water, as a function of incidence angle.

The reader can work through the algebra and show that if $n_t > n_i$, $r_\parallel = 0$ at one particular incident angle, θ_i. This angle is called *Brewster's angle*, θ_p, and is given by $\tan\theta_p = n_t/n_i$. All the parallel-polarized light is transmitted. What about the perpendicular polarization? One can show that there is *no* angle that gives $r_\perp = 0$. Therefore, shining randomly polarized light incident at the Brewster angle, the reflected light is completely polarized with its electric field perpendicular to the plane of incidence. This property, together with a desk lamp, a sink, and a large bowl, once saved your author from misfortune, as he found himself with a much-needed linear polarizer whose transmission axis was unlabeled. The reader can test his or her mystery-solving skills by figuring out how he determined the polarizer's axis and thereby saved the day.

The polarization-dependence of reflection at interfaces also underpins Brewster angle microscopy, in which an interface is imaged with parallel-polarized light incident at the Brewster angle of the two media. The presence of interfacial molecules distinct from those of the two media—for example, lipids organized at an air–water interface—alters the local index of refraction, leading to a nonzero reflection coefficient. The intensity of the reflected light therefore provides a sensitive measure of interfacial molecular organization.

3.7 Concluding Remarks

In the preceding pages, we have explored the basic elements of optics. All of these topics can be explored much further, uncovering still more depth and beauty than we have been able to sketch, and also illuminating applications of great importance to science and technology. As will be evident throughout this book, these two aspects of optics—its formal elegance and its practical utility—are intertwined. Demands from fields as diverse as biological imaging, astronomy, and telecommunications drive the search for deeper insights into the behavior of light. Conversely,

explorations of the intricacies of electromagnetic wave propagation have yielded, and will undoubtedly continue to yield, remarkable new tools.

References

Abbott, A. 2009. Microscopic marvels: The glorious resolution. *Nature* 459: 638–639.

Axelrod, D. 2001. Total internal reflection fluorescence microscopy in cell biology. *Traffic* 2: 764–774.

Axelrod, D., Burghardt, T. P., and Thompson, N. L. 1984. Total internal-reflection fluorescence. *Annu. Rev. Biophys. Bioeng*. 13: 247–268.

Bates, M., Huang, B., Dempsey, G. T., and Zhuang, X. 2007. Multicolor super-resolution imaging with photo-switchable fluorescent probes. *Science* 317: 1749–1753.

Betzig, E., Patterson, G. H., Sougrat, R. et al. 2006. Imaging intracellular fluorescent proteins at nanometer resolution. *Science* 313: 1642–1645.

Born, M. and Wolf, E. 1997. *Principles of Optics. Electromagnetic Theory of Propagation, Interference and Diffraction of Light*, 6th edn. Cambridge, U.K.: Cambridge University Press.

Crocker, J. C. and Grier, D. G. 1996. Methods of digital video microscopy for colloidal studies. *J. Coll. Interf. Sci*. 179: 298–310.

Crocker, J. C. and Hoffman, B. D. 2007. Multiple-particle tracking and two-point microrheology in cells. *Methods Cell. Biol*. 83: 141–178.

Groves, J. T., Parthasarathy, R., and Forstner, M. B. 2009. Fluorescence imaging of membrane dynamics. *Annu. Rev. Biomed. Eng*. 10: 311–338.

Gu, M. 1999. *Advanced Optical Imaging Theory*. Berlin, Germany: Springer.

Hecht, E. 2002. *Optics*, 4th edn. San Francisco, CA: Pearson Addison Wesley.

Hell, S. W. 2007. Far-field optical nanoscopy. *Science* 316: 1153–1158.

Kong, Y. and Parthasarathy, R. 2009. Modulation of attractive colloidal interactions by lipid membrane functionalization. *Soft Matter* 5: 2027–2029.

Roichman, Y., Sun, B., Roichman, Y., Amato-Grill, J., and Grier, D. G. 2008. Optical forces arising from phase gradients. *Phys. Rev. Lett*. 100: 013602–013604.

Weihs, D., Mason, T. G., and Teitell, M. A. 2006. Biomicrorheology: A frontier in microrheology. *Biophys. J*. 91: 4296–305.

4

Light Sources, Detectors, and Irradiation Guidelines

Carlo Amadeo Alonzo
Massachusetts General Hospital

Malte C. Gather
Massachusetts General Hospital

Jeon Woong Kang
Massachusetts General Hospital

Giuliano Scarcelli
Massachusetts General Hospital

Seok-Hyun Yun
Massachusetts General Hospital

4.1 Introduction

Biomedical optics is an interdisciplinary study that demands proficiency across a broad range of topics, stretching from fundamental concepts in biology and biochemistry to technical instrumentation in optics. The latter can sometimes be quite intimidating itself, particularly when one considers the wide diversity of devices associated with modern techniques and applications in biomedical optics, as seen in later chapters of this book. It is easier to make sense of various instrumentation schemes when guided by a basic understanding of the properties and operating principles of devices involved in producing and detecting light. Light sources and light detectors represent the beginning and end, respectively, of any optical system used to study biological structures or mechanisms. Different types of these devices provide access to the different properties of light that, in turn, reveal different aspects of the biological cells and tissues under study.

Section 4.2 briefly describes some nonlaser light sources: high-pressure arc lamps, low-pressure vapor lamps, incandescent lamps, and light-emitting diodes (LEDs). Greater emphasis is placed on laser light sources as these play a more significant role in state-of-art biomedical optics techniques. Some basic properties that make laser radiation particularly useful in biomedical optics applications are discussed. The fundamental components of a laser are explained in the terms of their function during laser operation. As much progress in biomedical optics has been enabled by nanosecond- to femtosecond-duration laser pulses, some basic concepts of pulsed laser operation are also introduced. Finally, four types of lasers—gas, solid-state, dye, and semiconductor—are described and differentiated. Some specific examples of each type are also enumerated in the context of relevant biomedical applications.

Section 4.3 offers a complementary discussion of devices used to measure the intensity and distribution of light. Photodiodes, photomultiplier tubes (PMTs), and arrayed detectors such as charge coupled devices (CCDs) lie at the heart of all schemes to measure properties of light, even in more complex instruments, such as spectrometers and streak cameras. Each of these types of detectors is described in terms of their respective operating principles, unique characteristics, and typical applications.

Section 4.4 describes how exposure to optical radiation can trigger adverse effects to the cells and tissue. Depending on optical frequency, power, and duration of exposure, these range from photochemical, thermal, and thermoacoustic damage in tissues, to photobleaching, photodamage, and phototoxicity in cells. Such phenomena are discussed in order to provide a guide to safe and effective use of optical radiation in biomedicine.

4.2 Light Sources

4.2.1 Introduction

Light is an electromagnetic wave. Although the region of major concern in optics extends from the infrared, across the visible, to the ultraviolet, it is only a small portion of the vast electromagnetic spectrum, as shown in Figure 4.1. While all electromagnetic waves are of the same fundamental physical phenomenon, the particular time, length, and energy scale of each spectral

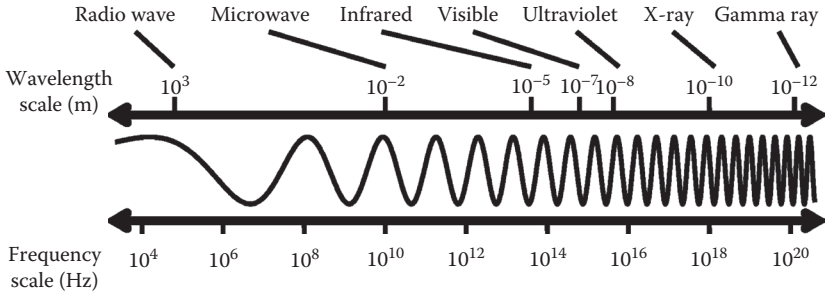

FIGURE 4.1 Electromagnetic spectrum. The upper scale presents a representative wavelength scale for each class of electromagnetic (EM) wave. A corresponding frequency scale is drawn below for reference. It should be noted that there are no specific boundaries between classes of EM waves. The most well-defined region would be for visible light, usually considered between 400 and 750 nm in wavelength. Otherwise, there is no particular wavelength that separates infrared from microwave radiation for example.

band lead to some specific properties that differentiate "light" from the shorter-wavelength x-rays and gamma rays, as well as the longer microwaves and radio waves, when propagation through and interaction with biological tissues are considered.

4.2.1.1 X-Rays and Gamma Rays

The application of x-rays and gamma rays in medicine is associated with the field of radiology and nuclear medicine. Short wavelengths ($<10^{-9}$ m) make these rays less prone to diffraction, yielding a ballistic trajectory through biological tissues—they behave more like streams of particles rather than propagating waves. The corresponding high oscillation frequencies ($>10^{17}$ Hz) also make them deeply penetrating, with minimal absorption and scattering in less dense materials. In fact, the propensity of x-rays to pass through soft tissues, but with different transmission coefficients, is what makes them so useful for diagnostic imaging. X-ray projection and computed tomography (CT) can visualize human anatomy using the tissue-dependent transmittance as contrast. However, when high-frequency radiation does interact with tissues, it is strongly ionizing—stripping away electrons from the constituent atoms and molecules. For gamma rays in particular, this destructive capacity in tissues is utilized in applications such as antimicrobial sterilization and the treatment of cancerous tumors. Gamma rays, at very low radiation level, are used in positron emission tomography (PET) and single-photon emission computer tomography (SPECT), widely used clinical imaging modalities.

4.2.1.2 Microwaves and Radiowaves

Microwaves and radio waves lie at the other end of the electromagnetic spectrum. These low frequency waves also interact weakly with biological tissues leading to deep penetration. Preparation of tissue with a strong magnetic field allows radio waves to yield images with high contrast between different tissue types through magnetic resonance imaging (MRI). Although they are nonionizing, radio frequency and microwave radiation can induce localized heating in tissues, and are thus useful for some surgical interventions.

4.2.1.3 Infrared, Visible, and Ultraviolet Light

Light, from infrared to visible and ultraviolet, occupies a window of particular relevance in biomedical applications. Electromagnetic waves at these optical frequencies are easily collected into beams, directed, and focused using mirrors and lenses. Light interacts in a number of ways with biological specimens and can yield a wealth of information. Absorption and scattering limit the penetration of light in tissues but provide good contrast to reveal detailed structures down to the cellular and subcellular level. Specialized optical components can also be used to visualize changes in the properties of light, such as phase and polarization, as it is transmitted through a sample. Photons of ultraviolet light, as well as the shorter wavelengths of visible light, carry just the right energy to induce the phenomena of fluorescence. Particular molecules, whether introduced (exogenous) or intrinsic (endogenous) to the tissue, convert short wavelength light to longer wavelengths of a particular signature, simultaneously revealing both their presence and identity.

In general, light is produced by the energy-level transitions of atomic electrons. As electrons leap down from more energetic to less energetic states, the energy difference is emitted as photons of light. Very large energy gaps correspond to high-frequency ultraviolet light. Smaller electronic transitions yield visible light. Low-frequency infrared light is produced by even closer transitions, or by changes in molecular vibrational and rotational states.

4.2.1.4 The Sun as a Light Source

The sun is the most universally accessible source of light. The heat of the sun generates a spectral pattern that approximately follows blackbody radiation at a temperature of about 5800 K (see Figure 4.2). Peak emission is in the region from 400 to 750 nm, corresponding to the human visual range.

Sunlight was a convenient and ubiquitous source of illumination during the development of the microscope—the earliest technology to harness light in studying biological tissues. The sun provided bright light and its paraxial rays were easy to collect and focus. However, obvious limitations such as inconsistent light quality and availability motivated the development of more controllable light sources. Over the past century, light sources and biomedical applications have progressed in parallel. Technological innovations in light sources have spurred further developments in biomedical applications, while ever more specialized applications have inspired more sophisticated new sources.

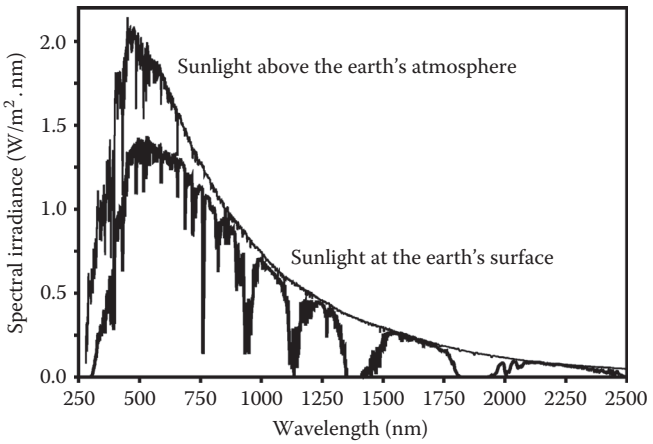

FIGURE 4.2 Spectral profile of sunlight above the Earth's atmosphere and at the Earth's surface. The difference between the two profiles is due to reflection and absorption of light within the atmosphere, primarily by molecules of water, oxygen, and carbon dioxide. (Based on data from American Society for Testing and Materials (ASTM) Terrestrial Reference Spectra for Photovoltaic Performance Evaluation, ASTM G-173-03.)

4.2.1.5 Man-Made Light Sources

Gas discharge and incandescent lamp sources were initially introduced as proxies for the incoherent radiation provided by the sun, but providing more consistent control and availability of light. These devices remain efficient work horses for many applications, particularly in microscopy and spectroscopy. However, it was the introduction of laser light sources that revolutionized the field of biomedical optics. Novel properties of coherent laser light opened new avenues of interrogation and interaction with biological tissues. As laser systems continue to advance not only in capability, but also in ease of use, optical techniques become more important and more pervasive in the practice of medicine. Man-made light sources may be categorized into two classes: nonlasers and lasers, as described in the following sections.

4.2.2 Nonlaser Sources

4.2.2.1 High-Pressure Arc Lamps

High-pressure arc lamps generate high intensity quasi-continuum spectra ranging from the ultraviolet to near-infrared. A gas (typically xenon, mercury, or a mixture of these) is filled into a quartz envelope with two tungsten electrodes. An applied voltage ionizes the gas and creates a bright arc between the electrodes. The electrical current driving these gas discharge lamps need to be regulated otherwise they would quickly burn out. The output power of commercially available high-pressure arc lamps ranges from a few watts to several kilowatts.

The xenon arc lamp is widely used as a steady state spectroscopy source because of its smooth spectral profile between 250 and 700 nm. Other arc lamps, such as the mercury-filled variant, feature prominent spectral lines more suitable for applications needing single wavelength excitation. Being an isotropic source, the emission from arc lamps is usually omnidirectional. Parabolic and ellipsoidal reflectors may be used to collect the generated light and focus it to a sample or specimen.

4.2.2.2 Low-Pressure Vapor Lamps

Low-pressure vapor lamps operate via gas discharge similar to high-pressure arc lamps, but can be operated with less complex power supplies. They are favored as high-intensity sources of stable spectral lines as each elemental gas generates a known set of characteristic lines. For example, a low-pressure mercury lamp has a dominant wavelength peak at 253.7 nm as well as a prominent triplet at 365.0/365.5/366.3 nm. Because of such characteristic discrete lines, these lamps are used in simple filter type spectrometers and also serve as calibration light sources.

4.2.2.3 Incandescent Lamps

An incandescent lamp generates light by heating a metal filament, typically tungsten. This provides a continuous blackbody radiation spectrum that is useful for, among other things, intensity calibration of various types of detectors including spectrometers. Incandescent lamps are inexpensive and very simple to operate with no external regulating requirements. They are very convenient sources for broadband illumination at infrared to visible wavelengths. However, ultraviolet output from these lamps is usually very low.

4.2.2.4 Light-Emitting Diodes

LEDs are solid-state, semiconductor-based light sources. Light is produced via electroluminescence when positive and negative charge carriers recombine at a semiconductor junction. Their intrinsic emission is close to monochromatic and wavelength is determined by the characteristic energy bandgap of the semiconductor or semiconductor alloy used. The range of available wavelengths from LEDs stretches from the near-infrared (~1050 nm) to the ultraviolet (UV) (~250 nm) and continues to be expanded. The addition of phosphors and integration of multiple emitters in single devices have also made broadband and white-light LEDs possible.

Since the output of LEDs can be modulated at high frequencies (>100 MHz), they are popular light sources in applications requiring fast response. Additional advantages are their compact size and the low level of heat generated during operation, which render them well suited for use in tight spaces and potentially even for in vivo use. Although LEDs are very energy efficient, they are still unable to match the high output powers of incandescent lamps and arc lamps.

4.2.3 Laser Sources

4.2.3.1 Properties of Lasers

Low-coherence lamps and LEDs are useful for many applications, particularly when low cost and simple operation are the primary concern. However, many modern techniques in

biomedical optics are practical only when applied with laser sources due to several important characteristics of coherent laser radiation.

4.2.3.1.1 Directionality

Directionality is the most immediately evident property of laser radiation. While light from incoherent lamp sources tends to spread out in all directions, light from a laser propagates as a beam along a particular direction. Laser beams can travel long distances with little loss of intensity and can be focused into very small beam spots, limited ultimately by diffraction. Tight focusing is essential for high resolution imaging, such as in confocal microscopy. Conservation of power along the beam also means that very high light intensities can be achieved at the focal spot. This property is key to achieving practical signal-to-noise ratios in many biomedical imaging applications.

4.2.3.1.2 Monochromaticity

Monochromaticity is another reason lasers appear much brighter than typical lamp sources for the same total output power. In general, lasers carry power over a very narrow spectral span, as opposed to broadband sources where power is spread over a wider range of wavelengths. This spectral concentration is very useful for techniques that require selective but efficient optical excitation. Laser hair removal is a specific example. Although "monochromatic" implies a single discrete frequency for laser oscillation, in reality all lasers posses a finite spectral bandwidth. Some lasers even operate with multiple modes, that is, multiple regularly spaced spectral lines. A laser may also be tunable such that the emission can be shifted across a range of wavelengths.

4.2.3.1.3 Coherence

Coherence refers to a consistent phase relationship between distant points on a light wave. In practice, coherence is determined by observing the interference patterns of overlapping light waves. Coherent light produces sharp contrast between dark and bright fringes. Fringes produced by partially coherent light appear washed out with poor contrast. Completely incoherent light does not produce any interference pattern at all.

We differentiate between spatial coherence and temporal coherence. Spatial coherence compares light from two points spatially separated across a single wavefront of light. The maximum separation for which interference can still be observed is called the spatial coherence length. Temporal coherence compares two sequential points along a wave train. Coherence time is defined as the maximum time interval over which a light wave can still interfere with a previous segment of itself. Since light is a traveling wave in both space and time, coherence time is often measured as an equivalent temporal coherence length.

Lasers can have spatial and temporal coherence lengths reaching several meters. In contrast, light from an incandescent lamp has coherence lengths in the order of just a few micrometers. Coherent laser sources enable the practical application of interferometry and holography that rely on the diffraction and interference of light. Coherence also leads to the formation of randomly distributed bright and dark interference spots, that is, speckle, when laser light illuminates an extended area. Speckle is often considered undesirable noise in imaging applications, but may sometimes be utilized to reveal information about surfaces.

4.2.3.1.4 Short Pulse Duration

Lasers may operate in either continuous wave (CW) or pulsed modes. A CW laser operates with uniform output power over time, while a pulsed laser emits light in a sequence of short bursts. Pulsed operation can be advantageous as it concentrates energy into a short window of time. This allows very high levels of instantaneous power even while maintaining modest average power. Ultrashort-pulsed lasers with pulse durations in the order of femtoseconds (10^{-15} s) have made high-resolution multiphoton microscopy techniques practical. High-energy pulsed lasers also enable very precise surgical procedures through very controlled and localized ablation of tissue.

4.2.3.2 Fundamental Laser Components

Lasers seem to come in a boundless variety of shapes, sizes, and materials—from microchip-sized semiconductor lasers, to fiber lasers tens of meters long, or even amplified laser systems that can fill a large room. However, regardless of the specific implementation, there are three basic components that can be identified in any operational laser—the gain medium, excitation or pump mechanism, and resonant cavity. These elements are diagrammed schematically in Figure 4.3.

4.2.3.2.1 Gain Medium

Lasers are often identified by specifying the gain media, for example, a helium–neon (HeNe) laser, or a neodymium:yttrium aluminum garnet (Nd:YAG) laser. The gain medium, alternatively referred to as the active medium, is where light amplification takes place. The material may be a solid, liquid, or gas. The electronic energy-level structure of a material determines its emission wavelengths and bandwidths; thus the gain medium partly determines the central wavelength and wavelength span of laser emission. Gas media, in particular, have narrow gain bandwidths. They amplify light over a very narrow spectral range (<10^{-3} nm) and are practically single-wavelength sources.

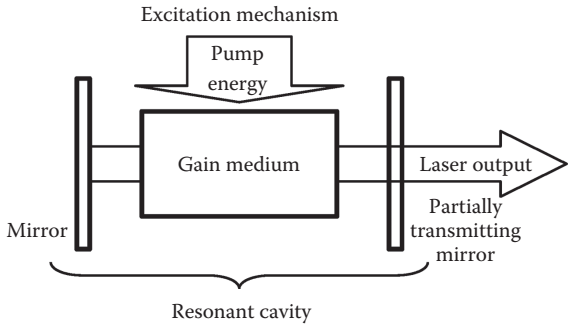

FIGURE 4.3 Fundamental laser components. The gain medium, excitation mechanism, and resonant cavity are the three basic components that can be identified in any operational laser.

Solid and liquid materials have broad gain bandwidths and can support lasing action over much wider intervals. Broad bandwidths enable the design of tunable lasers and ultrashort-pulsed lasers.

4.2.3.2.2 Excitation Mechanism

Before a gain medium can amplify light, atoms in the material must be suitably excited. An excitation mechanism is needed to pump energy into the gain medium. This energy is stored by atoms in their excited state, and released via either spontaneous or stimulated emission as the atoms relax back to their original ground state (see Figure 4.4). Excitation may be through electrical pumping, such as when injecting an electrical current into a semiconductor or gaseous gain medium. Alternatively, a laser may use an optical pumping scheme where the gain medium absorbs light, either from a flash lamp or another laser, and converts it to a different wavelength for emission.

4.2.3.2.3 Resonant Cavity

Spontaneous emission from an excited gain medium does not posses the coherent properties of laser radiation. A resonant cavity is necessary to provide optical feedback for stimulated emission to take place. Reintroducing previously emitted light knocks down excited atoms back to the ground state. These stimulated transitions emit coherent radiation without absorbing the optical feedback. Thus, the incident light is amplified, particularly if the excitation mechanism has induced population inversion in the gain medium, that is, there are more atoms in the excited state than the ground state. A resonant cavity is typically constructed by placing the gain medium between two mirrors, forming a Fabry–Perot resonator. The resonator may also be as simple as the cleaved ends of a semiconductor gain medium. Other configurations for optical feedback include ring resonators and distributed feedback via scattering.

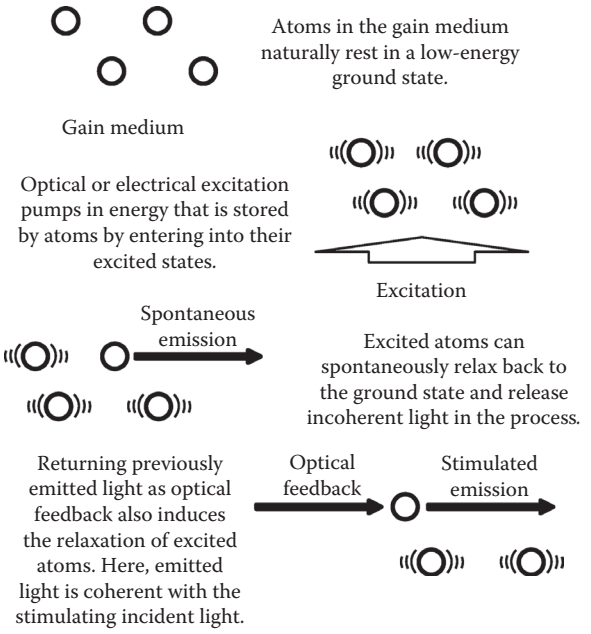

FIGURE 4.4 Fundamental laser operation. Optical or electrical excitation provides energy that is stored by atoms in the gain medium. This energy is released as incoherent light through spontaneous emission, or as coherent light through stimulated emission in the presence of optical feedback.

Ultimately, the resonant modes of the cavity, in conjunction with the gain profile of the active medium, determine the wavelengths at which a laser operates. This is referred to as the longitudinal modes of a laser. The resonator also determines the cross-sectional profile, or transverse modes, of the laser beam. Most often, the cavity is designed to produce a singular circular lobe of the fundamental Gaussian transverse mode.

Some laser designs may also feature other components for expanded functionality. For example, modulators can be used to induce pulsed laser operation. Spectral filters may be inserted to control the specific output wavelength in a tunable laser. However, the three fundamental components: gain medium, excitation mechanism, and resonant cavity remain the defining parts of a laser, and will always be present in any laser system.

4.2.3.3 Pulsed Laser Operation

Many biomedical applications require coherent light to be delivered as a series of laser pulses. Depending on the specific application and corresponding power density requirements, pulse durations may reach as short as several femtoseconds (10^{-15} s). Pulses longer than a few microseconds (10^{-6} s) are easily produced by directly modulating the CW output of a laser via mechanical, electro-optic, or acousto-optic shutter mechanisms. Shorter pulse durations, in the order of nanoseconds (10^{-9} s) or less, require more specialized laser designs. Two important methods for generating short laser pulses are Q-switching and modelocking.

4.2.3.3.1 Q-Switching

Q-switched lasers produce high-energy pulses that range from picoseconds to nanoseconds in duration. In preparation for pulse formation, optical feedback is prevented in the laser cavity by an absorbing or scattering component. The resonant cavity is then said to be in a low-Q state. Without optical feedback, energy builds up in the gain medium as there is no stimulated emission to deplete it. When the gain has reached its peak level (i.e., saturation), a high-Q state is restored by removing the source of loss. Stimulated emission then builds up exponentially while the stored energy in the gain medium is rapidly depleted. The result is an optical pulse that carries the maximum amount of energy that can be delivered by the active medium. Pulse energy is limited by the excited state lifetime of the gain medium. Longer lifetimes allow more energy to build up. Pulse duration is limited by cavity lifetime which measures the time it takes to extract energy from the resonant cavity. Q-switched lasers typically achieve nanosecond pulsewidths. In order to maximize pulse energy, repetition rates of Q-switched lasers are generally kept slow such that the gain has sufficient time to reach saturation.

4.2.3.3.2 Modelocking

Modelocking can produce laser pulses from several picoseconds to less than 10 fs in width. Pulse formation in modelocked lasers can be interpreted in terms of interference between multiple longitudinal modes. The greater the number of modes involved in the interference, the shorter the resulting pulses. Very broad gain bandwidth is thus necessary to achieve the narrowest pulses. The necessary interference can only occur if the different longitudinal laser modes maintain a fixed phase relationship between themselves. This is not an inherent behavior in lasers, but can be induced by creating modulation sidebands that link, that is, lock, adjacent modes to each other. The modulation frequency is determined by the round-trip travel time of pulses circulating in the cavity, typically in the order of tens of megahertz.

4.2.3.4 Types of Lasers

Lasers can be broadly categorized according to the type of gain media they utilize. The most generally recognized types of lasers are gas, dye, solid-state, and semiconductor. Below, some basic properties representative of each type are described. These are followed by some specific examples of laser systems with particular relevance in biomedical applications.

4.2.3.4.1 Gas Lasers

Gas lasers principally use electric current discharge as an excitation mechanism. Gas molecules receive energy via collisions with accelerating electrons. Often, a mixture of gases is used in the gain medium to improve the efficiency of energy transfer. However, emissions originate from just a single species. Different gas lasers cover a wide range of laser wavelengths from the infrared, across the visible, and into the ultraviolet. Very narrow spectral emission profiles are an important distinguishing property gas lasers have over other laser types. The sparse density of the gaseous medium minimizes perturbations to the internal energy levels of the active molecules. In general, gas lasers are also quite robust against thermal damage even when operating at high power levels, thanks to large volumes of gain media and gas flow systems that allow heat to dissipate quickly.

The CO_2 gas laser was one of the earliest lasers to be developed, but it remains relevant even today. CO_2 lasers typically operate between at 9.6 and 10.6 μm by accessing vibrational transitions of gaseous CO_2 molecules. High efficiency and powerful output levels—reaching over 100 kW in CW operation and 10 kJ per pulse in pulsed operation—are important advantages of this type of laser. The strong absorption by water at these wavelengths make the CO_2 laser particularly suited for application in laser surgery and other photothermal therapies such as skin resurfacing.

The HeNe laser has long served as a convenient source of red (632.8 nm) CW-laser radiation at low powers (1–100 mW). Although today it has been replaced in most applications by less expensive semiconductor lasers operating at similar wavelengths, this lasers still maintains one key advantage—long coherence length. HeNe lasers remain important light sources for holography and interferometry applications.

Excimer lasers are a class of gas lasers that use noble gas or noble gas halide complexes as gain media. The term excimer is a contraction of "excited dimer" that describes the excited state of the gain medium. Some specific examples of this type include: xenon fluoride (351 nm), xenon chloride (308 nm), krypton fluoride (248 nm), and argon fluoride (193 nm). Pulsed operation at ultraviolet wavelengths is characteristic of these lasers. Typical pulse durations are a few tens of nanoseconds with pulse energy falling between 0.2 and 1.0 J. The combination of short pulses, high energy, and UV wavelengths make excimer lasers useful for highly precise surgical procedures, such as sculpting the surface of the cornea to correct refractive errors in vision. The short UV pulses can remove very small amounts of tissue in a highly controlled fashion without any damage to surrounding areas. However, this approach is only suitable for exposed surfaces or those accessible with an endoscope as UV light does not penetrate very far into tissue.

4.2.3.4.2 Solid-State Lasers

Solid-state lasers refer specifically to lasers where the gain media are metal-ion dopants in crystalline or glass host materials. The dopants may be rare-earth ions such as neodymium (Nd) and erbium (Er), or transition-metal ions like titanium (Ti) and chromium (Cr). These lasers are optically pumped with either flashlamps or other laser devices. Laser emission typically centers about the near-infrared region, although some solid-state lasers may emit at slightly longer infrared wavelengths (~2 μm). Gain bandwidth is generally broad as the intrinsic electric field of the host material perturbs the electronic energy levels of the dopant ions. Thus, solid-state gain media are useful for building tunable lasers and pulsed lasers utilizing either Q-switching or modelocking mechanisms. Depending on the specific host material and bulk configuration used, solid-state lasers can also be designed to deliver high output power levels.

Erbium- and neodymium-based lasers are popular sources for near-infrared wavelengths close to 1 μm. Laser emission results from electronic transitions in the dopant ions, but energy level perturbations by the host material significantly affect the emission wavelength. For example, an Nd:YAG ($Y_3Al_5O_{12}$) laser will typically operate at 1.064 μm, while an Nd:glass laser operates at 1.054 μm. The Nd:YAG is a popular platform for producing Q-switched pulses because of its long excited state lifetime, as well as the good mechanical stability and thermal conductivity of the YAG crystal that allow it to withstand high energy pulses. Such lasers have become very useful for photodisruption of tissues in laser surgery. Erbium shows an even wider host-dependent variation of lasing wavelength. Er:YAG output is at 2.94 μm, while Er:glass emission centers at 1.54 μm. The longer wavelength of the Er:YAG laser makes it very useful for phototherapy on the surface of the skin where it is quickly absorbed by water molecules.

Solid-state lasers doped with titanium and chromium also operate in the near-infrared, but are distinguished by very

broad gain bandwidths, ranging from 100 to 400 nm. The Ti:sapphire ($Ti:Al_2O_3$) is perhaps the most widely used among these types of lasers today. It may be configured as a CW laser tunable from 700 to 980 nm, or as a modelocked laser capable of producing femtosecond pulses. The modelocked Ti:sapphire laser has become the workhorse for multiphoton microscopy as its wavelength range is suitable for two-photon excitation of many standard fluorescent dyes and fluorescent proteins—green fluorescent protein (GFP) stands out among these. Cr^{4+}:Fosterite($Cr^{4+}:Mg_2SiO_4$) is another significant solid-state laser of this type. It operates around 1.25 μm and provides deeper tissue penetration compared to the Ti:sapphire. These broadband solid-state lasers are also important light sources for optical coherence tomography.

4.2.3.4.3 Dye Lasers

Dye lasers have gain media that consist of organic dyes dissolved in a liquid such as ethyl or methyl alcohol, glycerol, or water. These dyes exhibit wide emission bands across the visible region, and sometimes stretching into near-infrared wavelengths. Similar to solid-state lasers, optical pumping is also achieved with flashlamps or other laser sources. Tunable laser operation across the visible wavelengths is an important feature of dye lasers, and it is broad bandwidth that enables pulsed operation via either Q-switching or modelocking.

The complex and potentially hazardous handling of liquid dyes and solvents has tended to limit general interest in dye lasers for applications beyond spectroscopic studies in research laboratories. However, the ability of dye lasers to produce high energy pulses at wavelengths strongly absorbed by hemoglobin has been proven useful in the treatment of vascular lesions and the removal of scars and tattoos from the skin.

4.2.3.4.4 Semiconductor Lasers

Semiconductor lasers, or laser diodes, differ significantly from the solid-state lasers described above in terms of the mechanisms for excitation and emission. Semiconductor gain media are pumped via direct injection of electrical current. Electrons and holes are delivered from opposite sides of a semiconductor junction. As in LEDs, the recombination of these negative and positive charge carriers results in the emission of light at an optical frequency proportional to the semiconductor band gap. Given sufficient current density and optical feedback, stimulated emission and laser action take place. Emission bandwidths are typically broad, with center wavelengths in the red to near-infrared region. More recently developed wide bandgap semiconductor lasers with blue-violet wavelengths have also become commercially important.

Simple operation, physically compact and robust devices, and low cost compared to other lasers are among the advantages that make semiconductor lasers very attractive sources for applications requiring low to moderate power (i.e., from milliwatts to several watts). Significant biomedical applications of semiconductor lasers range from laser hair removal to the treatment of macular degeneration and retinopathy in the eye.

4.3 Light Detectors

4.3.1 Introduction

The vast variety of light-sources and techniques to image, investigate, or manipulate biomedical samples, calls for light detectors tailored to specific requirements. The result is an equally abundant multitude of light detection schemes. But at the heart of these schemes, there actually lies just a handful of different detector device types.

The basic functionality of most light detectors is the same: They make use of interactions between radiation and matter to convert the light intensity into a proportional electrical signal. More sophisticated optical detection systems allow measurements of various other properties such as the spatial, spectral and temporal profile, polarization, and phase of the incident light. However, such systems still contain simple intensity detectors as their basic building blocks. The most common among these are photodiodes and PMTs, both of which are single channel detectors that only measure light at a single point. In order to simultaneously measure a spatial distribution of light intensity, an arrayed detector is necessary. CCDs are the most common example of these, although complementary metal-oxide-semiconductor (CMOS) detectors and photodiode or PMT arrays are gradually becoming more important.

4.3.2 Photodiodes

4.3.2.1 Principle of Operation

Photodiodes are solid-state photodetectors that operate on the basis of the internal photoelectric effect. The material providing the sensitivity to incident light is a semiconductor. In pure semiconductors, the electrons fill up all available energy states in the valence band, which is one of the two relevant manifolds of energetic states in semiconductors. The other relevant band, the conductance band, is separated from the valence band by the bandgap and is higher in energy. At room temperature and in the absence of any light, the lack of any unoccupied states in the valence band means that the semiconductor cannot support a net movement of charges. It then has a very low (ideally zero) conductivity. Thus, little or no current is measured if one applies an electric field across the material. However, incident photons with energies larger than the bandgap can excite electrons from the valence to the conduction band where they are free to move. This results in an increase of conductivity, which is (over a certain range) proportional to the number of incident photons. Consequently, the application of an electric field now results in a measurable current that is proportional to the number of the incident photons.

Most photodiodes used today are based on *p*–*i*–*n* structures. Here, a layer of an intrinsic semiconductor, that is, a semiconductor free of any impurities, is sandwiched between a *p*- and an *n*-doped region. In these doped regions a defined number of mobile holes (*p*-doped) or electrons (*n*-doped) are created by adding a well-controlled amount of impurities. The *p*–*i*–*n*

configuration induces a permanent electric field in the intrinsic region of the structure and thus facilitates generation of a photo-induced current, even if no external electric field is present. In this short-circuit or photovoltaic mode, *p–i–n* photodiodes feature low noise levels (due to the absence of significant shot-noise, see below) but have a limited response time. Application of an external voltage (reverse-bias mode) improves the response time at the cost of increased noise (also see APDs below).

The bandgap of the semiconductor on which a photodiode is based defines the wavelength range over which the device is sensitive. Photons with energies below the bandgap will not excite electrons into the conductance band and thus will not generate a signal. Independent of its energy, one photon only excites a single electron in a conventional photodiode. This reduces the response (relative to the incident power) for short wavelengths and usually defines the lower end of the spectral response curve of a photodiode.

4.3.2.2 Characteristics and Applications

Photodiodes are possibly the most widely used photodetector today. Among them, silicon-based devices are the most common. The prevalence of silicon is due to the extreme maturity of the silicon technology, which is the result of some 50 years of development. In particular, silicon devices are relatively low-cost, offer linear response, are very rugged, and can be easily integrated with other optical and electronic components. Silicon photodiodes can be used to detect light in the near UV, the entire visible, and some part of the near infrared (NIR) region of the spectrum. At the maximum of the spectral response curve, which is located at around 800 nm, *p–i–n* silicon photodiodes can have a quantum efficiency approaching unity, that is, nearly every incident photon causes excitation and subsequent extraction of an electron.

Silicon photodiodes feature a sharp drop-off in spectral response at wavelengths longer than 1.1 μm. Thus, alternative materials are required to access the full NIR region (0.75–1.6 μm). This spectral region is important for many biomedical optical techniques since most biological samples show relatively low absorption at these wavelengths. The band structure of germanium renders photodiodes based on this material sensitive to light with wavelengths up to 1.8 μm. Good response in the NIR can also be obtained in hetero-junction structures where the *p*- and the *n*-type part of the diode consists of different materials, for example, heterojunction photodiodes based on the semiconductor alloys InGaAsP and InP. Such devices achieve a quantum efficiency approaching 75%. Despite the significant progress in heterojunction technology over recent years, the use of different materials in one detector still considerably increases the cost of manufacturing.

The inherent spectral response of photodiodes is relatively broad. Various coatings and filters can be used to block certain spectral regions, thereby defining a narrow range of wavelengths to which the detector is sensitive. In addition, anti-reflection coatings are routinely employed to improve the sensitivity of photodiodes and various other photodetectors.

In reverse-bias configuration, *p–i–n* type photodiodes can achieve response times on the order of a few tens of picoseconds rendering them well suited for the detection of fast processes provided the absolute intensity of the incident light is sufficiently high to obtain reliable measurements. The ultrafast detection of small optical signals requires more sophisticated metrology.

An ideal photodiode generates an electrical signal that is simply proportional to the optical input. However, even in the absence of light, a photodiode will generate a finite electrical output, referred to as dark current. In addition, both the dark current and the real signal will have a random component, known as noise. The total noise level is a superposition of different contributions. The two most important to be aware of are the shot-noise and the thermal noise. The presence of shot-noise is a direct result of the photon nature of light. Each incident photon generates a tiny signal pulse. Even for a constant optical signal the incident photon stream will thus have random fluctuations that translate into electrical noise at the output of the detector. In this context, the term shot-limited noise level is sometimes used to indicate that the noise of the output signal is at or close to this fundamental limit. Thermal noise results from random motion of charges within the photodiode and can be reduced by cooling of the detector. To quantify the overall noise level for a photodiode or any other detector or detection system, one usually quotes the noise equivalent power (NEP). The NEP is defined as the radiant power of an incident signal that results in a signal-to-noise ratio of unity at the detector output. The NEP is measured for a sinusoidally modulated input signal and varies with the frequency of the input signal.

It is important to note that the properties of nearly all semiconductor-based photodetectors significantly depend on temperature. We have already seen that the noise level increases with temperature and that cooling of the photodiode might be necessary to measure small signals. However, even for fairly large signals an active temperature control might be required to ensure repeatable measurements if the photodiode is used in an environment with significant temperature variation. In addition, external mechanical stress may considerably increase the noise level and electromagnetic interference (EMI) can constitute a serious source of noise that calls for measures to protect the photodiode, for example, by shielding or adequate packaging.

4.3.2.3 Avalanche Photodiodes

Photodiodes that are designed for operation at a high reverse bias (usually 50–300 V) are known as avalanche photodiodes (APDs). The reverse bias generates a high electric field inside the photodiode that accelerates the photo-generated charges. Collision of these accelerated (primary) charges with the crystal lattice leads to ionization, which in turn generates secondary electron-hole pairs. In a chain reaction (avalanche process), these secondary charges are then also accelerated and thus generate additional charge by further ionization processes. Depending on the reverse bias, avalanche ionization amplifies the photocurrent by 50 to several 100 times. One usually refers to this amplification as the gain or multiplication factor of the APD.

If the reverse-bias is too high (i.e. above the breakdown voltage), the arrival of just a single photon triggers a continuing avalanche process (Geiger mode). While this is useful for single-photon counting experiments and allows measuring the arrival of a single photon with high time resolution (≤50 ps), the reset required to stop the avalanche process results in a dead time of approximately 50 ns. One can therefore not use the Geiger mode when dealing with large optical signals or considerable amounts of background light.

Both, the breakdown voltage and the gain are strongly temperature dependent. In addition, if the APD is operated close to the breakdown voltage, the generated photocurrent itself can lead to a significant reduction of the electric field in the avalanche region of the device. This creates an unwanted situation in which the photocurrent is not proportional to the input signal.

Due to high gain factors, APDs are particularly useful when measuring small optical signals that one might not be able to detect with conventional photodiodes. However, the increased sensitivity brings in some limitations: (1) operation close to the breakdown voltage results in nonlinear response; (2) the active area of APDs is very limited in size since defects in the crystal structure and strain must be avoided to achieve the high breakdown voltages required for an efficient avalanche process; (3) the noise level is often higher than in conventional photodiodes since amplification noise constitutes an additional source of noise; and (4) the readout circuit is generally more complex due to the need for a high operating voltage.

4.3.3 Photomultiplier Tubes

4.3.3.1 Principle of Operation

In contrast to photodiodes, photomultiplier tubes (PMTs) employ the external photoelectric effect. As illustrated in Figure 4.5, incident photons hit a photocathode and if their energy is sufficiently high, extract electrons from the material by ionization. The critical energy for this photoemission process depends on the photocathode material: Photoemission

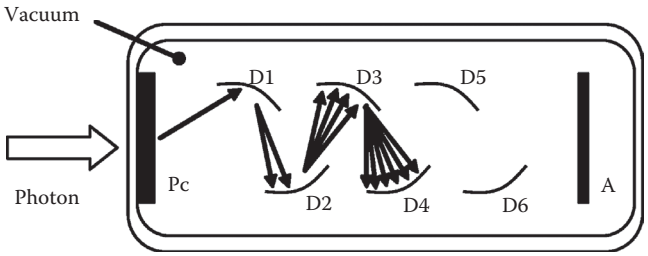

FIGURE 4.5 Schematic illustration of a photomultiplier tube. Incident photons extract electrons from the photocathode (PC). The electrons are accelerated toward the first dynode (D1). At each dynode (D1 to D6) the incident electrons stimulate emission of additional secondary electrons. This avalanche process results in the buildup of a large current pulse that is collected at the anode (A).

only takes place if the energy of the incident photon is above the work function W of the photocathode or in other words if the wavelength of the light is shorter than hc/W, where h is Planck's constant and c is the speed of light. However, even if the photons provide sufficient energy, electron-trapping in the bulk of the photocathode my still prevent their extraction. Therefore, the quantum efficiency of photoemission, that is, the ratio of extracted electrons over incident photons, is well below unity in real devices. The extracted electrons are focused by a pair of electrodes and then accelerated toward a secondary electrode, or *dynode*, which is kept at an attractive electric potential. Upon impact, each incident primary electron can extract several secondary electrons by ionization processes. Using again a drop in electric potential, the secondary electrons are collected by a third electrode, the *anode*. The flow of electrons from photocathode to anode gives rise to a photocurrent that forms the output signal. The components of a PMT are housed in a vacuum tube to allow electrons to travel over macroscopic distances once they have been extracted from the cathode.

4.3.3.2 Characteristics and Applications

The PMT may be regarded as the vacuum analogue of the APD. However, the possibility to add multiple amplification stages by integrating several dynodes (usually between 6 and 14) allows substantially higher amplification factors. Depending on the voltage drop between the electrodes and the number of dynodes, a gain between 10^5 and 10^8 can be achieved. The quantum efficiency of the primary photoemission process, however, is considerably lower than in semiconductor photodiodes. Traditionally, metals and metal alloys—especially alloys of alkali metals—have been used as photocathodes. Due to partial reflection of the incident light and trapping of electrons, these electrodes typically have quantum efficiencies around 10%. Today, metal photocathodes have been complemented by semiconductor alloy cathodes with lower reflectivity and quantum efficiency up to 30%.

As with photodiodes, the spectral response of PMTs is strongly dependent on the photocathode material used, but is also influenced by the transmission profile of the vacuum tube's glass window. PMTs can cover the UV range (down to about 120 nm) and the entire visible part of the spectrum. The sensitivity of most PMTs drops sharply at wavelengths above 1 μm, which limits their application in the IR. Exceptions are devices using p-n-junction photocathodes based on InP/InGaAs that provide reasonable sensitivity up to 1.6 μm.

Again in analogy to the situation for APDs, the design of the peripheral circuitry required to operate the PMT and to measure the output signal is not trivial. The overall voltage drop between photocathode and anode usually ranges between 500 and 3000 V and a set of voltage-dividing resistors is required to adjust the voltage gradient between the individual elements of the PMT. Special care must be taken to avoid introduction of additional noise from the peripheral circuitry. For instance, improper selection of the grounding scheme will result in a considerable

increase of the noise level and can even cause permanent damage of the PMT. Depending on the measurement scheme, PMTs can be used to either measure DC or AC signals or can be operated in photon-counting mode.

Assuming an appropriate driving circuit, the largest source of noise in PMTs is thermal electron extraction from the photocathode. This effect is most pronounced for materials with a low work function since the chance of electron extraction by thermal fluctuations is relatively high in this case. Vice versa, photocathodes with high work function and sensitivity in the UV or blue end of the visible spectrum are virtually immune to thermal noise. Cooling of the photocathode, and in certain cases also of the first dynode, with a Peltier element generally reduces thermal noise.

Additional potential sources of noise include external electric or magnetic fields that will distort the electron trajectories in the PMT. The strong signal amplification in PMTs and the fact that PMTs are usually housed in evacuated glass or quartz tubes also means that they are extremely susceptible to stray light. Therefore, appropriate housing and shielding are particularly important for PMTs. Often the outer surface of PMTs is coated with black conductive paint, also referred to as hydroxyapatite HA coating. "HA coating" is a term commonly used by manufacturers to refer to black conductive paint applied to the outer surface of the PMT glass envelope in order to reduce noise. It is not composed of hydroxyapatite.

Depending on the dynode geometry the *signal rise time* of PMTs varies between 1 and 20 ns. The limiting factor is the spread in electron transit time that results from the fact that electrons follow different trajectories on their way from the photocathode to the anode. Certain dynode designs can limit this effect. However, this usually comes at the cost an overall reduction in sensitivity or leads to reduced uniformity across the active area of the photocathode.

4.3.3.3 Microchannel Plates

Microchannel plates (MCPs) are a refinement of the PMT concept and offer improvements mainly in terms of response time, compactness, reduced power consumption, and lower sensitivity to magnetic fields. Instead of using a series of discrete dynodes, MCPs are based on a disc-shaped array of capillaries with diameters in the 10-μm range and lengths on the order of 1 mm. The inner walls of the capillaries have a high Ohmic resistance coating. If a voltage is applied to both ends of the capillary, a gradient in electric potential is generated between one end and the other. Primary electrons extracted from the photocathode of the MCP are absorbed by the capillary array and passed toward the opposite end. Due to the high aspect ratio of the capillaries, these electrons impinge on the capillary walls multiple times and thus generate an avalanche of secondary electrons that then exits the capillary at the opposite end.

MCPs can achieve a temporal resolution (signal *rise time*) on the order of 100 ps since all electron trajectories inside the microchannel array are nearly parallel and thus have almost equal lengths. In addition, the absolute path lengths are much shorter than in conventional PMTs. Therefore, all electrons have a similar transit time, which is a principal requirement for fast response.

4.3.4 Arrayed Detectors

The photodetectors discussed so far are single element detectors. However, biomedical optics often requires the simultaneous measurement of multiple spatially separated optical signals, for example, to create an image of a specimen or to measure the spectral distribution of an optical signal. If a large detector array is required, as it is the case for image generation, detection and read-out should be integrated on a single chip. Separate wiring of discrete detectors would be impractical in terms of cost and device size.

4.3.4.1 Charge Coupled Device—Principle of Operation

A very common detector array scheme based on a photodiode-like detection principle is the CCD. Instead of using the *p–i–n* diodes described before, CCD sensors consist of an array of metal–oxide–semiconductor (MOS) capacitors as photosensitive elements. These structures generate and store an amount of charge that is proportional to the number of photons collected since the detector was last read-out, which means that CCD chips are integrating detectors. During the measurement, all photocapacitors record the incident optical signal simultaneously. The subsequent read-out of the charge stored in each capacitor or pixel, however, is a sequential process. One common read-out scheme, which is to embed charge transfer channels under each row of the CCD array, is illustrated in Figure 4.6. These channels consist of a series of capacitors, referred to as wells, each of which is capable of storing the charge generated by one of the pixels during exposure. By manipulating the voltage applied to these wells, the generated charge can be transferred along the channel and is eventually collected at the end of the channel by an amplifier. The spatial origin of each charge package can be inferred from the sequence of arrival, which ultimately allows the reconstruction of the image. To avoid intermixing of the charge packages accumulated in neighboring pixels, at least three wells are required for every pixel.

4.3.4.2 Characteristics and Applications of CCDs

Most CCD chips are based on silicon as the active and photosensitive material. Therefore, they achieve good sensitivity in the visible and near IR (≤1 μm) range of the spectrum with quantum efficiencies between 20% and 40%. Conventional front-illuminated CCDs have limited sensitivity in the UV due to significant absorption of UV light in the top electrodes of the CCD chip. Good UV sensitivity can either be achieved by down-conversion of the UV light to visible wavelengths using phosphors or preferably by using a back-illumination scheme. In this case, the wafer on which the CCD is fabricated is back-thinned so that the light can be passed to the photodetectors on the chip through the backside of the substrate. This scheme greatly reduces absorption losses and yields good sensitivity over the UV, visible, and near IR range with quantum efficiency between 50% and 90% across the entire spectral range. CCD chips can also be used for direct detection of x-rays. In this case, the generation of multiple

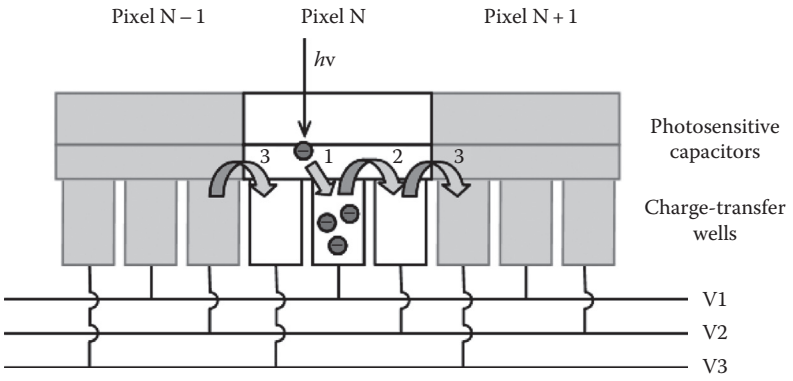

FIGURE 4.6 Illustration of the charge transfer mechanism used to read out the photo-generated charge from a CCD detector. Incident photons ($h\nu$) charge the photosensitive capacitor. Application of a voltage to line V1 transfers the charges to one of the three wells connected to each pixel. The charge is transported to the adjacent well when line V2 is switched on and V1 is switched off. The charge is then transferred to the first well of the neighboring pixel by activation of line V3. Repetitive voltage cycles applied to V1, V2, and V3 thus transport the charge generated in each pixel along the channel without intermixing charge from different pixels.

electron-hole pairs per incident photon even enables determining the energy of the incident photons.

Like conventional photodiodes, the noise generated in a CCD chip can be reduced by active cooling. Sensitive CCD chips are often equipped with integrated thermoelectric (TE) coolers to ensure a homogenous and constant temperature distribution of the chip.

The dynamic range of any detector is defined as the ratio between the largest input signal that can be measured without saturating the detector and the smallest signal that can be distinguished from noise. The dynamic range is of particular importance for image detectors as a low dynamic range will reduce the contrast of recorded images. For CCDs, the saturation limit is given by the capacitance of the wells storing the photo-generated charge, whereas the noise limit at low light intensities is usually dominated by read-out noise—at least if the device is TE-cooled. Simple CCDs typically have an 8 bit dynamic range, that is, the saturation level is 256 times larger than the noise level. More sophisticated devices can reach dynamic ranges of up to 16 bit or 65,000. The dynamic range must not be confused with the bit-depth, which refers to the number of digitization steps provided by the analogue-to-digital converter (ADC) on the chip. Ideally, the bit-depth should be equal or slightly larger than the dynamic range. However, CCDs with bit-depths that greatly exceed the actual dynamic range are frequently offered, in the hope that the large number of digitization steps gives the (false) impression of a large dynamic range.

CCD image sensors were developed in the 1970s, and have since then been tremendously improved with regard to their dynamic range, sensitivity, production cost, speed, and noise level. Today, they have replaced many other imaging sensors, including the photographic film, which was for a long time superior to electronic image detection in terms of dynamic range, sensitivity, and cost. In addition, electronic imaging schemes such as CCD imaging have enabled new kinds of data collection, in particular by allowing continuous data acquisition and convenient image analysis. It is worth noting that compared to other technological breakthroughs mediated by the semiconductor technology, this is a relatively recent advance and that the widespread use of high quality CCD image sensors in biomedical labs has only begun some 15 years ago.

4.3.4.3 CCD Spectrometers

Measurements of the spectral distribution of an optical signal can often provide very useful information about the composition or state of a specimen. Traditional techniques include absorption and fluorescence spectroscopy, but Raman-spectroscopy and various time-resolved and nonlinear spectroscopy techniques have complemented these methods and can often provide additional or more precise information. Similar to the situation with CCD image sensors, traditional spectroscopic measurement techniques such as scanning spectrometers were to a large extent replaced by CCD-based spectrometers over the past 10–15 years. In addition, CCD-based systems have enabled various measurements that have not been possible before, as their fast and parallel data acquisition allows the measurement of rapidly changing spectral profiles.

Spectrometers come in various configurations and refinements. A typical configuration is a CCD-based fiber spectrometer that consists of an entrance slit, a diffraction grating, and a linear CCD detector (see Figure 4.7). The light enters through the entrance slit, which restricts the spatial distribution of the incoming light to a narrow line. The light is then collimated and passed onto the diffraction grating, which is the main functional component of the spectrometer. The grating reflects light of different wavelengths under different angles, which enables spatial separation of different spectral components in the input signal. The diffracted light is subsequently focused onto the CCD detector, which consists of a single line of sensitive elements. The position of the focus on the CCD detector depends on the diffraction angle and thus on the wavelength of the incident light. The spectrometer is calibrated using a light

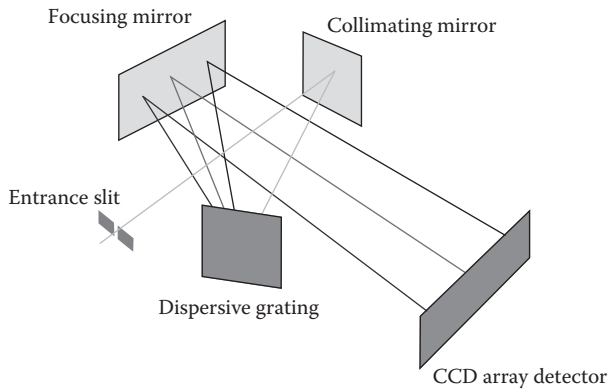

FIGURE 4.7 Schematic illustration of a crossed Czerny-Turner CCD spectrometer. Light enters through the slit, is collimated by a first mirror, and then divided into its spectral components by a dispersive grating. The focusing mirror creates an image of the entrance slit on the CCD detector. Since the position of the image depends on the wavelength of the light, the CCD signal yields the spectrum of the incident light.

source with a set of spectrally narrow emission lines at known wavelengths.

In addition to the previously discussed general performance parameters of photodetectors, spectrometers can be characterized by their spectral range and spectral resolution. The first refers to the wavelength range that can be unambiguously analyzed with a spectrometer; the latter describes the minimal wavelength difference of two spectral lines that can be resolved. It is technologically challenging to combine a large spectral range and a high spectral resolution in one device. In addition, improving the spectral resolution usually reduces the sensitivity of the spectrometer. The optimal trade-off between these conflicting performance parameters therefore depends very much on the intended application. For instance, Raman spectroscopy requires excellent spectral resolution and high sensitivity but only a limited spectral range. By contrast, a low spectral resolution is often acceptable for fluorescence spectroscopy since the fluorescence spectra of many dyes used in biomedical optics are spectrally broad.

4.3.4.4 Complementary Metal–Oxide–Semiconductor Detectors

CMOS is an alternative technology to integrate light measurement and read-out for a large array of detectors on a single chip. Instead of using charge collecting wells and a charge transfer process to transfer the information about the light intensity, each pixel of a CMOS sensor comprises a separate circuit to process and digitize the signal before transferring the data to the output interface of the chip. This scheme adds complexity to the design of the chip and sacrifices uniformity of detection and fill-factor (i.e., the fraction of the chip surface that is light sensitive). In the past, these issues prevented fabrication of CMOS-based image sensors with acceptable performance. Advances in the CMOS process technology, however, have resulted in a rediscovery of the CMOS sensor concept over recent years. The CMOS concept is very attractive since the required process technology is widely used in other fields of the electronics industry, which offers the potential of lower production cost. In addition, the use of standard semiconductor technology also enables integration of image processing functions on the chip, which reduces the complexity of off-chip circuitry. Finally, CMOS sensors usually consume less power than comparable CCD chips. While many of these features were originally thought to render CMOS sensors particularly suitable for consumer applications, they may also turn out useful for biomedical optics, especially in high-throughput experiments where cost is often a limiting factor or in situations where highly integrated electronics ("camera on a chip") can enable new types of measurements. Some examples are endoscopes containing ultraminiaturized cameras and data acquisition units, which are often also single use.

4.3.4.5 APD and PMT Arrays

Both the CCD and the CMOS sensor can be regarded as an arrayed version of conventional photodiodes combined with an efficient read-out scheme. Other sensor types, in particular APDs and PMTs, are also available as arrays. However, the number of detectors per array is significantly lower in these cases due to technical limitations. While CCDs and CMOS chips with pixel counts in excess of 10^7 are commercially available, APD and PMT arrays are usually limited to well below 100 detectors. Therefore, they cannot be directly used for imaging. However, they are very useful to simultaneously collect weak optical signals at different wavelengths, as required, for example, in scanning imaging techniques such as confocal or two-photon microscopy.

4.3.4.6 Intensified CCD Cameras

Intensified CCD cameras can be used to perform direct imaging of low-level optical signals. These devices are a combination of a MCP–PMT and a sensitive CCD chip. The optical signal is incident on the front surface of the MCP and the primary electrons generated at the photocathode are amplified as they pass through the capillaries of the MCP. Instead of collecting the photocurrent across the MCP, the electrons are directed toward a phosphor screen at the output facet of the MCP to reconstruct the spatial distribution of the input signal. Due to the large gain of the MCP (usually $\geq 10^5$), the image on the phosphor screen is orders of magnitude brighter and can therefore be recorded with the CCD chip. The spectral response of intensified CCD cameras is determined by the response of the photocathode. Different cameras with a range of photocathodes covering the UV, visible, and NIR are available today.

A prominent variant of the intensified CCD camera is the streak camera that can measure the temporal profile of fast optical signals with a subpicosecond time resolution. Since streak cameras can measure the temporal profile of many channels simultaneously, they can be combined with a spectrograph to perform ultrafast spectroscopic measurements. As shown

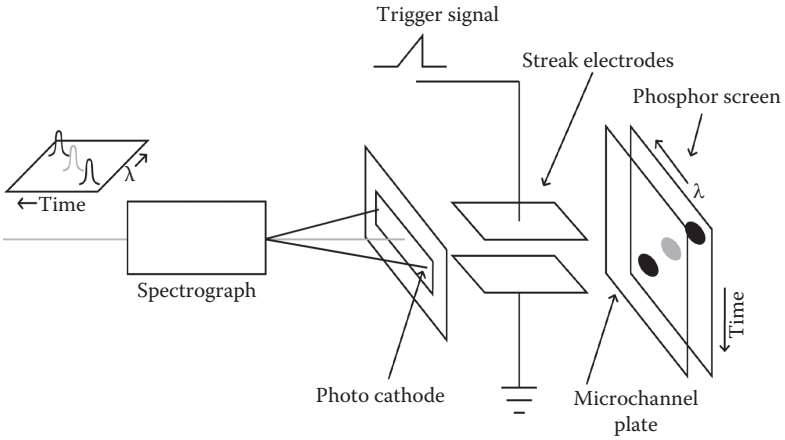

FIGURE 4.8 Schematic of a streak camera setup for ultrafast spectroscopic measurements. The optical signal is passed through a spectrograph in order to split up the different spectral components in *x*-axis. The photo cathode then converts the optical signal into electrons. The streak electrodes deflect the electrons along the *y*-axis by different amounts depending on their time of arrival. The electron signal is then amplified by the microchannel plate and converted back into optical information by a phosphor screen that is coupled to a sensitive CCD chip.

in Figure 4.8, a voltage ramp, which is synchronized with the optical signal, is applied to a pair of additional electrodes that is mounted in between the photocathode and the actual MCP. Depending on their arrival time at the detector, the primary electrons generated at the photocathode are exposed to a different electric field. Therefore, they are deflected by different distances, impinge on the MCP at different positions, and are thus laterally resolved on the CCD chip.

4.4 Irradiation Guidelines

4.4.1 Introduction

The common intent of all techniques described in this book is to use the interaction of light with tissue or cells to cure and diagnose diseases or to understand basic biological mechanisms. However, whenever biological tissue or living cells are exposed to optical radiation, adverse processes or responses can be triggered depending on the nature and state of cells and tissue, and most importantly depending on wavelength, power, and duration of optical exposure. In this section, we will deal with such phenomena to provide a guide to safe and effective use of optical radiation in biomedicine. In the process, we will see that often times, damage mechanisms have been turned into useful technologies to gather unique information or devise novel therapy treatments.

4.4.2 Radiation Effects at the Tissue Level

In this section, we will briefly explain the harmful phenomena that may occur when light is delivered to tissue. Such physical mechanisms are the basis for establishing damage thresholds for light applications. For an exact calculation of damage thresholds in the specific circumstances of a particular experimental situation, the American National Standards Institute (ANSI) publishes exhaustive guidelines for the safe use of lasers and other light sources both in clinical and in research settings (ANSI 2000).

4.4.2.1 Photochemical Damage

At low power densities, on the order of 1 W/cm^2, prolonged exposure (>1 s) to lasers or other light sources can cause unwanted chemical reactions in tissue that may produce damage identified as photochemical. If a tissue component is excited by high-energy visible radiation, the energy released upon de-excitation may split a bond in another molecule producing reactive oxygen species, the most common of which are singlet oxygen, hydrogen peroxide, hydroxyl radicals, and other free radicals. Reactive oxygen species are very aggressive and very dangerous to tissue especially because they can attack and break cell membranes.

In the retina, this type of damage is particularly interesting since the retina is devoted to visual perception and transduction of photons in this region, yet it is vulnerable to blue light because of photooxidation of photoreceptors and lipofuscin pigments in the retinal pigment epithelium (Wu et al. 2006, Delori et al. 2007). The damage threshold associated to this mechanism is independent of exposure duration time, although for very long exposure times (>1 day) one has to take into account incipient repair mechanisms.

4.4.2.2 UV-Induced Risks

Ultraviolet radiation covers the spectral range of 180–400 nm, the most harmful region of the optical spectrum because each photon carries sufficient energy (>3 eV) to induce molecular transitions inside a cell. UV-induced risks are more significant than any other radiation given the ubiquitous presence of UV in sunlight. A significant decrease in cell viability of all types has been reported after UV irradiation, even moderate, with evidence suggesting that DNA damage is the primary cause of