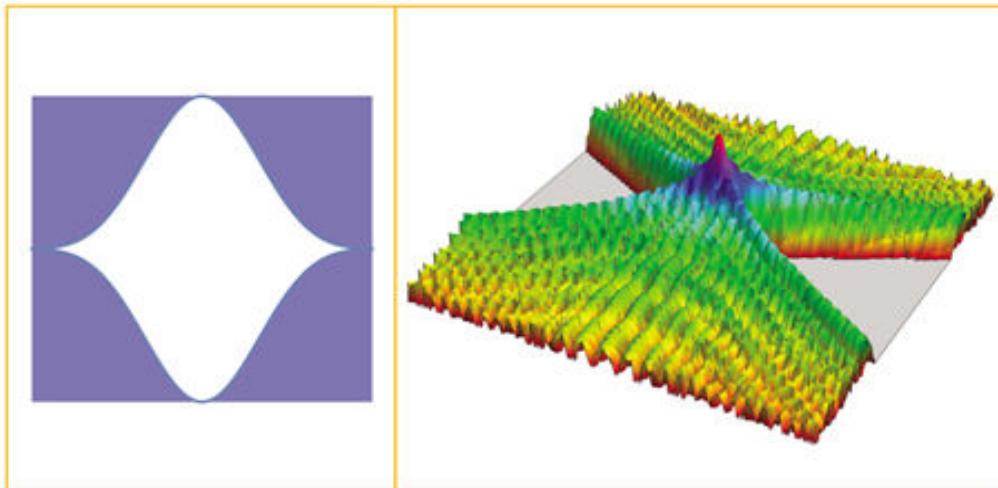


FOURTH EDITION

Introduction to

FOURIER OPTICS



Joseph W. Goodman

Introduction to Fourier Optics

FOURTH EDITION

Joseph W. Goodman
Stanford University



w.h.freeman
Macmillan Learning
New York

To the memory of my mother, Doris R. Goodman, and my father, Joseph Goodman, Jr.

Vice President, STEM: *Ben Roberts*
Acquisitions Editor: *Lori Stover*
Marketing Manager: *Maureen Rachford*
Director, Content Management Enhancement: *Tracey Kuehn*
Managing Editor: *Lisa Kinne*
Director of Design, Content Management: *Diana Blume*
Cover Designer: *Kevin Kall, Kall Design*
Project Editor: *Matthew Gervais, Lumina Datamatics, Inc.*
Permissions Manager: *Jennifer MacMillan*
Photo Editor: *Robin Fadool*
Production Manager: *Susan Wein*
Composition: *Lumina Datamatics, Inc.*
Cover Image: *Joseph W. Goodman*

Library of Congress Control Number: 2017930862

ISBN 978-1-319-15304-5 (epub)

© 2017 by W. H. Freeman and Company

All rights reserved

W. H. Freeman and Company
One New York Plaza
Suite 4500
New York, NY 10004-1562
www.macmillanlearning.com

The Author

Joseph W. Goodman came to Stanford in 1958 as a graduate student, and remained there his entire professional career. He was the primary dissertation advisor for 49 Ph.D. graduates, many of whom are now prominent in the field of optics. He held the William Ayer Chair in Electrical Engineering at Stanford, and also served in several administrative posts, including Chair of the Department of Electrical Engineering, and Senior Associate Dean of Engineering for Faculty Affairs. He is now the William Ayer Professor Emeritus. His work has been recognized by a variety of awards and honors, including the F.E. Terman award of the American Society for Engineering Education, the Dennis Gabor Award and the Gold Medal of the SPIE, the Max Born Award, the Esther Beller Hoffman Award, the Emmett Leith Award and the Frederic Ives Medal of the Optical Society of America, and the Education Medal of the Institute of Electrical and Electronic Engineers. He is a member of the National Academy of Engineering and has served as President of the Optical Society of America and the International Commission for Optics. He has been a cofounder of several companies, including Optivision, Inc., ONI Systems, NanoPrecision Products Inc., and Roberts & Company Publishers.

Preface

Fourier analysis is a ubiquitous tool that has found application to diverse areas of physics and engineering. This book deals with its applications in optics, and in particular with applications to diffraction, imaging, optical information processing, holography, and optical communications.

Since the subject covered is Fourier optics, it is natural that the methods of Fourier analysis play a key role as the underlying analytical structure of our treatment. Fourier analysis is a standard part of the background of most physicists and engineers. The theory of linear systems is also familiar, especially to electrical engineers. [Chapter 2](#) reviews the necessary mathematical background. For those not already familiar with Fourier analysis and linear systems theory, it can serve as the outline for a more detailed study that can be made with the help of other textbooks explicitly aimed at this subject. Ample references are given for more detailed treatments of this material. For those who have already been introduced to Fourier analysis and linear systems theory, that experience has usually been with functions of a single independent variable, namely time. The material presented in [Chapter 2](#) deals with the mathematics in two spatial dimensions (as is necessary for most problems in optics), yielding an extra richness not found in the standard treatments of the one-dimensional theory.

The book can be used as a textbook to satisfy the needs of several different types of courses. It is directed towards both physicists and engineers, and the portions of the book used in the course will in general vary depending on the audience. However, by properly selecting the material to be covered, the needs of any of a number of different audiences can be met.

There are many people to whom I owe a special word of thanks for their help with the earlier editions of the book. Early versions of the manuscript were used in courses at several different universities. I would in particular like to thank Profs. A.A. Sawchuk, J.F. Walkup, J. Leger, P. Pichon, D. Mehrl, and their many students for catching so many typographical errors and in some cases outright mistakes. Helpful comments for the second edition were also made by I. Erteza and M. Bashaw, for which I am grateful. Several useful suggestions were also made by anonymous manuscript reviewers engaged by the publisher. A special debt is owed to Prof. Emmett Leith and Prof. Adolf Lohmann, who provided many helpful suggestions. I would also like to thank the students in my 1995 Fourier Optics class, who competed fiercely to see who could find the most mistakes.

For [Chapter 12](#) of this edition, which first appeared in the third edition as [Chapter 10](#), I am indebted to Prof. Andrew Weiner, Mr. Gregory Brady, Dr. Dmitry Starodubov, and Dr. Jane Lam for their helpful comments and suggestions.

This fourth edition differs from the third edition through corrections of earlier typos and errors, and the addition of a great deal of new material, including but not limited to [Chapter 5](#) on computational propagation and diffraction and [Chapter 8](#) on engineered point-spread functions and transfer functions. For this edition, my largest debt is to Prof. James Fienup, as well as his students. Jim gave me suggestions and corrections on many, many occasions concerning different parts of the book, and I also received help from his students. I thank Dr. Jeffrey P. Wilde, who found many typos and mistakes for me to correct, especially in the two new chapters. I owe thanks

to Prof. Demetri Psaltis for educating me regarding lithography of diffractive optical elements. I would also like to thank Prof. David Voelz, who read and commented on [Chapter 5](#), Prof. Gouang Zheng, who explained to me one of the subtleties of ptychographic imaging, Prof. W.E. Moerner, who read the section on fluorescence microscopy and corrected one important error, Dr. Prasanna Pavani, who read and commented on the section on rotating point-spread functions, and Prof. A.A. Sawchuk, who reviewed [Chapter 5](#). Undoubtedly there are others to whom I owe thanks, and I apologize for not mentioning them explicitly here.

Finally, I thank Hon Mai, without whose patience, encouragement, and support this book would not have been possible.

Joseph W. Goodman

Note for Instructors

A complete manual with solutions to all of the problems found in this book is available from the publisher, but only to instructors.

Contents

1. [Preface](#)
2. [1 Introduction](#)
 1. [1.1 Optics, Information, and Communication](#)
 2. [1.2 The Book](#)
3. [2 Analysis of Two-Dimensional Signals and Systems](#)
 1. [2.1 Fourier Analysis in Two Dimensions](#)
 1. [2.1.1 Definition and Existence Conditions](#)
 2. [2.1.2 The Fourier Transform as a Decomposition](#)
 3. [2.1.3 Fourier Transform Theorems](#)
 4. [2.1.4 Separable Functions](#)
 5. [2.1.5 Functions with Circular Symmetry: Fourier-Bessel Transforms](#)
 6. [2.1.6 Some Frequently Used Functions and Some Useful Fourier Transform Pairs](#)
 2. [2.2 Spatial Frequency and Space-Frequency Localization](#)
 1. [2.2.1 Local Spatial Frequencies](#)
 2. [2.2.2 The Wigner Distribution Function](#)
 3. [2.3 Linear Systems](#)
 1. [2.3.1 Linearity and the Superposition Integral](#)
 2. [2.3.2 Invariant Linear Systems: Transfer Functions](#)
 4. [2.4 Two-Dimensional Sampling Theory](#)
 1. [2.4.1 The Whittaker-Shannon Sampling Theorem](#)
 2. [2.4.2 Oversampling, Undersampling and Aliasing](#)

- 3. [2.4.3 Space-Bandwidth Product](#)
- 5. [2.5 The Discrete Fourier Transform](#)
- 6. [2.6 The Projection-Slice Theorem](#)
- 7. [2.7 Phase Retrieval from Fourier Magnitude](#)

- 4. [3 Foundations of Scalar Diffraction Theory](#)
 - 1. [3.1 Historical Introduction](#)
 - 2. [3.2 From a Vector to a Scalar Theory](#)
 - 3. [3.3 Some Mathematical Preliminaries](#)
 - 1. [3.3.1 The Helmholtz Equation](#)
 - 2. [3.3.2 Green's Theorem](#)
 - 3. [3.3.3 The Integral Theorem of Helmholtz and Kirchhoff](#)
 - 4. [3.4 The Kirchhoff Formulation of Diffraction by a Planar Screen](#)
 - 1. [3.4.1 Application of the Integral Theorem](#)
 - 2. [3.4.2 The Kirchhoff Boundary Conditions](#)
 - 3. [3.4.3 The Fresnel-Kirchhoff Diffraction Formula](#)
 - 5. [3.5 The Rayleigh-Sommerfeld Formulation of Diffraction](#)
 - 1. [3.5.1 Choice of Alternative Green's Functions](#)
 - 2. [3.5.2 The Rayleigh-Sommerfeld Diffraction Formula](#)
 - 3. [3.5.3 Reproduction of Boundary Conditions](#)
 - 6. [3.6 Kirchhoff and Rayleigh-Sommerfeld Theories Compared](#)
 - 7. [3.7 Further Discussion of the Huygens-Fresnel Principle](#)
 - 8. [3.8 Generalization to Nonmonochromatic Waves](#)
 - 9. [3.9 Diffraction at Boundaries](#)
 - 10. [3.10 The Angular Spectrum of Plane Waves](#)
 - 1. [3.10.1 The Angular Spectrum and Its Physical Interpretation](#)
 - 2. [3.10.2 Propagation of the Angular Spectrum](#)
 - 3. [3.10.3 Effects of a Diffracting Aperture on the Angular Spectrum](#)
 - 4. [3.10.4 The Propagation Phenomenon as a Linear Spatial Filter](#)

5. 4 Fresnel and Fraunhofer Diffraction

1. 4.1 Background

1. 4.1.1 The Intensity of a Wave Field
2. 4.1.2 The Huygens-Fresnel Principle in Rectangular Coordinates

2. 4.2 The Fresnel Approximation

1. 4.2.1 Positive vs. Negative Phases
2. 4.2.2 Accuracy of the Fresnel Approximation
3. 4.2.3 Finite Integral of the Quadratic-Phase Exponential Function
4. 4.2.4 The Fresnel Approximation and the Angular Spectrum
5. 4.2.5 Fresnel Diffraction Between Confocal Spherical Surfaces
6. 4.2.6 Fresnel Diffraction in Terms of Ray Transfer Matrices

3. 4.3 The Fraunhofer Approximation

4. 4 Examples of Fraunhofer Diffraction Patterns

1. 4.4.1 Rectangular Aperture
2. 4.4.2 Circular Aperture
3. 4.4.3 Thin Sinusoidal Amplitude Grating
4. 4.4.4 Thin Sinusoidal Phase Grating
5. 4.4.5 General Method for Calculating Diffraction Efficiency of Gratings

5. 4.5 Examples of Fresnel Diffraction Calculations

1. 4.5.1 Fresnel Diffraction by a Square Aperture
2. 4.5.2 Fresnel Diffraction by a Circular Aperture
3. 4.5.3 Fresnel Diffraction by a Sinusoidal Amplitude Grating-Talbot Images

6. 4.6 Beam Optics

1. 4.6.1 Gaussian Beams
2. 4.6.2 Hermite-Gaussian Beams
3. 4.6.3 Laguerre-Gaussian Beams
4. 4.6.4 Bessel Beams

6. 5 Computational Diffraction and Propagation

1. 5.1 Approaches to Computational Diffraction
2. 5.2 Sampling a Space-Limited Quadratic-Phase Exponential
3. 5.3 The Convolution Approach
 1. 5.3.1 Bandwidth and Sampling Considerations
 2. 5.3.2 Discrete Convolution Equations
 3. 5.3.3 Simulation Results
 4. 5.3.4 Convolution by Fourier Transforms
4. 5.4 The Fresnel Transform Approach
 1. 5.4.1 Sampling Increments
 2. 5.4.2 Sampling Ratio Q
 3. 5.4.3 Finding the Required M, Q, and N
 4. 5.4.4 The Discrete Diffraction Formulas
 5. 5.4.5 Examples of the Dependence of M and N on N_F
 6. 5.4.6 Summary of Steps Using the Fresnel Transform Approach
 7. 5.4.7 Computational Complexity of the Fresnel Transform Approach
5. 5.5 The Fresnel Transfer Function Approach
 1. 5.5.1 Sampling Considerations
 2. 5.5.2 Finding N, M and Q for each N_F
 3. 5.5.3 The Discrete Diffraction Formulas
 4. 5.5.4 Examples of the Dependence of M, N and Q on N_F
 5. 5.5.5 Summary of Steps Using the Fresnel Transfer Function Approach
 6. 5.5.6 Computational Complexity of the Fresnel Transfer Function Approach
6. 5.6 The Exact Transfer Function Approach
 1. 5.6.1 Sampling in the Frequency Domain
 2. 5.6.2 Sampling in the Space Domain
 3. 5.6.3 Simulation Results
 4. 5.6.4 Computational Complexity of the Exact Transfer Function Approach

7. [5.7 Comparison of Computational Complexities](#)
8. [5.8 Extension to More Complex Apertures](#)
 1. [5.8.1 One-Dimensional Case](#)
 2. [5.8.2 Two-Dimensional Apertures Separable in \(x,y\) Coordinates](#)
 3. [5.8.3 Circularly-Symmetric Apertures](#)
 4. [5.8.4 More General Cases](#)
9. [5.9 Concluding Comments](#)

7. [6 Wave-Optics Analysis of Coherent Optical Systems](#)

1. [6.1 A Thin Lens as a Phase Transformation](#)
 1. [6.1.1 The Thickness Function](#)
 2. [6.1.2 The Paraxial Approximation](#)
 3. [6.1.3 The Phase Transformation and Its Physical Meaning](#)
2. [6.2 Fourier Transforming Properties of Lenses](#)
 1. [6.2.1 Input Placed against the Lens](#)
 2. [6.2.2 Input Placed in Front of the Lens](#)
 3. [6.2.3 Input Placed behind the Lens](#)
 4. [6.2.4 Example of an Optical Fourier Transform](#)
3. [6.3 Image Formation: Monochromatic Illumination](#)
 1. [6.3.1 The Impulse Response of a Positive Lens](#)
 2. [6.3.2 Eliminating Quadratic-Phase Factors: The Lens Law](#)
 3. [6.3.3 The Relation between Object and Image](#)
4. [6.4 Analysis of Complex Coherent Optical Systems](#)
 1. [6.4.1 The Ray Matrix Approach](#)
 2. [6.4.2 Analysis of Two Optical Systems Using Ray Matrices](#)

8. [7 Frequency Analysis of Optical Imaging Systems](#)

1. [7.1 Generalized Treatment of Imaging Systems](#)
 1. [7.1.1 A Generalized Model](#)

- 2. [7.1.2 Effects of Diffraction on the Image](#)
- 3. [7.1.3 Polychromatic Illumination: The Coherent and Incoherent Cases](#)
- 2. [7.2 Frequency Response for Diffraction-Limited Coherent Imaging](#)
 - 1. [7.2.1 The Amplitude Transfer Function](#)
 - 2. [7.2.2 Examples of Amplitude Transfer Functions](#)
- 3. [7.3 Frequency Response for Diffraction-Limited Incoherent Imaging](#)
 - 1. [7.3.1 The Optical Transfer Function](#)
 - 2. [7.3.2 General Properties of the OTF](#)
 - 3. [7.3.3 The OTF of an Aberration-Free System](#)
 - 4. [7.3.4 Examples of Diffraction-Limited OTFs](#)
- 4. [7.4 Aberrations and Their Effects on Frequency Response](#)
 - 1. [7.4.1 The Generalized Pupil Function](#)
 - 2. [7.4.2 Effects of Aberrations on the Amplitude Transfer Function](#)
 - 3. [7.4.3 Effects of Aberrations on the OTF](#)
 - 4. [7.4.4 Example of a Simple Aberration: A Focusing Error](#)
 - 5. [7.4.5 Apodization and Its Effects on Frequency Response](#)
- 5. [7.5 Comparison of Coherent and Incoherent Imaging](#)
 - 1. [7.5.1 Frequency Spectrum of the Image Intensity](#)
 - 2. [7.5.2 Two-Point Resolution](#)
 - 3. [7.5.3 Other Effects](#)
- 6. [7.6 Confocal Microscopy](#)
 - 1. [7.6.1 Coherent Case](#)
 - 2. [7.6.2 Incoherent Case](#)
 - 3. [7.6.3 Optical Sectioning](#)
- 9. [8 Point-Spread Function and Transfer Function Engineering](#)
 - 1. [8.1 Cubic Phase Mask for Increased Depth of Field](#)
 - 1. [8.1.1 Depth of Focus](#)

2. [8.1.2 Depth of Field](#)
3. [8.1.3 The Cubic Phase Mask](#)
2. [8.2 Rotating Point-Spread Functions for Depth Resolution](#)
3. [8.3 Point-Spread Function Engineering for Exoplanet Discovery](#)
 1. [8.3.1 The Lyot Coronagraph](#)
 2. [8.3.2 Apodization for Starlight Suppression](#)
4. [8.4 Resolution beyond the Classical Diffraction Limit](#)
 1. [8.4.1 Analytic Continuation](#)
 2. [8.4.2 Synthetic Aperture Fourier Holography](#)
 3. [8.4.3 Fourier Ptychography](#)
 4. [8.4.4 Coherent Spectral Multiplexing](#)
 5. [8.4.5 Incoherent Structured Illumination Imaging](#)
 6. [8.4.6 Super-Resolved Fluorescence Microscopy](#)
5. [8.5 Light Field Photography](#)

10. [9 Wavefront Modulation](#)

1. [9.1 Wavefront Modulation with Photographic Film](#)
 1. [9.1.1 The Physical Processes of Exposure, Development, and Fixing](#)
 2. [9.1.2 Definition of Terms](#)
 3. [9.1.3 Photographic Film or Plate in Coherent Optical Systems](#)
 4. [9.1.4 The Modulation Transfer Function](#)
 5. [9.1.5 Bleaching of Photographic Emulsions](#)
2. [9.2 Wavefront Modulation with Diffractive Optical Elements](#)
 1. [9.2.1 Single Step Lithography](#)
 2. [9.2.2 Multistep Lithography](#)
 3. [9.2.3 Other Types of Diffractive Optics](#)
 4. [9.2.4 A Word of Caution](#)
3. [9.3 Liquid Crystal Spatial Light Modulators](#)
 1. [9.3.1 Properties of Liquid Crystals](#)

- 2. [9.3.2 Spatial Light Modulators Based on Liquid Crystals](#)
 - 4. [9.4 Deformable Mirror Spatial Light Modulators](#)
 - 5. [9.5 Acousto-Optic Spatial Light Modulators](#)
 - 6. [9.6 Other Methods of Wavefront Modulation](#)
-
- 11. [10 Analog Optical Information Processing](#)
 - 1. [10.1 Historical Background](#)
 - 1. [10.1.1 The Abbe-Porter Experiments](#)
 - 2. [10.1.2 The Zernike Phase-Contrast Microscope](#)
 - 3. [10.1.3 Improvement of Photographs: Maréchal](#)
 - 4. [10.1.4 Application of Coherent Optics to More General Data Processing](#)
 - 2. [10.2 Coherent Optical Information Processing Systems](#)
 - 1. [10.2.1 Coherent System Architectures](#)
 - 2. [10.2.2 Constraints on Filter Realization](#)
 - 3. [10.3 The VanderLugt Filter](#)
 - 1. [10.3.1 Synthesis of the Frequency-Plane Mask](#)
 - 2. [10.3.2 Processing the Input Data](#)
 - 3. [10.3.3 Advantages of the VanderLugt Filter](#)
 - 4. [10.4 The Joint Transform Correlator](#)
 - 5. [10.5 Application to Character Recognition](#)
 - 1. [10.5.1 The Matched Filter](#)
 - 2. [10.5.2 A Character-Recognition Problem](#)
 - 3. [10.5.3 Optical Synthesis of a Character-Recognition Machine](#)
 - 4. [10.5.4 Sensitivity to Scale Size and Rotation](#)
 - 6. [10.6 Image Restoration](#)
 - 1. [10.6.1 The Inverse Filter](#)
 - 2. [10.6.2 The Wiener Filter, or the Least-Mean-Square-Error Filter](#)
 - 3. [10.6.3 Filter Realization](#)

7. [10.7 Acousto-Optic Signal Processing Systems](#)

1. [10.7.1 Bragg Cell Spectrum Analyzer](#)
2. [10.7.2 Space-Integrating Correlator](#)
3. [10.7.3 Time-Integrating Correlator](#)
4. [10.7.4 Other Acousto-Optic Signal Processing Architectures](#)

8. [10.8 Discrete Analog Optical Processors](#)

1. [10.8.1 Discrete Representation of Signals and Systems](#)
2. [10.8.2 A Parallel Incoherent Matrix-Vector Multiplier](#)
3. [10.8.3 Methods for Handling Bipolar and Complex Data](#)

12. [11 Holography](#)

1. [11.1 Historical Introduction](#)
2. [11.2 The Wavefront Reconstruction Problem](#)
 1. [11.2.1 Recording Amplitude and Phase](#)
 2. [11.2.2 The Recording Medium](#)
 3. [11.2.3 Reconstruction of the Original Wavefront](#)
 4. [11.2.4 Linearity of the Holographic Process](#)
 5. [11.2.5 Image Formation by Holography](#)
3. [11.3 The Gabor Hologram](#)
 1. [11.3.1 Origin of the Reference Wave](#)
 2. [11.3.2 The Twin Images](#)
 3. [11.3.3 Limitations of the Gabor Hologram](#)
4. [11.4 The Leith-Upatnieks Hologram](#)
 1. [11.4.1 Recording the Hologram](#)
 2. [11.4.2 Obtaining the Reconstructed Images](#)
 3. [11.4.3 The Minimum Reference Angle](#)
 4. [11.4.4 Holography of Three-Dimensional Scenes](#)
 5. [11.4.5 Practical Problems in Holography](#)

5. [11.5 Image Locations and Magnification](#)

1. [11.5.1 Image Locations](#)
2. [11.5.2 Axial and Transverse Magnifications](#)
3. [11.5.3 An Example](#)

6. [11.6 Some Different Types of Holograms](#)

1. [11.6.1 Fresnel, Fraunhofer, Image, and Fourier Holograms](#)
2. [11.6.2 Transmission and Reflection Holograms](#)
3. [11.6.3 Holographic Stereograms](#)
4. [11.6.4 Rainbow Holograms](#)
5. [11.6.5 Multiplex Holograms](#)
6. [11.6.6 Embossed Holograms](#)

7. [11.7 Thick Holograms](#)

1. [11.7.1 Recording a Volume Holographic Grating](#)
2. [11.7.2 Reconstructing Wavefronts from a Volume Grating](#)
3. [11.7.3 Fringe Orientations for More Complex Recording Geometries](#)
4. [11.7.4 Gratings of Finite Size](#)
5. [11.7.5 Diffraction Efficiency—Coupled Mode Theory](#)

8. [11.8 Recording Materials](#)

1. [11.8.1 Silver Halide Emulsions](#)
2. [11.8.2 Photopolymer Films](#)
3. [11.8.3 Dichromated Gelatin](#)
4. [11.8.4 Photorefractive Materials](#)

9. [11.9 Computer-Generated Holograms](#)

1. [11.9.1 The Sampling and Computation Problems](#)
2. [11.9.2 The Representational Problem](#)

10. [11.10 Degradations of Holographic Images](#)

1. [11.10.1 Effects of Film MTF](#)
2. [11.10.2 Effects of Film Nonlinearities](#)

- 3. [11.10.3 Effects of Film-Grain Noise](#)
- 4. [11.10.4 Speckle Noise](#)
- 11. [11.11 Digital Holography](#)
 - 1. [11.11.1 Offset Reference-Wave Digital Holography](#)
 - 2. [11.11.2 Phase-Shifting Digital Holography](#)
- 12. [11.12 Holography with Spatially Incoherent Light](#)
- 13. [11.13 Applications of Holography](#)
 - 1. [11.13.1 Microscopy and High-Resolution Volume Imagery](#)
 - 2. [11.13.2 Interferometry](#)
 - 3. [11.13.3 Imaging through Distorting Media](#)
 - 4. [11.13.4 Holographic Data Storage](#)
 - 5. [11.13.5 Holographic Weights for Artificial Neural Networks](#)
 - 6. [11.13.6 Other Applications](#)
- 13. [12 Fourier Optics in Optical Communications](#)
 - 1. [12.1 Introduction](#)
 - 2. [12.2 Fiber Bragg Gratings](#)
 - 1. [12.2.1 Introduction to Optical Fibers](#)
 - 2. [12.2.2 Recording Gratings in Optical Fibers](#)
 - 3. [12.2.3 Effects of an FBG on Light Propagating in the Fiber](#)
 - 4. [12.2.4 Applications of FBGs](#)
 - 5. [12.2.5 Gratings Operated in Transmission](#)
 - 3. [12.3 Ultrashort Pulse Shaping and Processing](#)
 - 1. [12.3.1 Mapping of Temporal Frequencies to Spatial Frequencies](#)
 - 2. [12.3.2 Pulse Shaping System](#)
 - 3. [12.3.3 Applications of Spectral Pulse Shaping](#)
 - 4. [12.4 Spectral Holography](#)
 - 1. [12.4.1 Recording the Hologram](#)

- 2. [12.4.2 Reconstructing the Signals](#)
- 3. [12.4.3 Effects of Delay between the Reference Pulse and the Signal Waveform](#)
- 5. [12.5 Arrayed Waveguide Gratings](#)
 - 1. [12.5.1 Component Parts of an Arrayed Waveguide Grating](#)
 - 2. [12.5.2 Applications of AWGs](#)
- 14. [A Delta Functions and Fourier Transform Theorems](#)
 - 1. [A.1 Delta Functions](#)
 - 2. [A.2 Derivation of Fourier Transform Theorems](#)
- 15. [B Introduction to Paraxial Geometrical Optics](#)
 - 1. [B.1 The Domain of Geometrical Optics](#)
 - 2. [B.2 Refraction, Snell's Law, and the Paraxial Approximation](#)
 - 3. [B.3 The Ray-Transfer Matrix](#)
 - 4. [B.4 Conjugate Planes, Focal Planes, and Principal Planes](#)
 - 5. [B.5 Entrance and Exit Pupils](#)
- 16. [C Polarization and Jones Matrices](#)
 - 1. [C.1 Definition of the Jones Matrix](#)
 - 2. [C.2 Examples of Simple Polarization Transformations](#)
 - 3. [C.3 Reflective Polarization Devices](#)
- 17. [D The Grating Equation](#)
- 18. [Bibliography](#)
- 19. [Index](#)

1 Introduction

1.1 Optics, Information, and Communication

Since the late 1930s, the venerable branch of physics known as optics has gradually developed ever-closer ties with the communication and information sciences of electrical engineering. The trend is understandable, for both communication systems and imaging systems are designed to collect or convey information. In the former case, the information is generally of a temporal nature (e.g. a modulated voltage or current waveform), while in the latter case it is of a spatial nature (e.g. a light amplitude or intensity distribution over space), but from an abstract point of view, this difference is a rather superficial one.

Perhaps the strongest tie between the two disciplines lies in the similar mathematics which can be used to describe the respective systems of interest—the mathematics of Fourier analysis and systems theory. The fundamental reason for the similarity is not merely the common subject of “information”, but rather certain basic properties that communication systems and imaging systems share. For example, many electronic networks and imaging devices share the properties called *linearity* and *invariance* (for definitions see [Chapter 2](#)). Any network or device (electronic, optical, or otherwise) that possesses these two properties can be described mathematically with considerable ease using the techniques of *frequency analysis*. Thus, just as it is convenient to describe an audio amplifier in terms of its (temporal) frequency response, so too it is often convenient to describe an imaging system in terms of its (spatial) frequency response.

The similarities do not end when the linearity and invariance properties are absent. Certain nonlinear optical elements (e.g. photographic film) have input-output relationships that are directly analogous to the corresponding characteristics of nonlinear electronic components (diodes, transistors, etc.), and similar mathematical analysis can be applied in both cases.

It is particularly important to recognize that the similarity of the mathematical structures can be exploited not only for analysis purposes but also for *synthesis* purposes. Thus, just as the spectrum of a temporal function can be intentionally manipulated in a prescribed fashion by filtering, so too can the spectrum of a spatial function be modified in various desired ways. The history of optics is rich with examples of important advances achieved by application of Fourier synthesis techniques—the Zernike phase-contrast microscope is an example that was worthy of a Nobel prize. Many other examples can be found in the fields of signal and image processing.

1.2 The Book

The readers of this book are assumed at the start to have a solid foundation in Fourier analysis and linear systems theory. [Chapter 2](#) reviews the required background; to avoid boring those who are well grounded in the analysis of temporal signals and systems, the review is conducted for functions of two independent variables. Such functions are, of course, of primary concern in optics, and the extension from one to two independent variables provides a new richness to the mathematical theory, introducing many new properties that have no direct counterpart in the theory of temporal signals and systems.

The phenomenon called *diffraction* is of the utmost importance in the theory of optical systems. [Chapter 3](#) treats the foundations of scalar diffraction theory, including the Kirchhoff, Rayleigh-Sommerfeld, and angular spectrum approaches. In [Chapter 4](#), certain approximations to the general results are introduced, namely the Fresnel and Fraunhofer approximations, and examples of diffraction-pattern calculations are presented.

[Chapter 5](#) considers the problem of digitally computing diffraction patterns, examining several different approaches and considering their computational efficiencies.

[Chapter 6](#) considers the analysis of coherent optical systems that consist of lenses and free-space propagation. The approach is that of wave optics, rather than the more common geometrical optics method of analysis. A thin lens is modeled as a quadratic-phase transformation; the usual lens law is derived from this model, as are certain Fourier transforming properties of lenses.

[Chapter 7](#) considers the application of frequency analysis techniques to both coherent and incoherent imaging systems. Appropriate transfer functions are defined and their properties discussed for systems with and without aberrations. Coherent and incoherent systems are compared from various points of view. The limits to achievable resolution are derived.

[Chapter 8](#) is devoted to engineered point-spread functions and transfer functions. Subjects covered include cubic phase masks for increased depth resolution, rotating point-spread functions for enhancing depth resolution, coronagraphs and apodization for exoplanet discovery, resolution beyond the diffraction limit by means of synthetic-aperture holography, Fourier ptychography, coherent spectral multiplexing, structured illumination, and super-resolved fluorescence microscopy. The chapter ends with a discussion of light-field photography.

In [Chapter 9](#) the subject of wavefront modulation is considered. The properties of photographic film as an input medium for optical systems are discussed. Diffractive optical elements are discussed in some detail. Attention is then turned to spatial light modulators, which are devices for entering information into optical systems in real time or near real time.

Analog optical information processing is discussed in [Chapter 10](#). Both continuous and discrete processing systems are considered. Applications to image enhancement and pattern recognition are treated.

[Chapter 11](#) is devoted to the subject of holography. The techniques developed by Gabor and by Leith and Upatnieks are considered in detail and compared. Both thin and thick holograms are treated. Extensions to three-dimensional imaging are presented. A section on digital holography has been added to this edition. Various applications of holography are described.

[Chapter 12](#) covers applications of Fourier optics to devices or techniques that are important for optical communications, including fiber Bragg gratings, ultrashort pulse shaping and

processing, spectral holography, and arrayed waveguide gratings.

Finally, several appendices provide further background information.

Problems are found at the end of each chapter, and a solution manual is available for instructors.

2 Analysis of Two-Dimensional Signals and Systems

Many physical phenomena are found experimentally to share the basic property that their response to several stimuli acting simultaneously is identically equal to the sum of the responses that each component stimulus would produce individually. Such phenomena are called *linear*, and the property they share is called *linearity*. Electrical networks composed of resistors, capacitors, and inductors are usually linear over a wide range of inputs. In addition, as we shall soon see, the wave equation describing the propagation of light through most media leads us naturally to regard optical imaging operations as linear mappings of “object” light distributions into “image” light distributions.

The single property of linearity leads to a vast simplification in the mathematical description of such phenomena and represents the foundation of a mathematical structure that we shall refer to here as *linear systems theory*. The great advantage afforded by linearity is the ability to express the response (be it voltage, current, light amplitude, or light intensity) to a complicated stimulus in terms of the responses to certain “elementary” stimuli. Thus if a stimulus is decomposed into a linear combination of elementary stimuli, each of which produces a known response of convenient form, then by virtue of linearity, the total response can be found as a corresponding linear combination of the responses to the elementary stimuli.

In this chapter we review some of the mathematical tools that are useful in describing linear phenomena and discuss some of the mathematical decompositions that are often employed in their analysis. Throughout the later chapters we shall be concerned with stimuli (system inputs) and responses (system outputs) that may be either of two different physical quantities. If the illumination used in an optical system exhibits a property called *spatial coherence*, then we shall find that it is appropriate to describe the light as a spatial distribution of *complex-valued* field amplitude. When the illumination is totally lacking in spatial coherence, it is appropriate to describe the light as a spatial distribution of *real-valued* intensity. Attention will be focused here on the analysis of linear systems with complex-valued inputs; the results for real-valued inputs are thus included as special cases of the theory.

2.1 Fourier Analysis in Two Dimensions

A mathematical tool of great utility in the analysis of both linear and nonlinear phenomena is *Fourier analysis*. This tool is widely used in the study of electrical networks and communication systems; it is assumed that the reader has encountered Fourier theory previously, and therefore that he or she is familiar with the analysis of functions of one independent variable (e.g. time). For a review of the fundamental mathematical concepts, see the books by [Papoulis \[275\]](#), [Bracewell \[37\]](#), and [Gray and Goodman \[146\]](#). A particularly relevant treatment is by [Bracewell \[38\]](#). Our purpose here is limited to extending the reader's familiarity to the analysis of functions of *two* independent variables. No attempt at great mathematical rigor will be made, but rather, an operational approach, characteristic of most engineering treatments of the subject, will be adopted.

2.1.1 Definition and Existence Conditions

The *Fourier transform* (alternatively the *Fourier spectrum* or *frequency spectrum*) of a (in general, complex-valued) function g of two independent variables x and y will be represented here by $\mathcal{F}\{g\}$ and is defined by¹

$$\mathcal{F}\{g\} = \int_{-\infty}^{\infty} \int g(x, y) \exp[-j2\pi(f_X x + f_Y y)] dx dy.$$

$$\mathcal{F}\{g\} = \int_{-\infty}^{\infty} \int g(x, y) \exp[-j2\pi(f_X x + f_Y y)] dx dy.$$

(2-1)

The transform so defined is itself a complex-valued function of two independent variables f_X and f_Y , which we generally refer to as *frequencies*. Similarly, the *inverse Fourier transform* of a function $G(f_X, f_Y)$ will be represented by $\mathcal{F}^{-1}\{G\}$ and is defined as

$$\mathcal{F}^{-1}\{G\} = \int_{-\infty}^{\infty} \int G(f_X, f_Y) \exp[j2\pi(f_X x + f_Y y)] df_X df_Y.$$

$$\mathcal{F}^{-1}\{G\} = \int_{-\infty}^{\infty} \int G(f_X, f_Y) \exp[j2\pi(f_X x + f_Y y)] df_X df_Y.$$

(2-2)

Note that as mathematical operations the transform and inverse transform are very similar, differing only in the sign of the exponent appearing in the integrand. The inverse Fourier transform is sometimes referred to as the *Fourier integral* representation of a function $g(x, y)$.

Before discussing the properties of the Fourier transform and its inverse, we must first decide when (2-1) and (2-2) are in fact meaningful. For certain functions, these integrals may not exist in the usual mathematical sense, and therefore this discussion would be incomplete without at least a

brief mention of “existence conditions.” While a variety of sets of *sufficient* conditions for the existence of (2-1) are possible, perhaps the most common set is the following:

1. $g^{\mathcal{F}}$ must be absolutely integrable over the infinite (x,y) plane.
2. $g^{\mathcal{F}}$ must have only a finite number of discontinuities and a finite number of maxima and minima in any finite rectangle.
3. $g^{\mathcal{F}}$ must have no infinite discontinuities.

In general, any one of these conditions can be weakened at the price of strengthening one or both of the companion conditions, but such considerations lead us rather far afield from our purposes here.

As [Bracewell \[37\]](#) has pointed out, “physical possibility is a valid sufficient condition for the existence of a transform.” However, it is often convenient in the analysis of systems to represent true physical waveforms by idealized mathematical functions, and for such functions one or more of the above existence conditions may be violated. For example, it is common to represent a strong, narrow time pulse by the so-called Dirac delta function² often represented by

$$\delta(t) = \lim_{N \rightarrow \infty} N \exp(-N^2 \pi t^2),$$

$$\delta(t) = \lim_{N \rightarrow \infty} N \exp(-N^2 \pi t^2),$$

(2-3)

where the limit operation provides a convenient mental construct but is not meant to be taken literally. See [Appendix A](#) for more details. Similarly, an idealized point source of light is often represented by the two-dimensional equivalent,

$$\delta(x,y) = \lim_{N \rightarrow \infty} N^2 \exp[-N^2 \pi(x^2 + y^2)].$$

$$\delta(x, y) = \lim_{N \rightarrow \infty} N^2 \exp[-N^2 \pi(x^2 + y^2)].$$

(2-4)

Such “functions”, being infinite at the origin and zero elsewhere, have an infinite discontinuity and therefore fail to satisfy existence condition 3. Other important examples are readily found; for example, the functions

$$f(x,y) = 1 \text{ and } f(x,y) = \cos(2\pi f_x x)$$

$$f(x, y) = 1 \text{ and } f(x, y) = \cos(2\pi f_x x)$$

(2-5)

both fail to satisfy existence condition 1.

If the majority of functions of interest are to be included within the framework of Fourier analysis, some generalization of the definition (2-1) is required. Fortunately, it is often possible to find a meaningful transform of functions that do not strictly satisfy the existence conditions, provided those functions can be defined as the limit of a sequence of functions that are

transformable. By transforming each member function of the defining sequence, a corresponding sequence of transforms is generated, and we call the limit of this new sequence the *generalized Fourier transform* of the original function. Generalized transforms can be manipulated in the same manner as conventional transforms, and the distinction between the two cases can generally be ignored, it being understood that when a function fails to satisfy the existence conditions and yet is said to have a transform, then the generalized transform is actually meant. For a more detailed discussion of this generalization of Fourier analysis the reader is referred to the book by [Lighthill \[227\]](#).

To illustrate the calculation of a generalized transform, consider the Dirac delta function, which has been seen to violate existence condition 3. Note that each member function of the defining sequence (2-4) does satisfy the existence requirements and that each, in fact, has a Fourier transform given by (see [Table 2.1](#))

Table 2.1: Transform pairs for some functions separable in rectangular coordinates.

Function	Transform
$\exp[-\pi(a^2x^2+b^2y^2)]$	$1 ab \exp-\pi f_X^2a^2+f_Y^2b^2 \frac{1}{ ab } \exp \left[-\pi \left(\frac{f_X^2}{a^2} + \frac{f_Y^2}{b^2} \right) \right]$
$\exp[-\pi(a^2x^2+b^2y^2)]$	
$\text{rect}(ax)\text{rect}(by)$	$1 ab \text{sinc}(f_X/a)\text{sinc}(f_Y/b) \frac{1}{ ab } \text{sinc}(f_X/a) \text{sinc}(f_Y/b)$
$\Lambda(ax)\Lambda(by)$	$1 ab \text{sinc}^2(f_X/a)\text{sinc}^2(f_Y/b) \frac{1}{ ab } \text{sinc}^2(f_X/a) \text{sinc}^2(f_Y/b)$
$\delta(ax, by)$	$1 ab \frac{1}{ ab }$
$\exp[j\pi(ax+by)]$	$\delta(f_X - a/2, f_Y - b/2)$
$\text{sgn}(ax)\text{sgn}(by)$	$ab ab 1j\pi f_X 1j\pi f_Y \frac{ab}{ ab } \frac{1}{j\pi f_X} \frac{1}{j\pi f_Y}$
$\text{comb}(ax)\text{comb}(by)$	$1 ab \text{comb}(f_X/a)\text{comb}(f_Y/b) \frac{1}{ ab } \text{comb}(f_X/a) \text{comb}(f_Y/b)$
$\exp[j\pi(a^2x^2+b^2y^2)]$	
$\exp[j\pi(a^2x^2+b^2y^2)]$	$j ab \exp-j\pi f_X^2a^2+f_Y^2b^2 \frac{j}{ ab } \exp \left[-j\pi \left(\frac{f_X^2}{a^2} + \frac{f_Y^2}{b^2} \right) \right]$
$\exp[-(a x +b y)]$	$1ab21+(2\pi f_X/a)221+(2\pi f_Y/b)2 \frac{1}{ab} \frac{2}{1+(2\pi f_X/a)^2} \frac{2}{1+(2\pi f_Y/b)^2}$
$(a>0, b>0)$	

$$\mathcal{F}\{N^2\exp[-N^2\pi(x^2+y^2)]\}=\exp-\pi(f_X^2+f_Y^2)N^2.$$

$$\mathcal{F}\{N^2 \exp[-N^2 \pi(x^2 + y^2)]\} = \exp \left[-\frac{\pi(f_X^2 + f_Y^2)}{N^2} \right].$$

(2-6)

Accordingly the generalized transform of $\delta(x, y)$ is found to be

$$\mathcal{F}\{\delta(x, y)\} = \lim_{N \rightarrow \infty} \exp[-\pi(f_X^2 + f_Y^2)N^2] = 1.$$

$$\mathcal{F}[\delta(x, y)] = \lim_{N \rightarrow \infty} \left\{ \exp \left[-\frac{\pi(f_X^2 + f_Y^2)}{N^2} \right] \right\} = 1.$$

(2-7)

Note that the spectrum of a delta function extends uniformly over the entire frequency domain.

For other examples of generalized transforms, see [Table 2.1](#).

2.1.2 The Fourier Transform as a Decomposition

As mentioned previously, when dealing with linear systems it is often useful to decompose a complicated input into a number of more simple inputs, to calculate the response of the system to each of these “elementary” functions, and to superimpose the individual responses to find the total response. Fourier analysis provides the basic means of performing such a decomposition. Consider the familiar inverse transform relationship

$$g(t) = \int_{-\infty}^{\infty} G(f) \exp(j2\pi ft) df$$

$$g(t) = \int_{-\infty}^{\infty} G(f) \exp(j2\pi ft) df$$

(2-8)

expressing the time function g^8 in terms of its frequency spectrum. We may regard this expression as a decomposition of the function $g(t)^8$ into a linear combination (in this case an integral) of elementary functions, each with a specific form $\exp(j2\pi ft)$. From this it is clear that the complex number $G(f)^G(f)$ is simply a weighting factor that must be applied to the elementary function of frequency f^f in order to synthesize the desired $g(t)^g(t)$.

In a similar fashion, we may regard the *two-dimensional* Fourier transform as a decomposition of a function $g(x, y)^g(x, y)$ into a linear combination of elementary functions of the form $\exp[j2\pi(f_X x + f_Y y)]^{\exp[j2\pi(f_X x + f_Y y)]}$. Such functions have a number of interesting properties. Note that for any particular frequency pair $(f_X, f_Y)^{(f_X, f_Y)}$ the corresponding elementary function has a phase that is zero or an integer multiple of $2\pi^{2\pi}$ radians along lines described by the equation

$$y = -f_X f_Y x + n f_Y,$$

$$y = -\frac{f_X}{f_Y} x + \frac{n}{f_Y},$$

(2-9)

where n^n is an integer. Thus, as indicated in [Fig. 2.1](#), this elementary function may be regarded as being “directed” in the $(x, y)^{(x, y)}$ plane at an angle θ^θ (with respect to the x^x axis) given by

$$\theta = \arctan f_Y / f_X.$$

$$\theta = \arctan \left(\frac{f_Y}{f_X} \right).$$

(2-10)

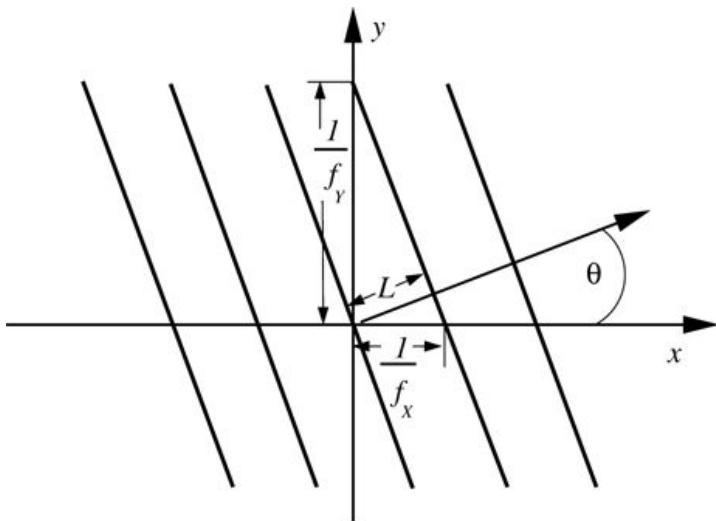


Figure 2.1

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 2.1 Lines of zero phase for the function $\exp[j2\pi(f_X x + f_Y y)]$.

The graph shows horizontal axis x and vertical axis y with 5 downward-sloping, equidistant, straight lines passing through the horizontal axis. Between two adjacent lines, the perpendicular distance is L and the distance along the x axis is $1/f_X$. Above the x axis, the lines extend up to a perpendicular height of $1/f_Y$ as marked on the y axis. An arrow extending in an upward slope in the first quadrant makes angle theta with the x axis.

In addition, the spatial period (i.e. the distance between zero-phase lines) is given by

$$L = 1/f_X^2 + f_Y^2.$$

$$L = \frac{1}{\sqrt{f_X^2 + f_Y^2}}.$$

(2-11)

In conclusion, then, we may again regard the inverse Fourier transform as providing a means for decomposing mathematical functions. The Fourier spectrum G of a function g is simply a description of the weighting factors that must be applied to each elementary function in order to synthesize the desired g . The real advantage obtained from using this decomposition will not be fully evident until our later discussion of invariant linear systems.

2.1.3 Fourier Transform Theorems

The basic definition (2-1) of the Fourier transform leads to a rich mathematical structure associated with the transform operation. We now consider a few of the basic mathematical properties of the transform, properties that will find wide use in later material. These properties are presented as mathematical theorems, followed by brief statements of their physical significance. Since these theorems are direct extensions of the analogous one-dimensional statements, the proofs are deferred to [Appendix A](#).

1. **Linearity theorem.** $\mathcal{F}\{\alpha g + \beta h\} = \alpha \mathcal{F}\{g\} + \beta \mathcal{F}\{h\}$; that is, the transform of a weighted sum of two (or more) functions is simply the identically weighted sum of their individual transforms.

2. **Similarity theorem.** If $\mathcal{F}\{g(x,y)\} = G(f_X, f_Y)$, then

$$\mathcal{F}\{g(ax, by)\} = 1/|ab|G(f_X/a, f_Y/b);$$

$$\mathcal{F}\{g(ax, by)\} = \frac{1}{|ab|}G\left(\frac{f_X}{a}, \frac{f_Y}{b}\right);$$

(2-12)

that is, a “stretch” of the coordinates in the space domain (x, y) results in a contraction of the coordinates in the frequency domain (f_X, f_Y) , plus a change in the overall amplitude of the spectrum.

3. **Shift theorem.** If $\mathcal{F}\{g(x,y)\} = G(f_X, f_Y)$, then

$$\mathcal{F}\{g(x-a, y-b)\} = G(f_X, f_Y) \exp[-j2\pi(f_X a + f_Y b)];$$

$$\mathcal{F}\{g(x - a, y - b)\} = G(f_X, f_Y) \exp[-j2\pi(f_X a + f_Y b)];$$

(2-13)

that is, translation in the space domain introduces a linear phase shift in the frequency domain.

4. **Rayleigh's theorem (Parseval's theorem).** If $\mathcal{F}\{g(x,y)\} = G(f_X, f_Y)$, then

$$\int_{-\infty}^{\infty} \int |g(x,y)|^2 dx dy = \int_{-\infty}^{\infty} \int |G(f_X, f_Y)|^2 df_X df_Y.$$

$$\int_{-\infty}^{\infty} \int |g(x, y)|^2 dx dy = \int_{-\infty}^{\infty} \int |G(f_X, f_Y)|^2 df_X df_Y.$$

(2-14)

The integral on the left-hand side of this theorem can be interpreted as the energy contained in the waveform $g(x,y)$. This in turn leads us to the idea that the quantity $|G(f_X, f_Y)|^2$ can be interpreted as an energy density in the frequency domain.

5. **Convolution theorem.** If $\mathcal{F}\{g(x,y)\} = G(f_X, f_Y)$ and $\mathcal{F}\{h(x,y)\} = H(f_X, f_Y)$, then

$$\mathcal{F}\int_{-\infty}^{\infty} \int g(\xi, \eta) h(x - \xi, y - \eta) d\xi d\eta = G(f_X, f_Y) H(f_X, f_Y).$$

$$\mathcal{F}\left\{ \int_{-\infty}^{\infty} \int g(\xi, \eta) h(x - \xi, y - \eta) d\xi d\eta \right\} = G(f_X, f_Y) H(f_X, f_Y).$$

(2-15)

The convolution of two functions in the space domain (an operation that will be found to arise frequently in the theory of linear systems) is entirely equivalent to the simpler operation of multiplying their individual transforms and inverse transforming.

6. **Autocorrelation theorem.** If $\mathcal{F}\{g(x,y)\} = G(f_X, f_Y)$, then

$$\mathcal{F}\int_{-\infty}^{\infty} \int g(\xi, \eta) g^*(\xi - x, \eta - y) d\xi d\eta = |G(f_X, f_Y)|^2.$$

$$\mathcal{F}\left\{ \int_{-\infty}^{\infty} \int g(\xi, \eta) g^*(\xi - x, \eta - y) d\xi d\eta \right\} = |G(f_X, f_Y)|^2.$$

(2-16)

Similarly,

$$\mathcal{F}|g(x,y)|^2 = \int_{-\infty}^{\infty} \int G(\xi, \eta) G^*(\xi - f_X, \eta - f_Y) d\xi d\eta.$$

$$\mathcal{F}\{|g(x, y)|^2\} = \int_{-\infty}^{\infty} \int G(\xi, \eta) G^*(\xi - f_X, \eta - f_Y) d\xi d\eta.$$

(2-17)

This theorem may be regarded as a special case of the convolution theorem in which we convolve $g(x, y)$ with $g^*(-x, -y)$.

7. **Rotation theorem.** Let $\mathcal{F}\{g(r, \theta)\} = G(\rho, \phi)$, with (r, θ) being radius and angle in the space domain and (ρ, ϕ) being radius and angle in the frequency domain. Then a rotation $g(r, \theta + \theta_0)$ by angle θ_0 in the space plane results in an identical rotation $G(\rho, \phi + \theta_0)$ in the frequency plane. In rectangular coordinates,

$$\mathcal{F}\{g(x \cos \theta - y \sin \theta, x \sin \theta + y \cos \theta)\} = G(f_X \cos \theta - f_Y \sin \theta, f_X \sin \theta + f_Y \cos \theta).$$

$$\mathcal{F}\{g(x \cos \theta - y \sin \theta, x \sin \theta + y \cos \theta)\} = G(f_X \cos \theta - f_Y \sin \theta, f_X \sin \theta + f_Y \cos \theta).$$

(2-18)

8. Shear theorem. The function $g(x+by, y)$ represents a horizontal shear of the function $g(x, y)$, while the function $g(x,y+cx)$ represents a vertical shear of $g(x, y)$. If $\mathcal{F}\{g(x,y)\}=G(f_X,f_Y)$, then

$$\mathcal{F}\{g(x+by,y)\}=G(f_X,f_Y-bf_X)$$

$$\mathcal{F}\{g(x + by, y)\} = G(f_X, f_Y - bf_X)$$

and

$$\mathcal{F}\{g(x,y+cx)\}=G(f_X-cf_Y,f_Y).$$

$$\mathcal{F}\{g(x, y + cx)\} = G(f_X - cf_Y, f_Y).$$

Thus a horizontal shear in the space domain results in a vertical shear in the frequency domain, while a vertical shear in the space domain results in a horizontal shear in the frequency domain.

9. Fourier integral theorem. At each point of continuity of g^S ,

$$\mathcal{F}^{-1}g(x,y)=\mathcal{F}^{-1}\mathcal{F}g(x,y)=g(x,y).$$

$$\mathcal{F}^{-1}\{g(x, y)\} = \mathcal{F}^{-1}\mathcal{F}\{g(x, y)\} = g(x, y).$$

(2-19)

At each point of discontinuity of g^S , the two successive transforms yield the angular average of the values of g^S in a small neighborhood of that point. That is, the successive transformation and inverse transformation of a function yields that function again, except at points of discontinuity.

The above transform theorems are of far more than just theoretical interest. They will be used frequently, since they provide the basic tools for the manipulation of Fourier transforms and can save enormous amounts of work in the solution of Fourier analysis problems.

2.1.4 Separable Functions

A function of two independent variables is called *separable* with respect to a specific coordinate system if it can be written as a product of two functions, each of which depends on only one of the independent variables. Thus the function g^S is separable in rectangular coordinates (x, y) if

$$g(x,y)=g_X(x)g_Y(y),$$

$$g(x, y) = g_X(x) g_Y(y),$$

(2-20)

while it is separable in polar coordinates (r, θ) if

$$g(r,\theta)=g_R(r)g_\Theta(\theta).$$

$$g(r, \theta) = g_R(r) g_\theta(\theta).$$

(2-21)

Separable functions are often more convenient to deal with than more general functions, for separability often allows complicated two-dimensional manipulations to be reduced to simpler one-dimensional manipulations. For example, a function separable in rectangular coordinates has the particularly simple property that its two-dimensional Fourier transform can be found as a product of one-dimensional Fourier transforms, as evidenced by the following relation:

$$\mathcal{F}g(x,y) = \iint_{-\infty}^{\infty} g(x,y) \exp[-j2\pi(f_X x + f_Y y)] dx dy = \int_{-\infty}^{\infty} g_X(x) \exp[-j2\pi f_X x] dx \int_{-\infty}^{\infty} g_Y(y) \exp[-j2\pi f_Y y] dy = \mathcal{F}_X\{g_X\} \mathcal{F}_Y\{g_Y\}.$$

$$\begin{aligned} \mathcal{F}\{g(x, y)\} &= \iint_{-\infty}^{\infty} g(x, y) \exp[-j2\pi(f_X x + f_Y y)] dx dy \\ &= \int_{-\infty}^{\infty} g_X(x) \exp[-j2\pi f_X x] dx \int_{-\infty}^{\infty} g_Y(y) \exp[-j2\pi f_Y y] dy \\ &= \mathcal{F}_X\{g_X\} \mathcal{F}_Y\{g_Y\}. \end{aligned}$$

(2-22)

Thus the transform of g is itself separable into a product of two factors, one a function of f_X only and the second a function of f_Y only, and the process of two-dimensional transformation simplifies to a succession of more familiar one-dimensional manipulations.

Functions separable in polar coordinates are not so easily handled as those separable in rectangular coordinates, but it is still generally possible to demonstrate that two-dimensional manipulations can be performed by a series of one-dimensional manipulations. For example, the reader is asked to verify in the problems that the Fourier transform of a general function separable in polar coordinates can be expressed as an infinite sum of weighted *Hankel* transforms

$$\begin{aligned} \mathcal{F}\{g(r, \theta)\} &= \sum_{k=-\infty}^{\infty} c_k (-j)^k \exp(jk\phi) \mathcal{H}_k\{g_R(r)\} \\ \mathcal{F}\{g(r, \theta)\} &= \sum_{k=-\infty}^{\infty} c_k (-j)^k \exp(jk\phi) \mathcal{H}_k\{g_R(r)\} \end{aligned}$$

(2-23)

where

$$c_k = \frac{1}{2\pi} \int_0^{2\pi} g_\theta(\theta) \exp(-jk\theta) d\theta$$

$$c_k = \frac{1}{2\pi} \int_0^{2\pi} g_\theta(\theta) \exp(-jk\theta) d\theta$$

and $\mathcal{H}_k\{\cdot\}$ is the Hankel transform operator of order k , defined by

$$\mathcal{H}_k g_R(r) = 2\pi \int_0^\infty r g_R(r) J_k(2\pi r\rho) dr.$$

$$\mathcal{H}_k \{g_R(r)\} = 2\pi \int_0^\infty r g_R(r) J_k(2\pi r \rho) dr.$$

(2-24)

The function J_k is the k th-order Bessel function of the first kind.

2.1.5 Functions with Circular Symmetry: Fourier-Bessel Transforms

Perhaps the simplest class of functions separable in polar coordinates is composed of those possessing *circular symmetry*. The function g^g is said to be circularly symmetric if it can be written as a function of r^r alone, that is,

$$g(r, \theta) = g_R(r).$$

$$g(r, \theta) = g_R(r).$$

(2-25)

Such functions play an important role in the problems of interest here, since most optical systems have precisely this type of symmetry. We accordingly devote special attention to the problem of Fourier transforming a circularly symmetric function.

The Fourier transform of g^g in a system of rectangular coordinates is, of course, given by

$$G(f_X, f_Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) \exp[-j2\pi(f_X x + f_Y y)] dx dy.$$

$$G(f_X, f_Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) \exp[-j2\pi(f_X x + f_Y y)] dx dy.$$

(2-26)

To fully exploit the circular symmetry of g^g , we make a transformation to polar coordinates in both the (x, y) and the (f_X, f_Y) planes as follows:

$$r = \sqrt{x^2 + y^2} \quad x = r \cos \theta \\ \theta = \arctan \left(\frac{y}{x} \right) \quad y = r \sin \theta \\ \rho = \sqrt{f_X^2 + f_Y^2} \quad f_X = \rho \cos \phi \\ \phi = \arctan \left(\frac{f_Y}{f_X} \right) \quad f_Y = \rho \sin \phi.$$

$$r = \sqrt{x^2 + y^2} \quad x = r \cos \theta \\ \theta = \arctan \left(\frac{y}{x} \right) \quad y = r \sin \theta \\ \rho = \sqrt{f_X^2 + f_Y^2} \quad f_X = \rho \cos \phi \\ \phi = \arctan \left(\frac{f_Y}{f_X} \right) \quad f_Y = \rho \sin \phi.$$

(2-27)

For the present we write the transform as a function of both radius and angle,³

$$\mathcal{F}g = G(\rho, \phi).$$

$$\mathcal{F}\{g\} = G_o(\rho, \phi).$$

(2-28)

Applying the coordinate transformations (2-27) to Eq.(2-26), the Fourier transform of g^g can be written

$$G_o(\rho, \phi) = \int_0^{2\pi} d\theta \int_0^\infty dr r g_R(r) \exp[-j2\pi r\rho(\cos\theta\cos\phi + \sin\theta\sin\phi)]$$

$$G_o(\rho, \phi) = \int_0^\infty dr r g_R(r) \int_0^{2\pi} d\theta \exp[-j2\pi r\rho(\cos\theta\cos\phi + \sin\theta\sin\phi)]$$

(2-29)

or equivalently,

$$G_o(\rho, \phi) = \int_0^\infty dr r g_R(r) \int_0^{2\pi} d\theta \exp[-j2\pi r\rho\cos(\theta - \phi)].$$

$$(2-30)$$

Finally we use the Bessel function identity

$$J_0(a) = \frac{1}{2\pi} \int_0^{2\pi} \exp[-ja\cos(\theta - \phi)] d\theta,$$

$$(2-31)$$

where J_0 is a Bessel function of the first kind, zero order, to simplify the expression for the transform. Substituting (2-31) in (2-30), the dependence of the transform on angle ϕ is seen to disappear, leaving G_o as the following function of radius ρ ,

$$G_o(\rho, \phi) = G_o(\rho) = 2\pi \int_0^\infty r g_R(r) J_0(2\pi r\rho) dr.$$

$$(2-32)$$

Thus the Fourier transform of a circularly symmetric function is itself circularly symmetric and can be found by performing the one-dimensional manipulation of (2-32). This particular form of the Fourier transform occurs frequently enough to warrant a special designation; it is accordingly referred to as the *Fourier-Bessel transform*, or alternatively as the *Hankel transform of zero order* (cf. (2-24)). For brevity, we adopt the former terminology.

By means of arguments identical with those used above, the *inverse* Fourier transform of a circularly symmetric spectrum $G_o(\rho)$ can be expressed as

$$gR(r) = 2\pi \int_0^\infty \rho G_o(\rho) J_0(2\pi r\rho) d\rho.$$

$$g_R(r) = 2\pi \int_0^\infty \rho G_o(\rho) J_0(2\pi r\rho) d\rho.$$

(2-33)

Thus for circularly symmetric functions there is no difference between the transform and the inverse-transform operations.

Using the notation $\mathcal{B}\{\cdot\}$ to represent the Fourier-Bessel transform operation, it follows directly from the Fourier integral theorem that

$$\mathcal{B}^{-1}gR(r) = \mathcal{B}^{-1}\mathcal{B}gR(r) = \mathcal{B}\mathcal{B}gR(r) = gR(r)$$

$$\mathcal{B}B^{-1}\{g_R(r)\} = \mathcal{B}^{-1}\mathcal{B}\{g_R(r)\} = \mathcal{B}\mathcal{B}\{g_R(r)\} = g_R(r)$$

(2-34)

at each value of r where $gR(r)$ is continuous. In addition, the *similarity* theorem can be straightforwardly applied (see [Prob. 2-6c](#)) to show that

$$\mathcal{B}gR(ar) = 1/a^2 G_o(a).$$

$$\mathcal{B}\{g_R(ar)\} = \frac{1}{a^2} G_o\left(\frac{\rho}{a}\right).$$

(2-35)

When using the expression [\(2-32\)](#) for the Fourier-Bessel transform, the reader should remember that it is no more than a special case of the two-dimensional Fourier transform, and therefore any familiar property of the Fourier transform has an entirely equivalent counterpart in the terminology of Fourier-Bessel transforms.

2.1.6 Some Frequently Used Functions and Some Useful Fourier Transform Pairs

A number of mathematical functions will find such extensive use in later material that considerable time and effort can be saved by assigning them special notations of their own. Accordingly, we adopt the following definitions of some frequently used functions:

Rectangle function

$$\text{rect}(x) = 1 \quad |x| < 1/2 \quad 1/2 \leq |x| \leq 1 \quad 0 \quad \text{otherwise}$$

$$\text{rect}(x) = \begin{cases} 1 & |x| < \frac{1}{2} \\ \frac{1}{2} & |x| = \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

Sinc function

$$\text{sinc}(x) = \sin(\pi x)/\pi x$$

$$\text{sinc}(x) = \frac{\sin(\pi x)}{\pi x}$$

Signum function

$$\text{sgn}(x) = 1 \text{ if } x > 0 \\ 0 \text{ if } x = 0 \\ -1 \text{ if } x < 0$$

$$\text{sgn}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}$$

Triangle function

$$\Lambda(x) = 1 - |x| \text{ if } |x| \leq 1 \\ 0 \text{ otherwise}$$

$$\Lambda(x) = \begin{cases} 1 - |x| & |x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Comb function

$$\text{comb}(x) = \sum_{n=-\infty}^{\infty} \delta(x - n)$$

$$\text{comb}(x) = \sum_{n=-\infty}^{\infty} \delta(x - n)$$

Circle function

$$\text{circ}(x^2 + y^2) = 1 \text{ if } x^2 + y^2 < 1 \\ 1/2 \text{ if } x^2 + y^2 = 1 \\ 0 \text{ otherwise}$$

$$\text{circ}(\sqrt{x^2 + y^2}) = \begin{cases} 1 & \sqrt{x^2 + y^2} < 1 \\ 1/2 & \sqrt{x^2 + y^2} = 1 \\ 0 & \text{otherwise} \end{cases}$$

The first five of these functions, depicted in [Fig. 2.2](#), are all functions of only one independent variable; however, a variety of separable functions can be formed in two dimensions by means of products of these functions. The circle function is, of course, unique to the case of two-dimensional variables; see [Fig. 2.3](#) for an illustration of its structure.

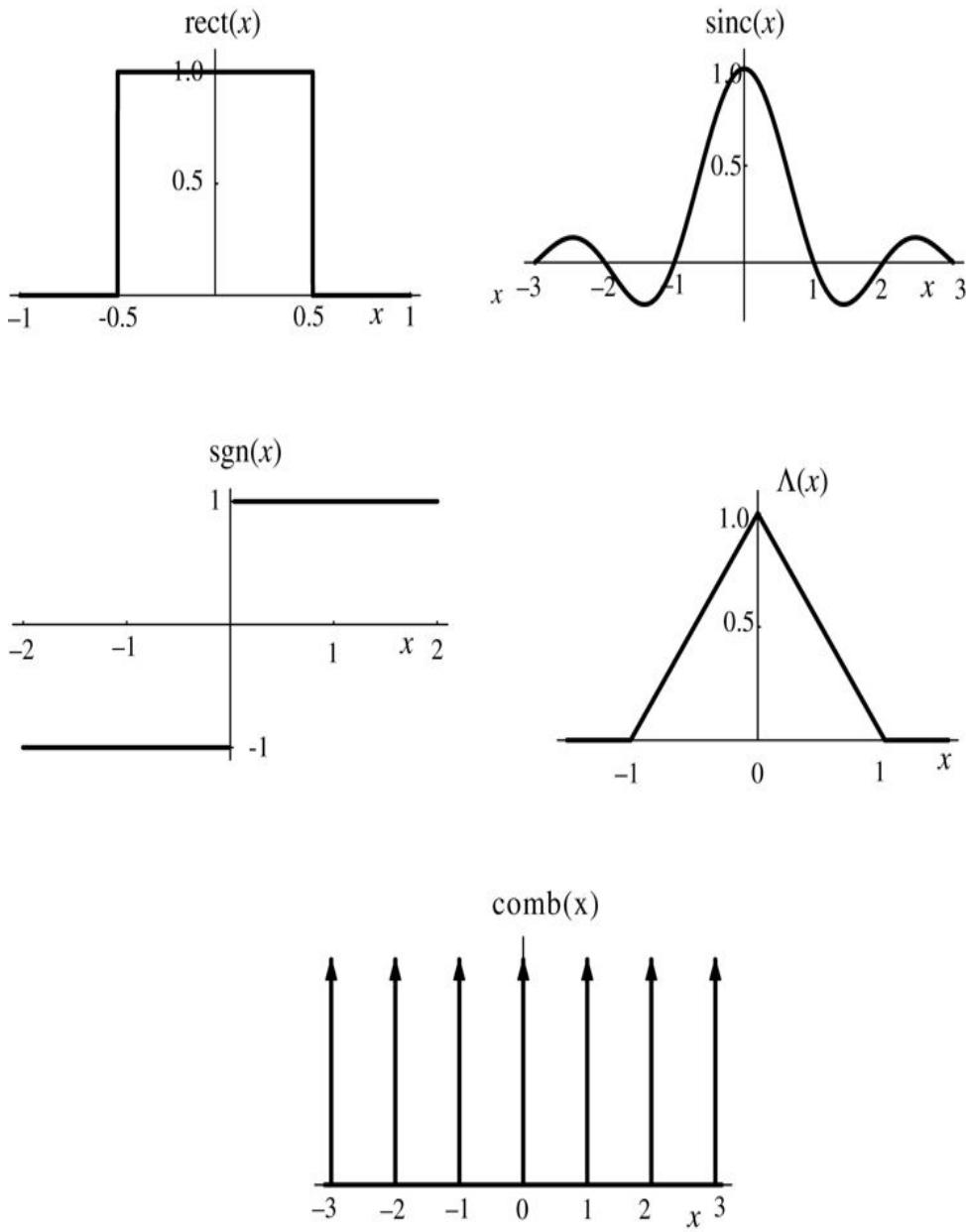


Figure 2.2
Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 2.2 Special functions.

“The 5 graphs are as follows. The first graph plots rectangular function (x) on the vertical axis, marked from 0 to 1, and horizontal axis x , marked from minus 1 to +1. The symmetric graph line has three horizontal lines and two vertical lines that connect the following points in the given order: (minus 1, 0), (minus 0.5, 0), (minus 0.5, 1), (0, 1), (0.5, 1), (0.5, 0), and (1, 0). The second graph plots sinc function (x) on the vertical axis, marked from 0 to 1, and horizontal axis, marked from minus 3 to +3. The wavelike graph begins at minus 3, rises in an upward slope and then turns downward to intersect the x axis and then rise again through the minus 1 mark, and then in a steep upward slope reach the 1 mark on the y axis. The path thus far is mirrored across the vertical axis on the other side. The third graph plots sign function (x) on the vertical axis, marked from minus 1

to 1, and horizontal axis x, marked from minus 2 to +2. The graph is a pair of straight lines parallel to the x axis, one extending to the right from sign (x) = 1 and the other extending to the left from sign (x) = minus 1. The fourth graph plots lambda function (x) on the vertical axis, marked from 0 to 1, and horizontal axis x, marked from minus 1 to +1. The graph extends along the x axis up to minus 1 and then in a straight upward slope extends up to the 1 mark on the vertical axis. The path thus far is mirrored across the vertical axis on the other side. The fourth graph plots comb function (x) on the vertical axis. The graph is a series of seven upward pointing equidistant arrows of equal length standing perpendicular at minus 3, minus 2, minus 1, 0 123 on the horizontal axis.”

We conclude our discussion of Fourier analysis by presenting some specific two-dimensional transform pairs. [Table 2.1](#) lists a number of transforms of functions separable in rectangular coordinates. For the convenience of the reader, the functions are presented with arbitrary scaling constants. Since the transforms of such functions can be found directly from products of familiar one-dimensional transforms, the proofs of these relations are left to the reader (cf. [Prob. 2-2](#)).

On the other hand, with a few exceptions (e.g. $\exp[-\pi(x^2+y^2)]$), which is both separable in rectangular coordinates and circularly symmetric), transforms of most circularly symmetric functions cannot be found simply from a knowledge of one-dimensional transforms. The most frequently encountered function with circular symmetry is:

$$\text{circ}(r) = \begin{cases} 1 & r < 1 \\ \frac{1}{2} & r = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{circ}(r) = \begin{cases} 1 & r < 1 \\ \frac{1}{2} & r = 1 \\ 0 & \text{otherwise} \end{cases}.$$

Accordingly, some effort is now devoted to finding the transform of this function. Using the Fourier-Bessel transform expression ([2-32](#)), the transform of the circle function can be written

$$\mathcal{B}\{\text{circ}(r)\} = 2\pi \int_0^1 r J_0(2\pi r\rho) dr.$$

$$\mathcal{B}\{\text{circ}(r)\} = 2\pi \int_0^1 r J_0(2\pi r\rho) dr.$$

Using a change of variables $r' = 2\pi r\rho$ and the identity

$$\int_0^\infty x \xi J_0(\xi) d\xi = x J_1(x),$$

$$\int_0^\infty \xi J_0(\xi) d\xi = x J_1(x),$$

we rewrite the transform as

$$\mathcal{B}\{\text{circ}(r)\} = 2\pi \rho \int_0^\infty 2\pi r \rho r' J_0(r') dr' = J_1(2\pi\rho)\rho,$$

$$\mathcal{B}\{\text{circ}(r)\} = \frac{1}{2\pi\rho^2} \int_0^{2\pi\rho} r' J_0(r') dr' = \frac{J_1(2\pi\rho)}{\rho},$$

(2-36)

where J_1 is a Bessel function of the first kind, order 1. [Figure 2.3](#) illustrates the circle function and its transform. Note that the transform is circularly symmetric, as expected, and consists of a central lobe and a series of concentric rings of diminishing amplitude. Its value at the origin is $\pi \pi$. As a matter of curiosity we note that the zeros of this transform are not equally spaced in radius. A convenient normalized version of this function, with value unity at the origin, is $2J_1(2\pi\rho)2\pi\rho$

$$2 \frac{J_1(2\pi\rho)}{2\pi\rho}$$

This particular function is called the “besinc” function or the “jinc” function.

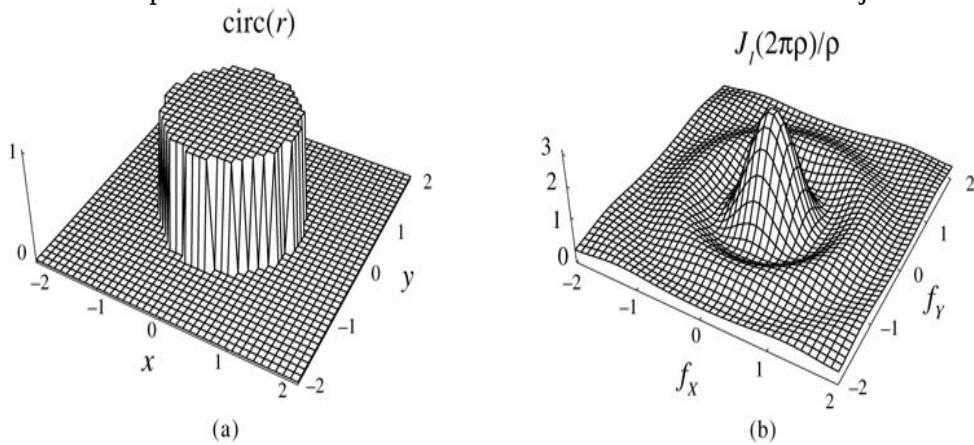


Figure 2.3

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 2.3 The circle function and its transform.

The projection in illustration a is cylindrical, marking circular function (r) rising along a vertical axis marked from 0 to 1 at the center of a square horizontal plane whose adjoining sides x and y are axes marked from minus 2 to +2. The projection in illustration b is a cone set in the middle of a ring. It marks the values of $J_1(2\pi\rho)/\rho$ rising along a vertical axis marked from 0 to 3 at the center of a square horizontal plane whose adjoining sides f_x and f_y are axes marked from minus 2 to +2.

For a number of additional Fourier-Bessel transform pairs, the reader is referred to the problems (see [Prob. 2-6](#)).

2.2 Spatial Frequency and Space-Frequency Localization

Each Fourier component of a function is a complex exponential of a unique spatial frequency. As such, every frequency component extends over the entire (x, y) domain. Therefore it is not possible to associate a spatial location with a particular spatial frequency. Nonetheless, we know that in practice certain portions of an image could contain parallel grid lines at a certain fixed spacing, and we are tempted to say that the particular frequency or frequencies represented by these grid lines are localized to certain spatial regions of the image. To help resolve this dilemma we introduce the idea of local spatial frequencies and their relation to Fourier components.

2.2.1 Local Spatial Frequencies

For the purpose of this discussion, we consider the general case of complex-valued functions, which we will later see represent the amplitude and phase distributions of monochromatic optical waves. For now, they are just complex functions. Any such function can be represented in the form

$$g(x, y) = a(x, y) \exp[j\phi(x, y)]$$

$$g(x, y) = a(x, y) \exp [j\phi(x, y)]$$

(2-37)

where $a(x, y)$ is a real and nonnegative amplitude distribution, while $\phi(x, y)$ is a real phase distribution. For this discussion we assume that the amplitude distribution $a(x, y)$ is a slowly varying function of (x, y) , so that we can concentrate on the behavior of the phase function $\phi(x, y)$.

We define the *local spatial frequency* of the function g^8 as a space-dependent frequency pair $f_X(\ell), f_Y(\ell)$ given by

$$f_X(\ell)(x, y) = 12\pi\partial\partial_x\phi(x, y), \quad f_Y(\ell)(x, y) = 12\pi\partial\partial_y\phi(x, y),$$

$$f_X^{(\ell)}(x, y) = \frac{1}{2\pi} \frac{\partial}{\partial x} \phi(x, y), \quad f_Y^{(\ell)}(x, y) = \frac{1}{2\pi} \frac{\partial}{\partial y} \phi(x, y),$$

(2-38)

or, in vector notation,

$$\mathbf{f}^{(\ell)} = 12\pi\nabla\phi(\mathbf{x}),$$

$$\vec{f}^{(\ell)} = \frac{1}{2\pi} \nabla \phi(\vec{x}),$$

(2-39)

where $f \rightarrow (\ell) = (f_X(\ell), f_Y(\ell))$ $\vec{f}^{(\ell)} = (f_X^{(\ell)}, f_Y^{(\ell)})$, $x \rightarrow = (x, y)$ $\vec{x} = (x, y)$ and $\nabla \nabla$ represents the gradient operation. In addition, $f \rightarrow (\ell) \vec{f}^{(\ell)}$ is defined to be zero in regions where the function $g(x, y)$ vanishes.

It is possible to define a single local spatial frequency having both a space dependent period $P(x, y)$ and a space-dependent angle $\psi(x, y)$, measured with respect to the x -axis,

$$P(x, y) = \sqrt{f_X^2 + f_Y^2}$$

$$P(x, y) = \sqrt{f_X^{(\ell)}^2 + f_Y^{(\ell)}^2}$$

(2-40)

and

$$\psi(x, y) = \arctan \frac{f_Y}{f_X}$$

$$\psi(x, y) = \arctan \left(\frac{f_Y^{(\ell)}}{f_X^{(\ell)}} \right)$$

(2-41)

Consider the result of applying these definitions to the particular complex function

$$g(x, y) = \exp[j2\pi(f_X x + f_Y y)]$$

$$g(x, y) = \exp[j2\pi(f_X x + f_Y y)]$$

representing a simple linear-phase exponential of frequencies (f_X, f_Y) . We obtain

$$f_X(\ell)(x, y) = 12\pi\partial_x 2\pi(f_X x + f_Y y) = f_X f_Y(\ell)(x, y) = 12\pi\partial_y 2\pi(f_X x + f_Y y) = f_Y.$$

$$\begin{aligned} f_X^{(\ell)}(x, y) &= \frac{1}{2\pi} \frac{\partial}{\partial x} [2\pi(f_X x + f_Y y)] = f_X \\ f_Y^{(\ell)}(x, y) &= \frac{1}{2\pi} \frac{\partial}{\partial y} [2\pi(f_X x + f_Y y)] = f_Y. \end{aligned}$$

(2-42)

Thus we see that for the case of a single Fourier component, the local frequencies do indeed reduce to the frequencies of that component, and those frequencies are constant over the entire (x, y) plane.

Next consider a space-limited version of a quadratic-phase exponential function,⁴ which we call a “finite chirp” function,⁵

$$g(x,y) = \exp[j\pi\beta(x^2 + y^2)] \operatorname{rect}\left(\frac{x}{L_X}\right) \operatorname{rect}\left(\frac{y}{L_Y}\right).$$

$$g(x, y) = \exp[j\pi\beta(x^2 + y^2)] \operatorname{rect}\left(\frac{x}{L_X}\right) \operatorname{rect}\left(\frac{y}{L_Y}\right). \quad (2-43)$$

Performing the differentiations called for by the definitions of local frequencies, we find that they can be expressed as

$$f_X(\ell)(x,y) = \beta x \operatorname{rect}(x/L_X) \quad f_Y(\ell)(x,y) = \beta y \operatorname{rect}(y/L_Y).$$

$$f_X^{(\ell)}(x, y) = \beta x \operatorname{rect}\left(\frac{x}{L_X}\right) \quad f_Y^{(\ell)}(x, y) = \beta y \operatorname{rect}\left(\frac{y}{L_Y}\right). \quad (2-44)$$

We see that in this case the local spatial frequencies *do* depend on location in the (x,y) plane; within a rectangle of dimensions $L_X \times L_Y$, $f_X(\ell)$ varies linearly with the x -coordinate while $f_Y(\ell)$ varies linearly with the y -coordinate. Thus there is indeed a dependence of local spatial frequency on position in the (x,y) plane.⁶

Since the local spatial frequencies are bounded to covering a rectangle of dimensions $L_X \times L_Y$, it would be tempting to conclude that the Fourier spectrum of $g(x,y)$ is also limited to the same rectangular region. In fact this is generally not true. The shape of the spectrum depends fundamentally on the product $(L_X/2)^2 \beta$, which later will be seen to be a quantity known as the *Fresnel number*. For Fresnel numbers greater than 1, the local spatial frequency distribution can yield good estimates of the shape and extent of the Fourier spectrum, but for Fresnel numbers less than 1 it does not. Under such conditions, the local spatial frequencies are changing too rapidly with the spatial coordinate to allow this relationship between local spatial frequency and spatial frequency to be accurate.

The Fourier transform of the finite chirp function function is given by the expression

$$G(f_X, f_Y) = \int_{-L_X/2}^{L_X/2} \int_{-L_Y/2}^{L_Y/2} e^{j\pi\beta(x^2 + y^2)} e^{-j2\pi(f_X x + f_Y y)} dx dy.$$

$$G(f_X, f_Y) = \int_{-L_X/2}^{L_X/2} \int_{-L_Y/2}^{L_Y/2} e^{j\pi\beta(x^2 + y^2)} e^{-j2\pi(f_X x + f_Y y)} dx dy.$$

This expression is separable in rectangular coordinates, so it suffices to find the one-dimensional spectrum

$$G_X(f_X) = \int_{-L_X/2}^{L_X/2} e^{j\pi\beta x^2} e^{-j2\pi f_X x} dx.$$

$$G_X(f_X) = \int_{-L_X/2}^{L_X/2} e^{j\pi\beta x^2} e^{-j2\pi f_X x} dx.$$

(2-45)

Completing the square in the exponent and making a change of variables of integration from x

$$t = \sqrt{2\beta} \left(x - \frac{f_X}{\beta} \right)$$

to $t=2\beta x-f_X\beta$ yields

$$G_X(f_X) = 12\beta e^{-j\pi f_X 2/\beta} - 2\beta L_X/2 + f_X/\beta 2\beta L_X/2 - f_X/\beta \exp j\pi t 22 dt.$$

$$G_X(f_X) = \frac{1}{\sqrt{2\beta}} e^{-j\pi f_X^2/\beta} \int_{-\sqrt{2\beta}(L_X/2 - f_X/\beta)}^{\sqrt{2\beta}(L_X/2 - f_X/\beta)} \exp \left[j\frac{\pi t^2}{2} \right] dt.$$

This integral can be expressed in terms of tabulated functions, the Fresnel integrals, which are defined by

$$C(z) = \int_0^z \cos \left(\frac{\pi t^2}{2} \right) dt \quad S(z) = \int_0^z \sin \left(\frac{\pi t^2}{2} \right) dt.$$

(2-46)

The spectrum G_X can then be expressed as

$$G_X(f_X) = e^{-j\pi f_X 2/\beta} 2\beta C 2\beta L_X/2 - f_X/\beta - C 2\beta L_X/2 - f_X/\beta + jS 2\beta L_X/2 - f_X/\beta - jS 2\beta L_X/2 - f_X/\beta.$$

$$G_X(f_X) = \frac{e^{-j\pi f_X^2/\beta}}{\sqrt{2\beta}} \left\{ C \left[\sqrt{2\beta} (L_X/2 - f_X/\beta) \right] - C \left[\sqrt{2\beta} (-L_X/2 - f_X/\beta) \right] + jS \left[\sqrt{2\beta} (L_X/2 - f_X/\beta) \right] - jS \left[\sqrt{2\beta} (-L_X/2 - f_X/\beta) \right] \right\}.$$

(2-47)

The expression for G_Y is of course identical, except the Y subscript replaces the X subscript. [Figure 2.4](#) shows a plot of $|G_X(f_X)|$ vs. f_X for the particular case of $L_X=20$, $\beta=1$ (a Fresnel number of 100), which is a case in which the local spatial frequency distribution is only somewhat different from the Fourier spectrum. Good agreement between the Fourier spectrum and the distribution of local spatial frequencies can be expected only when the variations of $\phi(x,y)$ are sufficiently “slow” in the (x,y) plane to allow $\phi(x,y)$ to be well approximated by only three terms of its Taylor series expansion about any point (x,y) , i.e. a constant term and two first-partial-derivative terms. See [Section 5.2](#) for a much more detailed discussion of the spectrum of the space-limited quadratic-phase function.

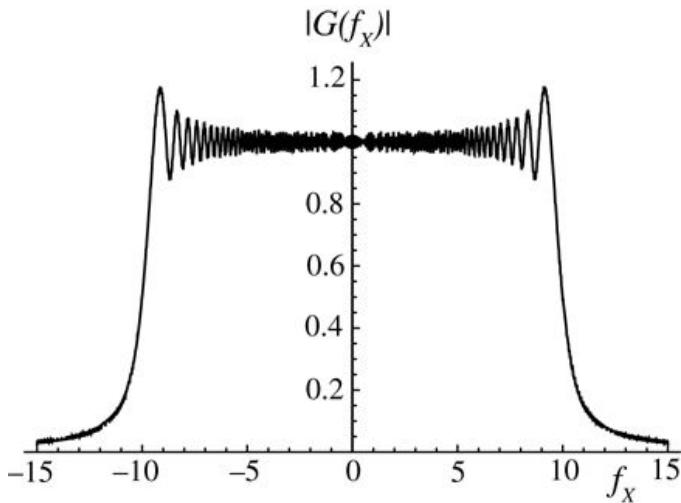


Figure 2.4

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 2.4 The spectrum of the finite chirp function, $L_X = 20$, $\beta = 1$.

The graph, plotting f_x along the horizontal axis and $|G(f_x)|$ along the vertical axis, shows a curve that rises in a gentle rightward slope up to $(-12, 0.1)$ and then rises almost vertically moving slightly rightward up to a height corresponding to the 1.0 mark on the vertical axis. The curve then extends horizontally rightward, dropping and rising in a regular wavelike pattern between the levels 0.9 and 1.1 on the vertical axis. The pattern grows denser and shorter as it approaches the 1.0 mark on the vertical axis, which mirrors the path so far on to the other side.

Local spatial frequencies are of special physical significance in optics. When the local spatial frequencies of the complex amplitude of a coherent optical wavefront are found at a particular point on the wavefront and multiplied by λ , the results correspond to the direction cosines of the geometrical optics ray at that particular point on the wavefront. If we consider the *occupancy distribution* of local spatial frequencies in the (f_x, f_y) plane, this distribution can be considered to be a *geometrical optics* estimate of the spectrum and will not include any diffraction effects. However, we are getting ahead of ourselves; we will return to this idea in later chapters and particularly in [Appendix B](#).

2.2.2 The Wigner Distribution Function

The concept of space-bandwidth occupancy can be put on a firmer footing by introducing the *Wigner distribution* [370], [18], [19], [274], [4]. Whereas the local frequencies will later be found to be analogous to the ray directions of geometrical optics at each point in space, the Wigner distribution provides a treatment that is valid within the formalism of wave optics.

The two-dimensional Wigner distribution $W_g(x, y; f_x, f_y)$ of a function $g(x, y)$ is defined on the four-dimensional space-frequency domain as

$$W_g(x, y; f_x, f_y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x + \Delta x, y + \Delta y) g^*(x - \Delta x, y - \Delta y) e^{-j2\pi(\Delta x f_x + \Delta y f_y)} d\Delta x d\Delta y.$$

$$W_g(x, y; f_X, f_Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g\left(x + \frac{\Delta x}{2}, y + \frac{\Delta y}{2}\right) g^*\left(x - \frac{\Delta x}{2}, y - \frac{\Delta y}{2}\right) \\ \times e^{-j2\pi(\Delta x f_X + \Delta y f_Y)} d\Delta x d\Delta y.$$

(2-48)

Replacing g and g^* by their Fourier integrals, using the shift theorem and the sifting property of delta functions, we can obtain the equivalent expression

$$Wg(x,y;fX,fY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} G\left(f_X + \frac{\Delta f_X}{2}, f_Y + \frac{\Delta f_Y}{2}\right) G^*\left(f_X - \frac{\Delta f_X}{2}, f_Y - \frac{\Delta f_Y}{2}\right) \\ \times e^{-j2\pi(x\Delta f_X + y\Delta f_Y)} d\Delta f_X d\Delta f_Y,$$

$$W_g(x, y; f_X, f_Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} G\left(f_X + \frac{\Delta f_X}{2}, f_Y + \frac{\Delta f_Y}{2}\right) G^*\left(f_X - \frac{\Delta f_X}{2}, f_Y - \frac{\Delta f_Y}{2}\right) \\ \times e^{-j2\pi(x\Delta f_X + y\Delta f_Y)} d\Delta f_X d\Delta f_Y,$$

(2-49)

where $G = \mathcal{F}\{g\}$.

The Wigner distribution as defined above has many interesting properties. For example, as can easily be proved from [Eq.\(2-48\)](#), the projection of the Wigner distribution onto the (x, y) plane yields the energy density in the (x, y) plane,

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} Wg(x, y; f_X, f_Y) df_X df_Y = g(x, y)^2.$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W_g(x, y; f_X, f_Y) df_X df_Y = |g(x, y)|^2.$$

(2-50)

Likewise, as can be proved from [Eq.\(2-49\)](#), the projection of Wg onto the (f_X, f_Y) plane yields the energy density in the (f_X, f_Y) plane,

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} Wg(x, y; f_X, f_Y) dx dy = G(f_X, f_Y)^2.$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W_g(x, y; f_X, f_Y) dx dy = |G(f_X, f_Y)|^2.$$

(2-51)

It follows that the volume under $Wg(x, y; f_X, f_Y)$ is the total energy in $g(x, y)$,

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} Wg(x, y; f_X, f_Y) dx dy df_X df_Y = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)^2 dx dy.$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int \int W_g(x, y; f_X, f_Y) dx dy df_X df_Y = \int_{-\infty}^{\infty} \int |g(x, y)|^2 dx dy.$$

(2-52)

These results would seem to imply that the Wigner distribution can be regarded as a measure of the energy density at any point $(x, y; f_X, f_Y)$ in the four-dimensional space-frequency domain. Unfortunately this is not quite true, for as we shall see, there are cases in which W_g can have a negative value at certain points in the space-frequency domain, thus eliminating the possibility that it can be regarded as a true energy density function. But before further discussing the possible negativity of W_g and ways to preserve its energy-density interpretation, we list, without proof, some of its additional properties:

Real-valued Property

$W_g^*(x, y; f_X, f_Y) = W_g(x, y; f_X, f_Y)$. That is, W_g is a real-valued function of its arguments.

Shift Property

$g(x-a, y-b)$ has a Wigner distribution $W_g(x-a, y-b; f_X, f_Y)$. That is, a shift of $g(x, y)$ results in a similar shift of the Wigner distribution in the (x, y) plane.

Multiplication by a Linear Exponential

$\exp[j2\pi(ax+by)] g(x, y)$ has a Wigner distribution $W_g(x, y; f_X-a, f_Y-b)$. That is, multiplying the function $g(x, y)$ by a complex exponential with a phase that depends linearly on x and y , shifts the Wigner distribution in the (f_X, f_Y) plane.

Convolution Property

$h(x, y) * g(x, y)$ has a Wigner distribution

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W_h(x-\xi, y-\eta; f_X, f_Y) W_g(\xi, \eta; f_X, f_Y) d\xi d\eta.$$

$$\int_{-\infty}^{\infty} \int W_h(x-\xi, y-\eta; f_X, f_Y) W_g(\xi, \eta; f_X, f_Y) d\xi d\eta.$$

That is, convolution of two functions h and g results in a Wigner distribution that is a convolution of their respective Wigner distributions with respect to the (x, y) variables.

Multiplication Property

$h(x,y)g(x,y)$ has a Wigner distribution given by

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} Wh(x,y; f_X - \xi, f_Y - \eta) Wg(x,y; \xi, \eta) d\xi d\eta.$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W_h(x, y; f_X - \xi, f_Y - \eta) W_g(x, y; \xi, \eta) d\xi d\eta.$$

That is, multiplication of two functions h and g results in a Wigner distribution that is a convolution of the two respective Wigner distributions with respect to the (f_X, f_Y) variables.

Magnification Property

The Wigner distribution of $1|M|gxM, yM$ is given by

$$W_{gxM, yM; Mf_X, Mf_Y}$$

$$W_s\left(\frac{x}{M}, \frac{y}{M}; Mf_X, Mf_Y\right).$$

In optical terms, the magnification or demagnification of the function g by a factor M results in a Wigner distribution for which the space variables are likewise magnified or demagnified by a factor M , while the frequency variables suffer the opposite demagnification or magnification by the same factor M . In other words, if the dependence on space is magnified, the dependence on frequency is demagnified, and visa versa. In optics, we shall see that the frequency variables represent angles, so when space is magnified or demagnified, the opposite happens to angles.

Fourier Transform Property

As a final property of interest, let $\mathcal{F}\{\cdot\}$ represent a two-dimensional Fourier transform of its argument function, and let $G(f_X, f_Y) = \mathcal{F}g(x, y)$. Then the Wigner distribution of G is given by $W_g(-f_X, -f_Y; x, y)$. For a one-dimensional function $g(x)$, this corresponds to a rotation in Wigner space by 90° in the clockwise direction. In the two-dimensional case, this is a more complicated rotation in the four-dimensional Wigner space.

Finally, we illustrate by presenting the Wigner distributions of some common and important functions. Because the results are so hard to visualize in four dimensions, we restrict our examples to one-dimensional cases.

Linear Exponential Let $g(x) = \exp(j2\pi\alpha x)$, i.e. an infinite-length complex exponential function with a linear phase dependence on x . Then

$$Wg(x; f_X) = \delta(f_X - \alpha).$$

$$W_g(x; f_X) = \delta(f_X - \alpha).$$

That is, Wg is a delta function sheet parallel to, and distance α from, the x -axis, which is consistent with the idea that this function has a constant local spatial frequency that is the same for all x .

Quadratic-Phase Exponential Let $g(x) = \exp(j\pi\beta x^2)$, i.e. an infinite-length complex exponential with a quadratic-phase dependence on x (a so-called “chirp” function). Then

$$Wg(x; f_X) = \delta(f_X - \beta x)$$

$$W_g(x; f_X) = \delta(f_X - \beta x)$$

That is, Wg is again a delta function sheet, but this time the base of the sheet lies on a line that is tilted in the $(x; f_X)$ domain, the tilt corresponding to a slope β . This result is consistent with the idea that the local spatial frequency of a chirp function varies linearly as a function of x .

Gaussian Function Let $g(x) = \exp(-\pi\gamma x^2)$, i.e. a Gaussian function of x . Then

$$Wg(x; f_X) = 2\gamma \exp(-2\pi(\gamma x^2 + f_X^2/\gamma)).$$

$$W_g(x; f_X) = \sqrt{\frac{2}{\gamma}} \exp\left[-2\pi\left(\gamma x^2 + f_X^2/\gamma\right)\right].$$

Rectangle Function Lastly, consider the function $g(x) = \text{rect}(x)$, i.e. a rectangle function. The Wigner distribution is found to be

$$Wg(x; f_X) = 2(1 - 2|x|)\text{rect}(x)\text{sinc}[2(1 - |2x|)f_X].$$

$$W_g(x; f_X) = 2(1 - 2|x|)\text{rect}(x)\text{sinc}[2(1 - |2x|)f_X].$$

f_X may have any value. Note that when $x=0$, the Wigner distribution reduces to

$$Wg(0; f_X) = 2\text{sinc } 2f_X,$$

$$W_g(0; f_X) = 2 \text{ sinc } (2f_X),$$

which has *negative sidelobes*, thus illustrating the earlier claim that the Wigner distribution can not be strictly interpreted as a distribution of energy density in the space-frequency domain.

While the possible negativity of Wg brings into question its interpretation as an energy density function, this is not as fundamental a problem as it may seem at first glance (see [274], p. 79). The reason lies in fact that it is not possible to *measure* the Wigner distribution at an isolated point in $(x; f_X)$ space. The uncertainty principle prevents this. The product of the

normalized second central moments of the energy density $|g(x)|^2$ in the space domain and the energy density $|G(f_X)|^2$ in the frequency domain must always be greater than or equal to $1/(4\pi^2)$ ([37], p. 160). That is, if $|g(x)|^2$ and $|G(f_X)|^2$ have first moments μ_g and μ_G , respectively, and we define

$$\sigma_g^2 \triangleq \int_{-\infty}^{\infty} (x - \mu_g)^2 |g(x)|^2 dx$$

$$\sigma_g^2 \triangleq \frac{\int_{-\infty}^{\infty} (x - \mu_g)^2 |g(x)|^2 dx}{\int_{-\infty}^{\infty} |g(x)|^2 dx}$$

(2-53)

and

$$\sigma_G^2 \triangleq \int_{-\infty}^{\infty} (f_X - \mu_G)^2 |G(f_X)|^2 df_X$$

$$\sigma_G^2 \triangleq \frac{\int_{-\infty}^{\infty} (f_X - \mu_G)^2 |G(f_X)|^2 df_X}{\int_{-\infty}^{\infty} |G(f_X)|^2 df_X},$$

(2-54)

then the square root of their product must satisfy

$$\sigma_g \sigma_G \geq 14\pi.$$

$$\sigma_g \sigma_G \geq \frac{1}{4\pi}.$$

(2-55)

Thus the higher the resolution in the space domain with which we measure energy density, the lower the resolution in the frequency domain for which we can determine energy density. In view of this fundamental uncertainty, the smallest region in the space-frequency domain within which we can measure energy is of the order of $1/(4\pi)$. If we convolve the Wigner distribution with a window that averages its value over such a space-frequency interval, the result will be a positive function that is an approximation to the signal's energy density distribution in the space-frequency domain. Of course this convolution will also blur the projections of the Wigner

distribution onto the x and f_X axes, thus yielding slightly inaccurate energy densities in the two domains individually. So while the Wigner distribution is a more general concept than local spatial frequency, like the local spatial frequency approach, it is not a perfect representation of space-frequency energy occupation. Nonetheless, it does yield important insights into the character of signals.

In closing, we mention without proof that the local frequency distribution can be derived from the Wigner distribution through the relationship ([20], p. 14)

$$f \rightarrow (l)(x \rightarrow) = 12\pi \nabla \phi(x \rightarrow) = 12\pi \int_{-\infty}^{\infty} f \rightarrow Wg(x \rightarrow; f \rightarrow) df \rightarrow \int_{-\infty}^{\infty} Wg(x \rightarrow; f \rightarrow) df \rightarrow$$

$$\vec{f}^{(l)}(\vec{x}) = \frac{1}{2\pi} \nabla \phi(\vec{x}) = \frac{1}{2\pi} \frac{\int_{-\infty}^{\infty} \int \vec{f} W_g(\vec{x}; \vec{f}) d\vec{f}}{\int_{-\infty}^{\infty} \int W_g(\vec{x}; \vec{f}) d\vec{f}}$$

(2-56)

2.3 Linear Systems

For the purposes of discussion here, we seek to define the word *system* in a way sufficiently general to include both the familiar case of electrical networks and the less-familiar case of optical imaging systems. Accordingly, a system is defined to be a mapping of a set of input functions into a set of output functions. For the case of electrical networks, the inputs and outputs are real-valued functions (voltages or currents) of a one-dimensional independent variable (time); for the case of imaging systems, the inputs and outputs can be real-valued functions (intensity) or complex-valued functions (field amplitude) of a two-dimensional independent variable (space). As mentioned previously, the question of whether intensity or field amplitude should be considered the relevant quantity will be treated at a later time.

If attention is restricted to deterministic (nonrandom) systems, then a specified input must map to a unique output. It is not necessary, however, that each output correspond to a unique input, for as we shall see, a variety of input functions can produce *no* output. Thus we restrict attention at the outset to systems characterized by many-to-one mappings.

A convenient representation of a system is a mathematical operator, $\square\{\cdot\}$, which we imagine to operate on input functions to produce output functions. Thus if the function $g_1(x_1, y_1)$ represents the input to a system, and $g_2(x_2, y_2)$ represents the corresponding output, then by the definition of $\square\{\cdot\}$, the two functions are related through

$$g_2(x_2, y_2) = \square g_1(x_1, y_1).$$

$$g_2(x_2, y_2) = \{g_1(x_1, y_1)\}.$$

(2-57)

Without specifying more detailed properties of the operator $\square\{\cdot\}$, it is difficult to state more specific properties of the general system than those expressed by Eq.(2-57). In the material that follows, we shall be concerned primarily, though not exclusively, with a restricted class of systems that are said to be *linear*. The assumption of linearity will be found to yield simple and physically meaningful representations of such systems; it will also allow useful relations between inputs and outputs to be developed.

2.3.1 Linearity and the Superposition Integral

A system is said to be *linear* if the following superposition property is obeyed for all input functions p^P and q^Q and all complex constants a^A and b^B :

$$\square a p(x_1, y_1) + b q(x_1, y_1) = a \square p(x_1, y_1) + b \square q(x_1, y_1).$$

$$\{a p(x_1, y_1) + b q(x_1, y_1)\} = a \{p(x_1, y_1)\} + b \{q(x_1, y_1)\}.$$

(2-58)

As mentioned previously, the great advantage afforded by linearity is the ability to express the response of a system to an arbitrary input in terms of the responses to certain “elementary” functions into which the input has been decomposed. It is most important, then, to find a simple and convenient means of decomposing the input. Such a decomposition is offered by the so-called *sifting property* of the δ function (cf. [Section 1 of Appendix A](#)), which states that

$$g_1(x_1, y_1) = \int_{-\infty}^{\infty} g_1(\xi, \eta) \delta(x_1 - \xi, y_1 - \eta) d\xi d\eta.$$

$$g_1(x_1, y_1) = \int_{-\infty}^{\infty} \int g_1(\xi, \eta) \delta(x_1 - \xi, y_1 - \eta) d\xi d\eta.$$

(2-59)

This equation may be regarded as expressing g_1 as a linear combination of weighted and displaced δ functions; the elementary functions of the decomposition are, of course, just these δ functions.

To find the response of the system to the input g_1 , substitute (2-59) in (2-57):

$$g_2(x_2, y_2) = \int_{-\infty}^{\infty} g_1(\xi, \eta) \delta(x_1 - \xi, y_1 - \eta) d\xi d\eta.$$

$$g_2(x_2, y_2) = \left\{ \int_{-\infty}^{\infty} \int g_1(\xi, \eta) \delta(x_1 - \xi, y_1 - \eta) d\xi d\eta \right\}.$$

(2-60)

Now, regarding the number $g_1(\xi, \eta) \delta(x_1 - \xi, y_1 - \eta)$ as simply a weighting factor applied to the elementary function $\delta(x_1 - \xi, y_1 - \eta)$, the linearity property (2-58) is invoked to allow \square to operate on the individual elementary functions; thus the operator \square is brought within the integral, yielding

$$g_2(x_2, y_2) = \int_{-\infty}^{\infty} g_1(\xi, \eta) \delta(x_1 - \xi, y_1 - \eta) d\xi d\eta.$$

$$g_2(x_2, y_2) = \int_{-\infty}^{\infty} \int g_1(\xi, \eta) \{\delta(x_1 - \xi, y_1 - \eta)\} d\xi d\eta.$$

(2-61)

As a final step we let the symbol $h(x_2, y_2; \xi, \eta)$ denote the response of the system at point (x_2, y_2) of the output space to a δ function input at coordinates $(x_1 = \xi, y_1 = \eta)$ of the input space; that is,

$$h(x_2, y_2; \xi, \eta) = \delta(x_1 - \xi, y_1 - \eta).$$

$$h(x_2, y_2; \xi, \eta) = \{\delta(x_1 - \xi, y_1 - \eta)\}.$$

(2-62)

The function h is called the *impulse response* (or in optics, the *point-spread function*) of the system. The system input and output can now be related by the simple equation

$$g_2(x_2, y_2) = \int_{-\infty}^{\infty} \int g_1(\xi, \eta) h(x_2, y_2; \xi, \eta) d\xi d\eta.$$

$$g_2(x_2, y_2) = \int_{-\infty}^{\infty} \int g_1(\xi, \eta) h(x_2, y_2; \xi, \eta) d\xi d\eta. \quad (2-63)$$

Note that if the dimensions of g_1 and g_2 are the same, while dx and dy each have the dimensions of length, then the dimensions of the impulse response h must be $1/\text{length}^2$.

This fundamental expression, known as the *superposition integral*, demonstrates the very important fact that a linear system is completely characterized by its responses to unit impulses. To completely specify the output, the responses must in general be known for impulses located at all possible points in the input plane. For the case of a linear *imaging* system, this result has the interesting physical interpretation that the effects of imaging elements (lenses, stops, etc.) can be fully described by specifying the (possibly complex-valued) images of *point sources* located throughout the object field.

2.3.2 Invariant Linear Systems: Transfer Functions

Having examined the input-output relations for a general linear system, we turn now to an important subclass of linear systems, namely *invariant* linear systems. An electrical network is said to be *time-invariant* if its impulse response $h(t; \tau)$ (that is, its response at time t to a unit impulse excitation applied at time τ) depends only on the time difference $(t - \tau)$. Electrical networks composed of fixed resistors, capacitors, and inductors are time-invariant since their characteristics do not change with time.

In a similar fashion, a linear imaging system is *space-invariant* (or equivalently, *isoplanatic*) if its impulse response $h(x_2, y_2; \xi, \eta)$ depends only on the distances $(x_2 - \xi)$ and $(y_2 - \eta)$ (i.e. the x and y distances between the excitation point and the response point). For such a system we can, of course, write

$$h(x_2, y_2; \xi, \eta) = h(x_2 - \xi, y_2 - \eta).$$

$$h(x_2, y_2; \xi, \eta) = h(x_2 - \xi, y_2 - \eta).$$

(2-64)

Thus an imaging system is space-invariant if the image of a point source object changes only in location, not in functional form, as the point source explores the object field. In practice, imaging systems are seldom isoplanatic over their entire object field, but it is usually possible to divide that field into small regions (*isoplanatic patches*), within which the system is approximately invariant. To completely describe the imaging system, the impulse response appropriate for each isoplanatic patch should be specified; but if the particular portion of the object field of interest is sufficiently

small, it often suffices to consider only the isoplanatic patch on the optical axis of the system. Note that for an invariant system the superposition integral (2-63) takes on the particularly simple form

$$g_2(x_2, y_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g_1(\xi, \eta) h(x_2 - \xi, y_2 - \eta) d\xi d\eta$$

$$g_2(x_2, y_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g_1(\xi, \eta) h(x_2 - \xi, y_2 - \eta) d\xi d\eta$$

(2-65)

which we recognize as a two-dimensional *convolution* of the object function with the impulse response of the system. In the future it will be convenient to have a short-hand notation for a convolution relation such as (2-65), and accordingly this equation is written symbolically as

$$g_2 = g_1 * h$$

$$g_2 = g_1 * h$$

where a $*$ symbol between any two functions indicates that those functions are to be convolved.

The class of invariant linear systems has associated with it a far more detailed mathematical structure than the more general class of all linear systems, and it is precisely because of this structure that invariant systems are so easily dealt with. The simplicity of invariant systems begins to be evident when we note that the convolution relation (2-65) takes a particularly simple form after Fourier transformation. Specifically, transforming both sides of (2-65) and invoking the convolution theorem, the spectra $G_2(f_X, f_Y)$ and $G_1(f_X, f_Y)$ of the system output and input are seen to be related by the simple equation

$$G_2(f_X, f_Y) = H(f_X, f_Y) G_1(f_X, f_Y),$$

$$G_2(f_X, f_Y) = H(f_X, f_Y) G_1(f_X, f_Y),$$

(2-66)

where H is the Fourier transform of the impulse response

$$H(f_X, f_Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(\xi, \eta) \exp[-j2\pi(f_X \xi + f_Y \eta)] d\xi d\eta.$$

$$H(f_X, f_Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(\xi, \eta) \exp[-j2\pi(f_X \xi + f_Y \eta)] d\xi d\eta.$$

(2-67)

The function H , called the *transfer function* of the system, indicates the effects of the system in the “frequency domain.” Note that the relatively tedious convolution operation of (2-65) required to find the system output is replaced in (2-66) by the often more simple sequence of Fourier transformation, multiplication of transforms, and inverse Fourier transformation.

From another point of view, we may regard the relations (2-66) and (2-67) as indicating that, for a linear invariant system, the input can be decomposed into elementary functions that are more

convenient than the δ functions of [Eq.\(2-59\)](#). These alternative elementary functions are, of course, the complex-exponential functions of the Fourier integral representation. By transforming g_1 ⁸ we are simply decomposing the input into complex-exponential functions of various spatial frequencies (f_X, f_Y) . Multiplication of the input spectrum G_1 by the transfer function H then takes into account the effects of the system on each elementary function. Note that these effects are limited to an amplitude change and a phase shift, as evidenced by the fact that we simply multiply the input spectrum by a complex number $H(f_X, f_Y)$ at each (f_X, f_Y) . Inverse transformation of the output spectrum G_2 synthesizes the output g_2 by adding up the modified elementary functions.

The mathematical term *eigenfunction* is used for a function that retains its original form (up to a multiplicative complex constant) after passage through a system. Thus we see that *complex-exponential functions are eigenfunctions of linear, invariant systems*. The weighting applied by the system to an eigenfunction input is called the *eigenvalue* corresponding to that input. Hence the transfer function describes a continuum of eigenvalues of the system⁸.

Finally, it should be strongly emphasized that the simplifications afforded by transfer-function theory are only applicable for *invariant* linear systems. For applications of Fourier theory in the analysis of time-varying electrical networks, the reader may consult [\[186\]](#); applications of Fourier analysis to space-variant imaging systems can be found in [\[232\]](#).

2.4 Two-Dimensional Sampling Theory

It is often convenient, both for data processing and for mathematical analysis purposes, to represent a function $g(x,y)$ by an array of its sampled values taken on a discrete set of points in the (x,y) plane. Intuitively, it is clear that if these samples are taken sufficiently close to each other, the sampled data are an accurate representation of the original function, in the sense that g_s can be reconstructed with considerable accuracy by simple interpolation. It is a less obvious fact that for a particular class of functions (known as *bandlimited* functions) the reconstruction can be accomplished *exactly*, provided only that the interval between samples is not greater than a certain limit. This result was originally pointed out by [Whittaker \[368\]](#) and was later popularized by [Shannon \[314\]](#) in his studies of information theory.

The sampling theorem applies to the class of bandlimited functions, by which we mean functions with Fourier transforms that are nonzero over only a finite region \mathcal{R} of the frequency space. We consider first a form of this theorem that is directly analogous to the one-dimensional theorem used by Shannon. Later we very briefly indicate improvements of the theorem that can be made in some two-dimensional cases.

2.4.1 The Whittaker-Shannon Sampling Theorem

To derive what is perhaps the simplest version of the sampling theorem, we consider a rectangular lattice of samples of the function g_s , as defined by

$$g_s(x,y) = \text{comb}(x/X)\text{comb}(y/Y)g(x,y).$$

$$g_s(x, y) = \text{comb}\left(\frac{x}{X}\right)\text{comb}\left(\frac{y}{Y}\right)g(x, y).$$

(2-68)

The sampled function g_s thus consists of an array of δ functions, spaced at intervals of width X in the x direction and width Y in the y direction, as illustrated in [Fig. 2.5](#). The area under each δ function is proportional to the value of the function g at that particular point in the rectangular sampling lattice. As implied by the convolution theorem, the spectrum G_s of g_s can be found by convolving the transform of $\text{comb}(x/X)\text{comb}(y/Y)$ $\text{comb}(x/X)\text{comb}(y/Y)$ with the transform of g , or

$$G_s(f_X, f_Y) = \mathcal{F}[\text{comb}(x/X)\text{comb}(y/Y)] * G(f_X, f_Y)$$

$$G_s(f_X, f_Y) = \mathcal{F}\left\{\text{comb}\left(\frac{x}{X}\right)\text{comb}\left(\frac{y}{Y}\right)\right\} * G(f_X, f_Y)$$

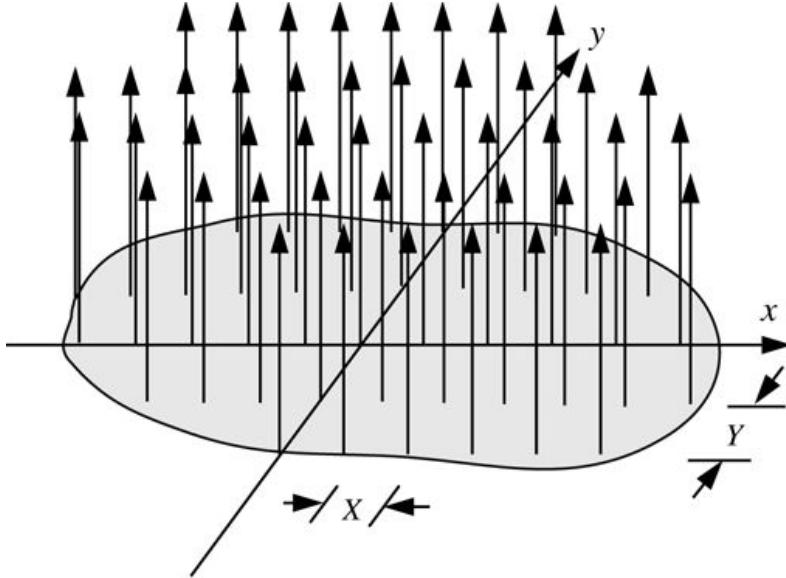


Figure 2.5

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 2.5 The sampled function.

where the $\ast \ast$ again indicates that a two-dimensional convolution is to be performed. Now using [Table 2.1](#) we have

$$\mathcal{F} \text{comb}_x X \text{comb}_y Y = XY \text{comb}(Xf_X) \text{comb}(Yf_Y)$$

$$\mathcal{F} \left\{ \text{comb} \left(\frac{x}{X} \right) \text{comb} \left(\frac{y}{Y} \right) \right\} = XY \text{comb}(Xf_X) \text{comb}(Yf_Y)$$

while from the results of [Prob. 2-1\(b\)](#),

$$XY \text{comb}(Xf_X) \text{comb}(Yf_Y) = \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \delta(f_X - nX, f_Y - mY).$$

$$XY \text{comb}(Xf_X) \text{comb}(Yf_Y) = \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \delta \left(f_X - \frac{n}{X}, f_Y - \frac{m}{Y} \right).$$

It follows that

$$G_s(f_X, f_Y) = \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} G(f_X - nX, f_Y - mY).$$

$$G_s(f_X, f_Y) = \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} G \left(f_X - \frac{n}{X}, f_Y - \frac{m}{Y} \right).$$

(2-69)

Evidently the spectrum of g_s can be found simply by erecting the spectrum of g about each point $(n/X, m/Y)$ in the (f_X, f_Y) plane as shown in [Fig. 2.6](#).

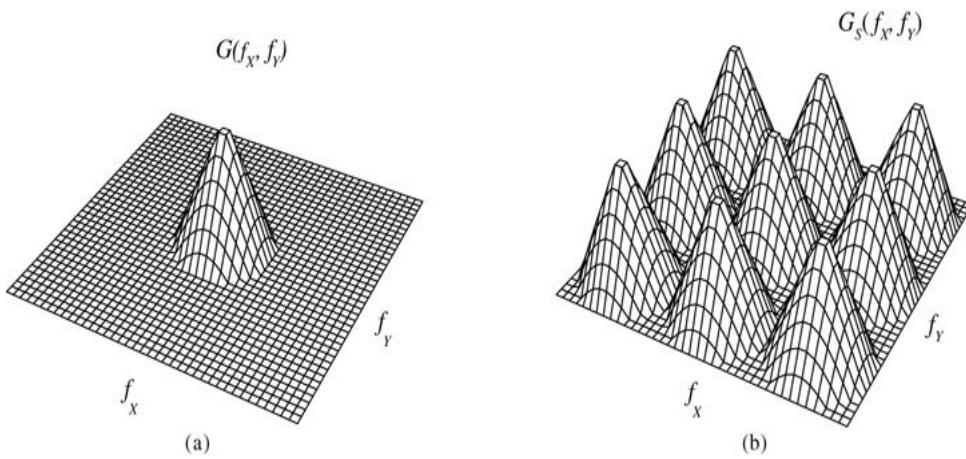


Figure 2.6

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 2.6 Spectra of (a) the original function and (b) the sampled data (only three periods are shown in each direction for this infinitely periodic function).

The two graphs show two 3 dimensional projections on a square horizontal plane whose adjoining axes are f_x and f_y . The projection in illustration a is a single cone at the center representing $G(f_x, f_y)$ and the projection in illustration b has three parallel rows of 3 cones each representing $G_s(f_x, f_y)$.

Since the function g is assumed to be bandlimited, its spectrum G is nonzero over only a finite region \mathcal{R} of the frequency space. As implied by (2-69), the region over which the spectrum of the *sampled* function is nonzero can be found by constructing the region \mathcal{R}_s about each point $(n/X, m/Y)$ in the frequency plane. Now it becomes clear that if X and Y are sufficiently small (i.e. the samples are sufficiently close together), then the separations $1/X$ and $1/Y$ of the various spectral islands will be great enough to ensure that the adjacent regions do not overlap (see Fig. 2.6). Thus the recovery of the original spectrum G from G_s can be accomplished *exactly* by passing the sampled function g_s through a linear invariant filter that transmits the term $(n=0, m=0)$ of (2-69) without distortion, while perfectly excluding all other terms. Thus, at the output of this filter we find an exact replica of the original data $g(x, y)$.

As stated in the above discussion, to successfully recover the original data it is necessary to take samples close enough together to enable separation of the various spectral regions of G_s . To determine the maximum allowable separation between samples, let B_X and B_Y (i.e. the widths of the intervals $(-B_X/2, B_X/2)$ and $(-B_Y/2, B_Y/2)$) represent the widths in the f_x and f_y directions, respectively, of the *smallest* rectangle⁹ that completely encloses the region \mathcal{R} . Since the various terms in the spectrum (2-69) of the sampled data are separated by distances $1/X$ and $1/Y$ in the f_x and f_y directions, respectively, separation of the spectral regions is ensured if

$$X \leq 1/B_X \text{ and } Y \leq 1/B_Y.$$

$$X \leq \frac{1}{B_X} \text{ and } Y \leq \frac{1}{B_Y}.$$

(2-70)

The *maximum* spacings of the sampling lattice for exact recovery of the original function are thus $(BX)^{-1}$ and $(BY)^{-1}$. When the function is sampled with these separations we say that it is being sampled at the *Nyquist rate*.

Having determined the maximum allowable distances between samples, it remains to specify the exact transfer function of the filter through which the data should be passed. In many cases there is considerable latitude of choice here, since for many possible shapes of the region \mathcal{R} there are a multitude of transfer functions that will pass the $(n=0, m=0)$ term of G_s and exclude all other terms. For our purposes, however, it suffices to note that if the relations (2-70) are satisfied, there is one transfer function that will always yield the desired result regardless of the shape of \mathcal{R} , namely

$$H(f_X, f_Y) = \text{rect}(f_X/B_X) \text{rect}(f_Y/B_Y).$$

$$H(f_X, f_Y) = \text{rect}\left(\frac{f_X}{B_X}\right) \text{rect}\left(\frac{f_Y}{B_Y}\right).$$

(2-71)

The exact recovery of G from G_s is seen by noting that the spectrum of the output of such a filter is

$$G_s(f_X, f_Y) \text{rect}(f_X/B_X) \text{rect}(f_Y/B_Y) = G(f_X, f_Y).$$

$$G_s(f_X, f_Y) \text{rect}\left(\frac{f_X}{B_X}\right) \text{rect}\left(\frac{f_Y}{B_Y}\right) = G(f_X, f_Y).$$

The equivalent identity in the space domain is

$$\begin{aligned} & \text{comb}(x/X) \text{comb}(y/Y) * h(x, y) = g(x, y) \\ & \left[\text{comb}\left(\frac{x}{X}\right) \text{comb}\left(\frac{y}{Y}\right) g(x, y) \right] * h(x, y) = g(x, y) \end{aligned}$$

(2-72)

where h is the impulse response of the filter,

$$h(x, y) = \mathcal{F}^{-1}[\text{rect}(f_X/B_X) \text{rect}(f_Y/B_Y)] = B_X B_Y \text{sinc}(B_X x) \text{sinc}(B_Y y).$$

$$h(x, y) = \mathcal{F}^{-1}\left\{\text{rect}\left(\frac{f_X}{B_X}\right) \text{rect}\left(\frac{f_Y}{B_Y}\right)\right\} = B_X B_Y \text{sinc}(B_X x) \text{sinc}(B_Y y).$$

Noting that

$$\text{comb}x\text{X}\text{comby}Yg(x,y)=XY\sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty}g(nX,mY)\delta(x-nX,y-mY),$$

$$\text{comb}\left(\frac{x}{X}\right)\text{comb}\left(\frac{y}{Y}\right)g(x, y) = XY \sum_{n = -\infty}^{\infty} \sum_{m = -\infty}^{\infty} g(nX, mY) \delta(x - nX, y - mY),$$

[Eq. \(2-72\)](#) becomes

$$g(x,y)=B_XB_YXY\sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty}g(nX,mY)\text{sinc}[BX(x-nX)]\text{sinc}[BY(y-mY)].$$

$$g(x, y) = B_X B_Y X Y \sum_{n = -\infty}^{\infty} \sum_{m = -\infty}^{\infty} g(nX, mY) \text{sinc}[B_X(x - nX)] \text{sinc}[B_Y(y - mY)].$$

Finally, when the sampling intervals X and Y are taken to have their maximum allowable values, the identity becomes

$$g(x,y)=\sum_{n=-\infty}^{\infty}\sum_{m=-\infty}^{\infty}g(nBX,mBY)\times\text{sinc}[BX(x-nBX)]\text{sinc}[BY(y-mBY)].$$

$$g(x, y) = \sum_{n = -\infty}^{\infty} \sum_{m = -\infty}^{\infty} g\left(\frac{n}{B_X}, \frac{m}{B_Y}\right) \times \text{sinc}\left[B_X\left(x - \frac{n}{B_X}\right)\right] \text{sinc}\left[B_Y\left(y - \frac{m}{B_Y}\right)\right].$$

(2-73)

Equation (2-73) represents a fundamental result which we shall refer to as the *Whittaker-Shannon sampling theorem*. It implies that exact recovery of a bandlimited function can be achieved from an appropriately spaced rectangular array of its sampled values; the recovery is accomplished by injecting, at each sampling point, an interpolation function consisting of a product of sinc functions, where each interpolation function is weighted according to the sampled value of g at the corresponding point.

The above result is by no means the only possible sampling theorem. Two rather arbitrary choices were made in the analysis, and alternative choices at these two points will yield alternative sampling theorems. The first arbitrary choice, appearing early in the analysis, was the use of a *rectangular* sampling lattice. The second, somewhat later in the analysis, was the choice of the particular filter transfer function (2-71). Alternative theorems derived by making different choices at these two points are no less valid than [Eq. \(2-73\)](#); in fact, in some cases alternative theorems are more “efficient” in the sense that fewer samples per unit area are required to ensure complete recovery. The reader interested in pursuing this extra richness of multidimensional sampling theory is referred to the works of [Bracewell \[36\]](#) and of [Peterson and Middleton \[281\]](#). A more modern treatment of multidimensional sampling theory is found in [Dudgeon and Mersereau \[96\]](#).

One final subtlety should be mentioned when dealing with sampling problems in optics.

Often the function $g(x,y)$ may represent a complex-valued optical field, while we are interested in reconstructing the *intensity* of that field, $|g(x,y)|^2$. The bandwidth of the intensity is twice the bandwidth of the field. To reconstruct the intensity from samples of the field, two different approaches are possible. First, we can take samples of the field at the Nyquist density appropriate for its bandwidth, reconstruct the field by interpolating with appropriate sinc functions, and then square the interpolated field, yielding the distribution of intensity.

Alternatively, we can sample at twice the density required for the field, take the squared magnitude of these samples, and interpolate the resulting samples with sinc functions that are half the width of the sinc functions used in the previous approach. Both approaches are correct, but the former approach is more efficient in the sense that the number of samples to be calculated and interpolated is smaller.

2.4.2 Oversampling, Undersampling and Aliasing

We have seen that, for a function with a bandlimited spectrum, there exists a critical sampling density, the Nyquist density, which is the smallest sampling density that allows complete recovery of the signal. Very often, to be sure that we are isolating the spectral island centered at frequency $(0, 0)$, as a matter of precaution it is wise to sample a bit more densely than the Nyquist rate, perhaps even twice as many samples as would strictly be required. Sampling more densely than the Nyquist rate is referred to as *oversampling*. By oversampling, we are creating more empty space between the spectral islands of the sampled data, not only ensuring that the spectral islands do not overlap, but also making it easier for a proper filter to isolate the spectral island of interest.

If the sampling density is less than the Nyquist density, then we say that we have *undersampled* the data. When the data is undersampled, the spectral islands will overlap, making perfect recovery of the signal impossible. Undersampling occurs inevitably when the signal we are sampling is not strictly bandlimited, but rather has a spectrum that tapers off gradually. If the spectrum continues with finite values and only isolated zeros for all frequencies, then there is no sampling density that will allow exact recovery, but rather there will always be some error in the reconstruction, no matter how dense the samples. The goal in sampling is then to sample densely enough that the overlap between spectral islands is sufficiently small to ensure that the signal reconstruction error will be tolerably small.

When the data is undersampled, the spectral islands will partially overlap. This leads to a phenomenon called *aliasing*, as illustrated in [Fig. 2.7](#). The details are found in the caption.

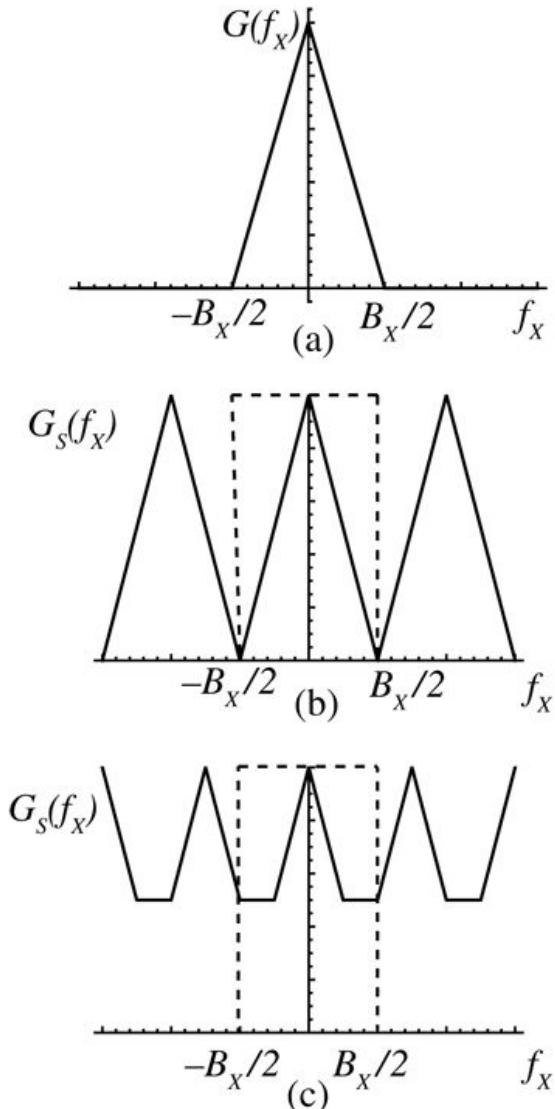


Figure 2.7
Goodman, Introduction to Fourier Optics, 4e,
 © 2017 W. H. Freeman and Company

Figure 2.7 Illustration of aliasing. (a) A central slice through the two-dimensional spectrum shown in Fig. 2.6 (a). (b) A central slice through the spectral islands shown in Fig. 2.6 (b), representing the spectrum that results from sampling at the Nyquist rate. (c) The spectrum that results from sampling at 3/4 the Nyquist rate, showing the result of overlap of the spectral islands. In both (b) and (c), the dotted rectangles represent the transfer function of the interpolation filter. Clearly the original spectrum is not recovered in (c).

Graph a plots f_x along the horizontal axis and $G(f_x)$ along the vertical axis. Points $B_x/2$ and $-B_x/2$ are marked on the horizontal axis. The graph includes two sloping straight lines originating at a point on the vertical axis and mirroring each other across the vertical axis, one sloping to $B_x/2$ and the other to $-B_x/2$. The graph continues in a straight path along the horizontal axis, rightward from $B_x/2$ and leftward from $-B_x/2$. Graph b plots $G_s(f_x)$ along the vertical axis. Points $B_x/2$ and $-B_x/2$ are marked on the horizontal axis. The graph includes two sloping straight lines originating at a point on the vertical axis and mirroring each other across the vertical axis, one sloping to $B_x/2$ and the other to $-B_x/2$.

minus B subscript X/2. The triangle thus formed is replicated once to the right and once to the left to form a continuous pattern of three triangles. Two perpendicular dotted lines of height equal to that of the triangles are dropped at B subscript X/ 2 and minus B subscript X/2. A horizontal dotted line connects the top extremes of the perpendiculars. Graph c plots f subscript x along the horizontal axis and G subscript (f subscript X) along the vertical axis. Points B subscript X/ 2 and minus B subscript X/2 are marked on the horizontal axis. The graph includes two sloping straight lines originating at a point on the vertical axis and mirroring each other across the vertical axis. Both extend up to a point approximately midway between their start point and the horizontal axis. Both lines then extend horizontally, the left one extends leftward and the right one extends rightward, up to points through which dotted perpendicular lines to the horizontal axis are dropped to B subscript X/ 2 and minus B subscript X/2. The top extremes of the perpendiculars are connected by a horizontal dotted line passing through the point where the sloping lines begin on the vertical axis. On the right side the graph extends in an upward slope and then in a downward slope followed by a horizontal segment and an upward sloping line. This pattern is reflected onto the other side of the vertical axis. Thus we have a repetition of a cuplike pattern, two on each side of the vertical axis, forming a symmetrical graph.

Aliasing is especially important when the spectrum of the data signal is not bandlimited, as is the case for every signal that has finite occupancy in the (x, y) plane. Some degree of overlap of the spectral islands is then inevitable, but it can be minimized if the sampling density is chosen to be high enough. When estimating the required sampling rate for a non-bandlimited signal, it should be kept in mind that the spectral islands in the frequency domain add on a complex amplitude basis, rather than an energy basis, and therefore can interfere.

2.4.3 Space-Bandwidth Product

It is possible to show that no function that is bandlimited can be perfectly space-limited as well. That is, if the spectrum G^G of a function g^g is nonzero over only a limited region \mathcal{R}^R in the (f_X, f_Y) plane, then it is not possible for g^g to be nonzero over only a finite region in the (x, y) plane simultaneously¹⁰. Nonetheless, in practice most functions do eventually fall to very small values, and therefore from a practical point of view it is usually possible to say that g^g has *significant* values only in some finite region. Exceptions are functions that do not have Fourier transforms in the usual sense, and have to be dealt with in terms of generalized Fourier transforms (e.g. $g(x,y)=1, g(x,y)=\cos[2\pi(f_X x + f_Y y)]$, etc.).

If $g(x,y)^g$ is bandlimited and indeed has *significant* value over only a finite region of the (x, y) plane, then it is possible to represent g^g with good accuracy by a *finite number* of samples. If g^g is of significant value only in the region $-L_X/2 \leq x < L_X/2, -L_Y/2 \leq y < L_Y/2$ $-L_X/2 \leq x < L_X/2, -L_Y/2 \leq y < L_Y/2$, and if g^g is sampled, in accord with the sampling theorem, on a rectangular lattice with spacings $(B_X)^{-1}, (B_Y)^{-1}$ in the x and y directions, respectively, then the total number of significant samples required to represent $g(x,y)^g$ is seen to be

$$N = L_X L_Y B_X B_Y,$$

$$N = L_X L_Y B_X B_Y,$$

(2-74)

which we call the *space-bandwidth product* of the function g^8 . The space-bandwidth product can be regarded as the number of degrees of freedom of the given function.

The concept of space-bandwidth product is also useful for many functions that are not strictly bandlimited. If the function is approximately space-limited and approximately bandlimited, then a rectangle (size $B_X \times B_Y$) within which most of the spectrum is contained can be defined in the frequency domain, and a rectangle (size $L_X \times L_Y$) within which most of the function is contained can be defined in the space domain. The space-bandwidth product of the function is then approximately given by [Eq.\(2-74\)](#).

The space-bandwidth product of a function is a measure of its complexity. The ability of an optical system to accurately handle inputs and outputs having large space-bandwidth products is a measure of performance, and is directly related to the quality of the system.

2.5 The Discrete Fourier Transform

The reader is by now quite familiar with the Fourier transform in continuous form. In this section we turn attention to the issue of computation of the Fourier transform from sampled data, the so-called *discrete Fourier transform*.

As a starting point, the continuous Fourier transform is, of course, given by

$$G(f_X, f_Y) = \int_{-\infty}^{\infty} \int g(x, y) \exp[-j2\pi(f_X x + f_Y y)] dx dy.$$

$$G(f_X, f_Y) = \int_{-\infty}^{\infty} \int g(x, y) \exp[-j2\pi(f_X x + f_Y y)] dx dy. \quad (2-75)$$

Now if the signal $g(x, y)$, which is assumed to have finite lengths (L_X, L_Y) , is sampled with spacing $(\Delta x, \Delta y)$ between samples, we obtain another continuous transform

$$G^{\wedge}(f_X, f_Y) = \sum_{n=0}^{N_X-1} \sum_{m=0}^{N_Y-1} g(n\Delta x, m\Delta y) \exp[-j2\pi(n\Delta x f_X + m\Delta y f_Y)],$$

$$\hat{G}(f_X, f_Y) = \sum_{n=0}^{N_X-1} \sum_{m=0}^{N_Y-1} g(n\Delta x, m\Delta y) \exp[-j2\pi(n\Delta x f_X + m\Delta y f_Y)], \quad (2-76)$$

where $N_X = L_X / \Delta x$ and $N_Y = L_Y / \Delta y$. Note that the spectrum $G^{\wedge}(f_X, f_Y)$ now consists of spectral islands due to the sampling in the space domain, and presumably Δx and Δy should be chosen small enough to minimize aliasing in the frequency domain. If the width of the spectrum is (B_X, B_Y) , then for Nyquist sampling of g^{\wedge} one would choose

$$\Delta x = 1/B_X \Delta y = 1/B_Y.$$

$$\begin{aligned} \Delta x &= 1 / B_X \\ \Delta y &= 1 / B_Y. \end{aligned}$$

$$(2-77)$$

We wish to calculate the spectrum of the sampled function g^{\wedge} , but we cannot find values at all arguments (f_X, f_Y) of \hat{G} , for there would be an infinite number of frequencies for

which the calculation would have to be performed. Rather, we calculate \hat{G} on an $N_X \times N_Y$ discrete array of frequencies spaced by increments $(\Delta f_X, \Delta f_Y)$,

$$G^\wedge(p\Delta f_X, q\Delta f_Y) = \sum_{n=0}^{N_X-1} \sum_{m=0}^{N_Y-1} g(n\Delta x, m\Delta y) \exp[-j2\pi np\Delta x \Delta f_X + mq\Delta y \Delta f_Y],$$

$$\hat{G}(p\Delta f_X, q\Delta f_Y) = \sum_{n=0}^{N_X-1} \sum_{m=0}^{N_Y-1} g(n\Delta x, m\Delta y) \exp[-j2\pi(np\Delta x \Delta f_X + mq\Delta y \Delta f_Y)],$$

(2-78)

where $p=0, 1, \dots, N_X-1$ and $q=0, 1, \dots, N_Y-1$. To be most efficient but to avoid aliasing in the space domain, one would choose

$$\Delta f_X = 1/L_X, \Delta f_Y = 1/L_Y.$$

$$\begin{aligned}\Delta f_X &= 1/L_X \\ \Delta f_Y &= 1/L_Y.\end{aligned}$$

(2-79)

Now note that

$$\Delta f_X \Delta x = 1/L_X B_X = 1/N_X \Delta f_Y \Delta y = 1/L_Y B_Y = 1/N_Y,$$

$$\begin{aligned}\Delta f_X \Delta x &= 1/L_X B_X = 1/N_X \\ \Delta f_Y \Delta y &= 1/L_Y B_Y = 1/N_Y,\end{aligned}$$

(2-80)

where $N = N_X N_Y$ is the space-bandwidth product of the function $g(x, y)$. It follows that

$$G^\wedge(p/L_X, q/L_Y) = \sum_{n=0}^{N_X-1} \sum_{m=0}^{N_Y-1} g(n/B_X, m/B_Y) \exp[-j2\pi np/N_X + mq/N_Y],$$

$$\hat{G}(p/L_X, q/L_Y) = \sum_{n=0}^{N_X-1} \sum_{m=0}^{N_Y-1} g(n/B_X, m/B_Y) \exp\left[-j2\pi\left(\frac{np}{N_X} + \frac{mq}{N_Y}\right)\right],$$

(2-81)

or, in a shorthand notation,

$$G^\sim(p, q) = \sum_{n=0}^{N_X-1} \sum_{m=0}^{N_Y-1} g(n, m) \exp[-j2\pi np/N_X + mq/N_Y].$$

$$\tilde{G}_{p,q} = \sum_{n=0}^{N_X-1} \sum_{m=0}^{N_Y-1} \tilde{g}_{n,m} \exp\left[-j2\pi\left(\frac{np}{N_X} + \frac{mq}{N_Y}\right)\right].$$

(2-82)

Equation 2-82 represents the *discrete Fourier transform* (DFT) of the sequence $\tilde{g}_{n,m}$ and is widely used in Fourier analysis of sampled data. We represent this relationship in shorthand by writing

$$\tilde{G}_{p,q} = \text{FT}\{\tilde{g}_{n,m}\}.$$

$$\tilde{G}_{p,q} = \text{FT}\{\tilde{g}_{n,m}\}.$$

(2-83)

The inverse DFT is given by

$$\tilde{g}_{n,m} = \frac{1}{N_X N_Y} \sum_{p=0}^{N_X-1} \sum_{q=0}^{N_Y-1} \tilde{G}_{p,q} \exp\left[j2\pi\left(\frac{np}{N_X} + \frac{mq}{N_Y}\right)\right],$$

$$\tilde{g}_{n,m} = \frac{1}{N_X N_Y} \sum_{p=0}^{N_X-1} \sum_{q=0}^{N_Y-1} \tilde{G}_{p,q} \exp\left[j2\pi\left(\frac{np}{N_X} + \frac{mq}{N_Y}\right)\right].$$

(2-84)

and is represented symbolically by

$$\text{IFT}\{\tilde{G}_{p,q}\}.$$

$$\text{IFT}^{-1}\{\tilde{G}_{p,q}\}.$$

(2-85)

Before closing this section, several points are worth mentioning. First, the “zero frequency” component of the spectrum is represented by $\tilde{G}_{0,0}$ and appears in the lower left corner of the DFT array. It is a simple matter to circularly shift¹¹ the DFT array so that this coefficient appears in the center. Second, the two-dimensional DFT can be performed as a succession of one-dimensional DFTs, first transforming each of the rows, and then transforming each of the columns of partially transformed data.

Thirdly, note that to directly calculate the DFT by brute force, requires $N_X N_Y$ complex multiplies and adds for each DFT coefficient. It follows that to calculate $N_X N_Y$ DFT coefficients would require $N_X^2 N_Y^2$ complex multiply and adds, or for a square array ($N_X = N_Y = N$), N^4 operations. An alternative approach to direct computation would be to perform one-dimensional transforms across each of the N_Y rows of the two-dimensional array, each such transform requiring N_X^2 operations, followed by one-dimensional transforms down each of the resulting columns, requiring $N_X N_Y^2$ operations for a total of $(N_X + N_Y)N_X N_Y$ operations. For a square array, the total operation count would then be $2N^3$. Fortunately there are algorithms known as *fast Fourier*

transform (FFT) algorithms that use special properties of the complex exponential to calculate $\tilde{G}_{p,q}$ in far fewer operations. In this case, transforming each row of the $\tilde{g}_{n,m}$ array with the FFT algorithm requires the order of $N_X \log_2 N_X$ complex multiplies and adds, and since there are N_Y rows, the row operations require the order $N_X N_Y \log_2 N_X$ operations. The column operations then require the order of $N_X N_Y \log_2 N_Y$ operations. The total operation count for the two-dimensional FFT is therefore the order of $N_X N_Y (\log_2 N_X + \log_2 N_Y)$, or, for a square array, $2N^2 \log_2 N$ operations. This number is to be compared with $2N^3$ operations required for the most efficient brute-force method. There is thus a great advantage to using the FFT algorithm when computing DFTs.

Other fast algorithms exist for computing the DFT, but we will not delve further into this subject here. The reader interested in going further may wish to consult, for example, [41].

2.6 The Projection-Slice Theorem

The *projection* of a two-dimensional function $g(x,y)$ onto a line that passes through the origin and is oriented at angle θ to the x -axis is defined by

$$p_\theta(x') = \int_{-\infty}^{\infty} g(x' \cos \theta - y' \sin \theta, x' \sin \theta + y' \cos \theta) dy'$$

$$p_\theta(x') = \int_{-\infty}^{\infty} g(x' \cos \theta - y' \sin \theta, x' \sin \theta + y' \cos \theta) dy' \quad (2-86)$$

where the projection is onto the $x'x'$ axis, the $y'y'$ axis is normal to the $x'x'$ axis, where

$$x' = x \cos \theta + y \sin \theta \quad y' = -x \sin \theta + y \cos \theta$$

$$\begin{aligned} x' &= x \cos \theta + y \sin \theta \\ y' &= -x \sin \theta + y \cos \theta \end{aligned}$$

(2-87)

and

$$x = x' \cos \theta - y' \sin \theta \quad y = x' \sin \theta + y' \cos \theta.$$

$$\begin{aligned} x &= x' \cos \theta - y' \sin \theta \\ y &= x' \sin \theta + y' \cos \theta. \end{aligned}$$

(2-88)

The projection-slice theorem can now be stated as follows:

The one-dimensional Fourier transform $P_\theta(f)$ of a projection $p_\theta(x')$ is identical to the two-dimensional Fourier transform $G(f_X, f_Y)$ of $g(x,y)$, evaluated along a slice through the origin at angle θ to the f_X axis:

$$P_\theta(f) = G(f \cos \theta, f \sin \theta).$$

$$P_\theta(f) = G(f \cos \theta, f \sin \theta). \quad (2-89)$$

To prove this theorem we start with an expression for $P_\theta(f)$,

$$P_\theta(f) = \int_{-\infty}^{\infty} p_\theta(x') e^{-j2\pi f x'} dx' = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x' \cos \theta - y' \sin \theta, x' \sin \theta + y' \cos \theta) e^{-j2\pi f x'} dx' dy',$$

$$\begin{aligned}
P_\theta(f) &= \int_{-\infty}^{\infty} p_\theta(x') e^{-j2\pi f x'} dx' \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} g(x' \cos \theta - y' \sin \theta, x' \sin \theta + y' \cos \theta) \right) e^{-j2\pi f x'} dx' dy',
\end{aligned}
\tag{2-90}$$

where the integral is to be carried out over the entire infinite (x', y') plane. It is equivalent to integrate over the entire (x, y) plane,

$$P_\theta(f) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) e^{-j2\pi(f \cos \theta + y \sin \theta)} dx dy.$$

$$P_\theta(f) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) e^{-j2\pi(f \cos \theta + y \sin \theta)} dx dy.$$

(2-91)

Thus we see that

$$\begin{aligned}
P_\theta(f) &= G(f \cos \theta, f \sin \theta), \\
P_\theta(f) &= G(f \cos \theta, f \sin \theta),
\end{aligned}
\tag{2-92}$$

where the right-hand side represents a central slice through the two-dimensional spectrum at angle θ to the f_x axis.

For some examples of the use of the projection-slice theorem, see [Probs. 2-16](#) and [2-17](#) and also [Section 5.8.3](#).

In closing this section, it should be mentioned that it is possible to recover a function from a set of projections taken from a multitude of different angles. Such methods are the basis of computerized tomography. For details, see [\[167\]](#).

2.7 Phase Retrieval from Fourier Magnitude

Suppose that in a certain experiment it is possible to measure only the magnitude of the Fourier transform of a function but not the phase associated with that transform. It is natural, then, to inquire as to whether it is possible to determine the Fourier-domain phase from knowledge of Fourier-domain magnitude. If the answer were to be “yes,” then it would be possible to recover the original function from knowledge of only the magnitude of its Fourier transform.

In general, the answer to this question is “no,” but there are important cases in which the answer is “yes.” It has been known for many years that when the original function being Fourier transformed is of limited extent, that is, zero outside a finite region of support, then the Fourier transform of that function has certain mathematical analyticity properties that suggest that phase retrieval may be possible. This question was investigated by [Wolf \[375\]](#) in 1962 for one-dimensional functions, for example functions of time or of one space variable. He found that there are fundamental ambiguities that lead to a multitude of possible phase functions, and therefore that unique phase recovery is in general not possible in one dimension.

A breakthrough in this field occurred in 1978 when Fienup demonstrated [\[111\]](#) that in two (and higher) dimensions, phase retrieval is possible for objects with bounded support. For more details see [\[112\]](#) and [\[113\]](#).

The procedure for finding the desired phase function is generally an iterative one, reinforcing prior knowledge of the spatial bound and the known modulus of the transform. An ambiguity exists if the support of the original function $g(x,y)$ is centrosymmetric, since the function and its conjugate reflected about the origin, $g^*(-x,-y)$, both have the same magnitudes of their Fourier transforms and both have the same region of support [\[152\]](#). In addition, there is no absolute position information recovered about the object, since linear phase tilts in the Fourier transform are not evident in the Fourier modulus information. If the object is real and non-negative, this fact provides further apriori knowledge that can be exploited in the reconstruction of phase.

Various iterative algorithms have been applied to the phase retrieval problem. A common difficulty is that many algorithms stagnate after a certain number of iterations, yielding no further improvements of the phase estimate with more iterations. A particularly successful algorithm developed by Fienup is the so-called “hybrid input-output” algorithm, which is based on the following:

1. Start with a guess $g_1(x,y)$ for the object $g(x,y)$; for example, fill the region of support with random numbers;
2. Fourier transform g_1 and set the magnitude of its Fourier transform G_1 equal to the known modulus information, keeping the computed phase;
3. Inverse Fourier transform the new spectrum, yielding a second estimate g_2 of the object;
4. At the $k+1$ st iteration, create a new input image g_{k+1} according to the prescription

$g_{k+1} = g_k$ where g_k satisfies the constraints
 $g_{k+1} = \begin{cases} g_k & \text{where } g_k \text{ satisfies the constraints} \\ g_k - \beta g_{k+1} & \text{where } g_{k+1} \text{ violates the constraints} \end{cases}$

(2-93)

where β is an empirically determined constant.

5. Continue iterating steps 2 through 4 until the mean-square difference between g_{k-1} and g_k is no longer changing significantly.

A multitude of other algorithms exist, and often best results are obtained if different algorithms are alternated to drive the mean-square difference even lower than one algorithm alone could obtain. For a history of phase-retrieval work in optics, see [113].

Problems - Chapter 2

1. 2-1. Prove the following properties of δ functions:

$$1. \delta(ax, by) = 1|ab|\delta(x, y) = \frac{1}{|ab|}\delta(x, y).$$

$$2. \text{comb}(ax) \text{comb}(by) = 1|ab|\sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \delta(x - \frac{n}{a}, y - \frac{m}{b}).$$

$$\text{comb}(ax) \text{comb}(by) = \frac{1}{|ab|} \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \delta\left(x - \frac{n}{a}, y - \frac{m}{b}\right).$$

2. 2-2. Prove the following Fourier transform relations:

$$1. \mathcal{F}\{\text{rect}(x)\text{rect}(y)\} = \text{sinc}(f_X)\text{sinc}(f_Y).$$

$$2. \mathcal{F}\{\Lambda(x)\Lambda(y)\} = \text{sinc}^2(f_X)\text{sinc}^2(f_Y).$$

Prove the following generalized Fourier transform relations:

$$3. \mathcal{F}\{1\} = \delta(f_X, f_Y).$$

$$4. \mathcal{F}\{\text{sgn}(x)\text{sgn}(y)\} = 1j\pi f_X 1j\pi f_Y \mathcal{F}\{\text{sgn}(x)\text{sgn}(y)\} = \left(\frac{1}{j\pi f_X}\right) \left(\frac{1}{j\pi f_Y}\right).$$

3. 2-3. Prove the following Fourier transform theorems:

$$1. \mathcal{F}\{Fg(x, y)\} = \mathcal{F}\{g(-x, -y)\} \quad \text{at all points of continuity of } g.$$

$$2. \mathcal{F}\{g(x, y)h(x, y)\} = \mathcal{F}\{g(x, y)\} * \mathcal{F}\{h(x, y)\}.$$

$$3. \mathcal{F}\{\nabla^2 g(x, y)\} = -4\pi^2(f_X^2 + f_Y^2)\mathcal{F}\{g(x, y)\}$$

where ∇^2 is the Laplacian operator

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}.$$

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}.$$

4. 2-4. Let the transform operators $\mathcal{F}_A \{ \cdot \}$ and $\mathcal{F}_B \{ \cdot \}$ be defined by

$$\mathcal{F}_A g = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(\xi, \eta) \exp[-j2\pi(f_X \xi + f_Y \eta)] d\xi d\eta$$

$$\mathcal{F}_B g = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(\xi, \eta) \exp[-j2\pi(x\xi + y\eta)] d\xi d\eta$$

1. Find a simple interpretation for

$$\mathcal{F}B\mathcal{F}Ag(x, y).$$

$$\mathcal{F}_B[\mathcal{F}_A[g(x, y)]].$$

2. Interpret the result for $a > b$ $a > b$ and $a < b$ $a < b$.

5. 2-5. The “equivalent area” Δ_{XY} of a function $g(x, y)$ $g(x, y)$ can be defined by

$$\Delta_{XY} = \int_{-\infty}^{\infty} \int g(x, y) dx dy g(0, 0),$$

$$\Delta_{XY} = \frac{\int_{-\infty}^{\infty} \int g(x, y) dx dy}{g(0, 0)},$$

while the “equivalent bandwidth” $\Delta_{f_X f_Y}$ of g g is defined in terms of its transform G G by

$$\Delta_{f_X f_Y} = \int_{-\infty}^{\infty} \int G(f_X, f_Y) df_X df_Y G(0, 0).$$

$$\Delta_{f_X f_Y} = \frac{\int_{-\infty}^{\infty} \int G(f_X, f_Y) df_X df_Y}{G(0, 0)}.$$

Show that $\Delta_{XY} \Delta_{f_X f_Y} = 1$.

6. 2-6. Prove the following Fourier-Bessel transform relations:

1. If $g_R(r) = \delta(r - r_0)$, then

$$\mathcal{B}g_R(r) = 2\pi r_0 J_0(2\pi r_0 \rho).$$

$$\mathcal{B}\{g_R(r)\} = 2\pi r_0 J_0(2\pi r_0 \rho).$$

2. If $g_R(r) = 1$ for $a \leq r \leq 1$ $a \leq r \leq 1$ and zero otherwise, then

$$\mathcal{B}g_R(r) = J_1(2\pi\rho) - a J_1(2\pi a\rho) \rho.$$

$$\mathcal{B}\{g_R(r)\} = \frac{J_1(2\pi\rho) - a J_1(2\pi a\rho)}{\rho}.$$

3. If $\mathcal{B}g_R(r) = G(\rho)$, then

$$\mathcal{B}g_R(ar) = 1 a^2 G(a\rho).$$

$$\mathcal{B}\{g_R(ar)\} = \frac{1}{a^2} G\left(\frac{\rho}{a}\right).$$

$$4. \mathcal{B}\{\exp(-\pi r^2)\} = \exp(-\pi\rho^2).$$

7. 2-7. Let $g(r, \theta)$ be separable in polar coordinates.

1. Show that if $g(r, \theta) = g_R(r)e^{jm\theta}$, then

$$\mathcal{F}g(r, \theta) = (-j)m e^{jm\phi} \mathcal{H}_m[g_R(r)]$$

$$\mathcal{F}\{g(r, \theta)\} = (-j)^m e^{jm\phi} \mathcal{H}_m[g_R(r)]$$

where $\mathcal{H}_m\{\}$ is the Hankel transform of order m ,

$$\mathcal{H}_m[g_R(r)] = 2\pi \int_0^\infty r g_R(r) J_m(2\pi r\rho) dr$$

$$\mathcal{H}_m[g_R(r)] = 2\pi \int_0^\infty r g_R(r) J_m(2\pi r\rho) dr$$

and (ρ, ϕ) are polar coordinates in the frequency space.

(Hint: $\exp(ja\sin x) = \sum_{k=-\infty}^{\infty} J_k(a) \exp(jkx)$)

$$\exp(ja\sin x) = \sum_{k=-\infty}^{\infty} J_k(a) \exp(jkx)$$

2. With the help of part (a), prove the general relation presented in [Eq.\(2-23\)](#) for functions separable in polar coordinates.

8. 2-8. Show that if a function $g(x, y)$ is separable in rectangular coordinates, its Wigner distribution $W_g(x, y; f_X, f_Y)$ is separable into the product of a function depending only on $(x; f_X)$ and a function depending only on $(y; f_Y)$.

9. 2-9. Show that the Wigner distribution of the one-dimensional function $g(x) = \text{sinc}(x)$ is given by

$$W_g(x; f_X) = 2(1 - 2|f_X|) \text{rect}(f_X) \text{sinc}[2(1 - |2f_X|)|x|]$$

$$W_g(x; f_X) = 2(1 - 2|f_X|) \text{rect}(f_X) \text{sinc}[2(1 - |2f_X|)|x|]$$

Hint: Be clever. Don't use brute force.

10. 2-10. Suppose that a sinusoidal input

$$g(x, y) = \cos[2\pi(f_X x + f_Y y)]$$

$$g(x, y) = \cos[2\pi(f_X x + f_Y y)]$$

is applied to a linear system. Under what (sufficient) conditions is the output a real sinusoidal function of the same spatial frequency as the input? Express the amplitude and phase of that output in terms of an appropriate characteristic of the system.

11. 2-11. Show that the zero-order Bessel function $J_0(2\pi\rho_o r)$ is an eigenfunction of any invariant linear system with a circularly symmetric impulse response. What is the corresponding eigenvalue?

12. 2-12. The Fourier transform operator may be regarded as a mapping of functions into their transforms and therefore satisfies the definition of a system as presented in this chapter.

1. Is this system *linear*?

2. Can you specify a *transfer function* for this system? If yes, what is it? If no, why not?

13. 2-13. The expression

$$p(x,y) = g(x,y) * \text{comb}xX\text{comb}yY$$

$$p(x, y) = g(x, y) * \left[\text{comb}\left(\frac{x}{X}\right) \text{comb}\left(\frac{y}{Y}\right) \right]$$

defines a periodic function, with period X in the x direction and period Y in the y direction.

1. Show that the Fourier transform of p^P can be written

$$P(f_X, f_Y) = \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} G\left(\frac{n}{X}, \frac{m}{Y}\right) \delta(f_X - \frac{n}{X}, f_Y - \frac{m}{Y})$$

where G is the Fourier transform of g .

2. Sketch the function $p(x,y)$ when

$$g(x,y) = \text{rect}\left(2\frac{x}{X}\right) \text{rect}\left(2\frac{y}{Y}\right)$$

and find the corresponding Fourier transform $P(f_X, f_Y)$.

14. 2-14. Show that a function with no nonzero spectral components outside a circle of radius $B/2$ in the frequency plane obeys the following sampling theorem:

$$g(x,y) = \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} g\left(\frac{n}{B}, \frac{m}{B}\right) \frac{\pi}{4} \left\{ 2 \frac{J_1\left[\pi B \sqrt{\left(x - \frac{n}{B}\right)^2 + \left(y - \frac{m}{B}\right)^2}\right]}{\pi B \sqrt{\left(x - \frac{n}{B}\right)^2 + \left(y - \frac{m}{B}\right)^2}} \right\}$$

15. 2-15. The input to a certain imaging system is an *object* complex field distribution $U_o(x,y)$ of unlimited spatial frequency content, while the output of the system is an *image*

field distribution $U_i(x, y)$. The imaging system can be assumed to act as a linear, invariant lowpass filter with a transfer function that is identically zero outside the region $|f_X| \leq B_X/2$, $|f_Y| \leq B_Y/2$ in the frequency domain. Show that there exists an “equivalent” object $U_o'(x, y)$ consisting of a rectangular array of point sources that produces exactly the same image U_i as does the true object U_o , and that the field distribution across the equivalent object can be written

$$U'_o(x, y) = \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \left[\int_{-\infty}^{\infty} \int U_o(\xi, \eta) \text{sinc}(n - BX\xi) \text{sinc}(m - BY\eta) d\xi d\eta \right] \times \delta(x - \frac{n}{B_X}, y - \frac{m}{B_Y})$$

$$U'_o(x, y) = \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \left[\int_{-\infty}^{\infty} \int U_o(\xi, \eta) \text{sinc}(n - BX\xi) \text{sinc}(m - BY\eta) d\xi d\eta \right] \times \delta\left(x - \frac{n}{B_X}, y - \frac{m}{B_Y}\right)$$

16. 2-16. Consider a circularly symmetric function $g(r)$. Suppose that a projection onto a line through the origin at any angle yields a projection

$$p_\theta(x') = 2 \text{sinc}(2x').$$

$$p_\theta(x') = 2 \text{sinc}(2x').$$

Find the radial profile $g(r)$ of the two-dimensional function.

17. 2-17. Show that any function $g(x, y)$ that is separable into $g_X(x)g_Y(y)$ can be recovered from only two different projections.

18. 2-18. Consider an aperture amplitude transmittance $a(x, y)$ defined in the (x, y) domain to be unity within the boundary $y = \pm g(x)$ and zero elsewhere, where $g(x)$ is a real-valued non-negative function symmetric about the y -axis.

1. Find an expression for the projection of this aperture function onto the x axis.

2. Show that the Fourier transform $A(f_X, f_Y)$ of this function, evaluated along the f_X axis, is given by

$$A(f_X, 0) = 2 \int_{-\infty}^{\infty} g(x) \exp(-j2\pi f_X x) dx.$$

$$A(f_X, 0) = 2 \int_{-\infty}^{\infty} g(x) \exp(-j2\pi f_X x) dx.$$

19. 2-19. Consider the one-dimensional function

$$g(x) = J_1(2\pi x)x.$$

$$g(x) = \frac{J_1(2\pi x)}{x}.$$

Using the projection-slice theorem, show that its Fourier transform is given by

$$G(fX) = 21 - fX2 |fX| \leq 10 \text{ otherwise}$$

$$\delta\left(ax,by\right)=\tfrac{1}{\left|ab\right|}\delta\left(x,y\right)$$

3 Foundations of Scalar Diffraction Theory

The phenomenon known as *diffraction* plays a role of utmost importance in the branches of physics and engineering that deal with wave propagation. In this chapter we consider some of the foundations of scalar diffraction theory. While the theory discussed here is sufficiently general to be applied in other fields, such as acoustic-wave and radio-wave propagation, the applications of primary concern will be in the realm of physical optics. To fully understand the properties of optical imaging and data processing systems, it is essential that diffraction and the limitations it imposes on system performance be appreciated. A variety of references to more comprehensive treatments of diffraction theory will be found in the material that follows.

3.1 Historical Introduction

Before beginning a discussion of diffraction, it is first necessary to mention another phenomenon with which diffraction should not be confused, namely *refraction*. Refraction can be defined as the bending of light rays that takes place when they pass through a region in which there is a change of the local speed of propagation of the wave. The most common example occurs when a light wave encounters a sharp boundary between two regions having different refractive indices. The propagation speed in the first medium, having refractive index n_1 , is $v_1 = c/n_1$ $v_1 = c / n_1$, c being the vacuum speed of light. The speed of propagation in the second medium is $v_2 = c/n_2$ $v_2 = c / n_2$.

As shown in [Fig. 3.1](#), the incident light rays are bent at the interface. The angles of incidence and refraction are related by *Snell's law*, which is the foundation of geometrical optics,

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

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

(3-1)

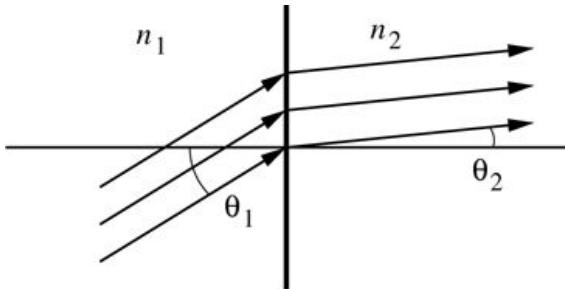


Figure 3.1

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 3.1 Snell's law at a sharp boundary ($n_2 > n_1$).

The illustration shows three upward sloping rays in a medium of refraction index n_1 incident on a vertical interface. The rays pass through the interface and continue on a path less steep than before in a medium of refraction index n_2 . A horizontal line perpendicular at the point of incidence of the lowest ray makes angle θ_1 with the incident ray and θ_2 with the refracted ray.

where in this example, $n_2 > n_1$ $n_2 > n_1$ and therefore $\theta_2 < \theta_1$ $\theta_2 < \theta_1$. Light rays are also bent upon *reflection*, which can occur at a metallic or dielectric interface. The fundamental relation governing this phenomenon is that the angle of reflection is always equal to the angle of incidence.

The term *diffraction* has been defined by [Sommerfeld \[328\]](#) as “any deviation of light rays from rectilinear paths which cannot be interpreted as reflection or refraction.” Diffraction is caused by the confinement of the lateral extent of a wave, and is most appreciable when that confinement is to sizes comparable with a wavelength of the radiation being used. The diffraction phenomenon

should also not be confused with the *penumbra effect*, for which the finite extent of a source causes the light transmitted by a small aperture to spread as it propagates away from that aperture (see [Fig. 3.2](#)). As can be seen in the figure, the penumbra effect does not involve any bending of the light rays.

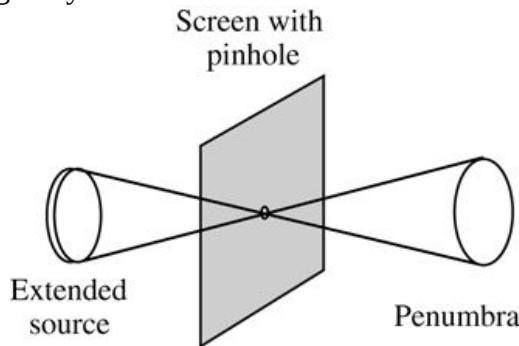


Figure 3.2

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 3.2 The penumbra effect.

The illustration shows a vertical rectangular screen with a pinhole. On the left of the screen is a circular extended source and to the right is a similarly shaped penumbra. Both are equidistant from the pinhole. A light ray from the top of the source passes through the pinhole and reaches the bottom end of the penumbra. Another ray begins at the bottom end of the source and reaches the top end of the penumbra.

There is a fascinating history associated with the discovery and explanation of diffraction effects. The first accurate report and description of such a phenomenon was made by Grimaldi and was published in the year 1665, shortly after his death. The measurements reported were made with an experimental apparatus similar to that shown in [Fig. 3.3](#). An aperture in an opaque screen was illuminated by a light source, chosen small enough to introduce a negligible penumbra effect; the light intensity was observed across a plane some distance behind the screen. The corpuscular theory of light propagation, which was the accepted means of explaining optical phenomena at the time, predicted that the shadow behind the screen should be well defined, with sharp borders. Grimaldi's observations indicated, however, that the transition from light to shadow was gradual rather than abrupt. If the spectral purity of the light source had been better, he might have observed even more striking results, such as the presence of light and dark fringes extending far into the geometrical shadow of the screen. Such effects cannot be explained by a corpuscular theory of light, which requires rectilinear propagation of light rays in the absence of reflection and refraction.

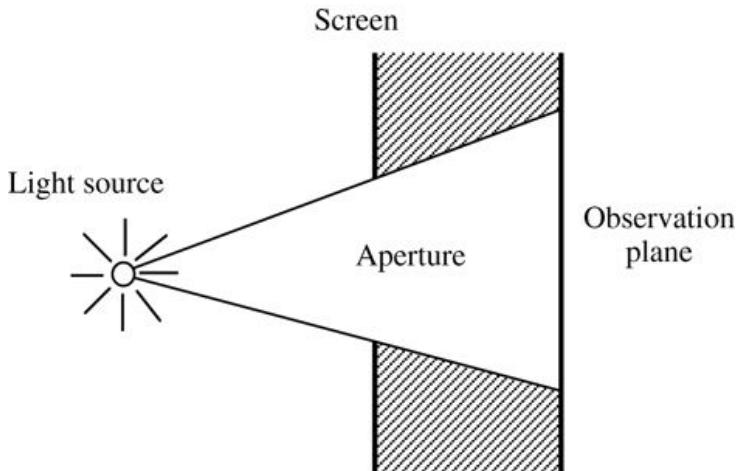


Figure 3.3

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 3.3 Arrangement used for observing diffraction of light.

The initial step in the evolution of a theory that would explain such effects was made by the first proponent of the wave theory of light, Christian Huygens, in the year 1678. Huygens expressed the intuitive conviction that if each point on the wavefront of a disturbance were considered to be a new source of a “secondary” spherical disturbance, then the wavefront at a later instant could be found by constructing the “envelope” of the secondary wavelets, as illustrated in Fig. 3.4.

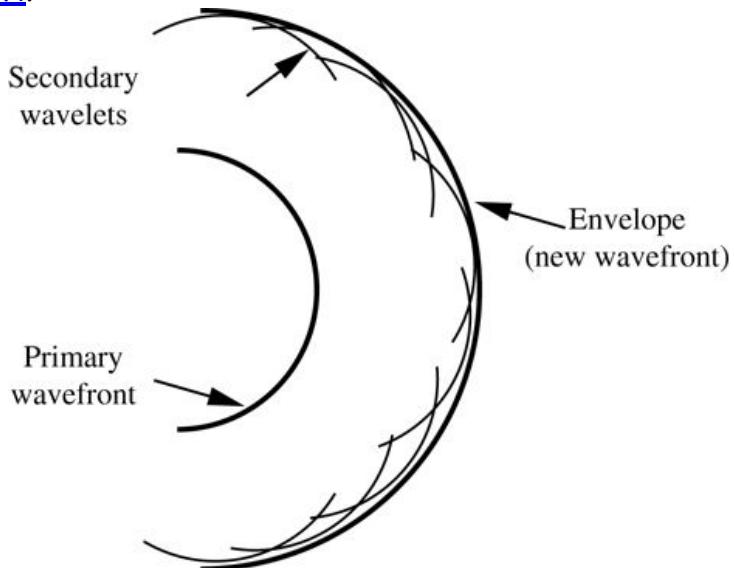


Figure 3.4

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 3.4 Huygens envelope construction.

The illustration shows two concentric semicircles, the inner semicircle is the primary wave front while the outer semicircle is the envelope, the new wave front. The inner side of the envelope is lined with short overlapping arcs opening toward the center; these are secondary wavelets.

Progress on further understanding diffraction was impeded throughout the entire 18th century by the fact that Isaac Newton, a scientist with an enormous reputation for his many contributions to physics in general and to optics in particular, favored the corpuscular theory of light as early as 1704. His followers supported this view adamantly. It was not until 1804 that further significant progress occurred. In that year, Thomas Young, an English physician, strengthened the wave theory of light by introducing the critical concept of *interference*. The idea was a radical one at the time, for it stated that under proper conditions, light could be added to light and produce darkness.

The ideas of Huygens and Young were brought together in 1818 in the famous memoir of Augustin Jean Fresnel. By making some rather arbitrary assumptions about the amplitudes and phases of Huygens' secondary sources, and by allowing the various wavelets to mutually interfere, Fresnel was able to calculate the distribution of light in diffraction patterns with excellent accuracy.

At Fresnel's presentation of his paper to a prize committee of the French Academy of Sciences, his theory was strongly disputed by the great French mathematician S. Poisson, a member of the committee. He demonstrated the absurdity of the theory by showing that it predicted the existence of a bright spot at the center of the shadow of an opaque disk. F. Arago, who chaired the prize committee, performed such an experiment and found the predicted spot. Fresnel won the prize, and since then the effect has been known as "Poisson's spot."

In 1860 Maxwell identified light as an electromagnetic wave, a step of enormous importance. But it was not until 1882 that the ideas of Huygens and Fresnel were put on a firmer mathematical foundation by Gustav Kirchhoff, who succeeded in showing that the amplitudes and phases ascribed to the secondary sources by Fresnel were indeed logical consequences of the wave nature of light. Kirchhoff based his mathematical formulation upon two assumptions about the boundary values of the light incident on the surface of an obstacle placed in the way of propagation of light. These assumptions were later proved to be inconsistent with each other, by Poincaré in 1892 and by Sommerfeld in 1894.¹ As a consequence of these criticisms, Kirchhoff's formulation of the so-called *Huygens-Fresnel* principle must be regarded as a first approximation, although under most conditions it yields results that agree amazingly well with experiment. [Kottler \[209\]](#) attempted to resolve the contradictions by reinterpreting Kirchhoff's boundary value problem as a *saltus* problem, where *saltus* is a Latin word signifying a discontinuity or jump. The Kirchhoff theory was also modified by Sommerfeld, who eliminated one of the aforementioned assumptions concerning the light amplitude at the boundary by making use of the theory of Green's functions. This so-called *Rayleigh-Sommerfeld diffraction theory* will be treated in [Section 3.5](#).

It should be emphasized from the start that the Kirchhoff and Rayleigh-Sommerfeld theories share certain major simplifications and approximations. Most important, light is treated as a *scalar* phenomenon, neglecting the fundamentally vectorial nature of the electromagnetic fields. Such an approach neglects the fact that, at boundaries, the various components of the electric and magnetic fields are coupled through Maxwell's equations and cannot be treated independently. Fortunately, experiments in the microwave region of the spectrum [\[317\]](#) have shown that the scalar theory yields very accurate results in most problems if two conditions are met: (1) the diffracting aperture must be large compared with a wavelength, and (2) the diffracting fields must not be observed too close to the aperture. These conditions will be well satisfied in the problems treated here. For a more complete discussion of the applicability of scalar theory in instrumental optics the reader may consult [\[34\]](#) ([Section 8.4](#)). Nonetheless, there do exist important problems for which the required conditions are *not* satisfied, for example in the theory of diffraction from high-resolution gratings and from extremely small pits on optical recording media, as well in the theory of electromagnetic *metamaterials* (materials with subwavelength periodic or quasiperiodic

structures). Such problems are excluded from consideration here, since the vectorial nature of the field (i.e. its polarization properties) *must* be taken into account if reasonably accurate results are to be obtained. Vectorial generalizations of diffraction theory do exist, the first satisfactory treatment being due to [Kottler \[207\]](#). However, in a majority of cases, solutions of complex vectorial electromagnetic problems are performed numerically.

The first truly rigorous solution of a diffraction problem was given in 1896 by [Sommerfeld \[326\]](#), who treated the two-dimensional case of a plane wave incident on an infinitesimally thin, perfectly conducting half plane. [Kottler \[208\]](#) later compared Sommerfeld's solution with the corresponding results of Kirchhoff's scalar treatment. For a more recent rigorous analysis of vector diffraction by a circular aperture in a thin metal screen, see [\[150\]](#).

Needless to say, a historic introduction to a subject so widely mentioned in the literature can hardly be considered complete. The reader is therefore referred to more comprehensive treatments of diffraction theory, for example [\[13\]](#), [\[35\]](#), and [\[172\]](#).

3.2 From a Vector to a Scalar Theory

The most fundamental beginning for our analysis is Maxwell's equations. In MKS units and in the absence of free charge, the equations are given by

$$\nabla \times \vec{\mathcal{E}} \rightarrow = -\mu \partial \vec{\mathcal{H}} \rightarrow / \partial t \quad \nabla \times \vec{\mathcal{H}} \rightarrow = \epsilon \partial \vec{\mathcal{E}} \rightarrow / \partial t \quad \nabla \cdot \epsilon \vec{\mathcal{E}} \rightarrow = 0 \quad \nabla \cdot \mu \vec{\mathcal{H}} \rightarrow = 0.$$

$$\begin{aligned} \nabla \times \vec{\mathcal{E}} \rightarrow &= -\mu \frac{\partial \vec{\mathcal{H}} \rightarrow}{\partial t} \\ \nabla \times \vec{\mathcal{H}} \rightarrow &= \epsilon \frac{\partial \vec{\mathcal{E}} \rightarrow}{\partial t} \\ \nabla \cdot \epsilon \vec{\mathcal{E}} \rightarrow &= 0 \\ \nabla \cdot \mu \vec{\mathcal{H}} \rightarrow &= 0. \end{aligned}$$

(3-2)

Here $\vec{\mathcal{E}} \rightarrow$ is the electric field, with rectilinear components ($\mathcal{E}_X, \mathcal{E}_Y, \mathcal{E}_Z$) ($\mathcal{E}_X, \mathcal{E}_Y, \mathcal{E}_Z$), and $\vec{\mathcal{H}} \rightarrow$ is the magnetic field, with components ($\mathcal{H}_X, \mathcal{H}_Y, \mathcal{H}_Z$) ($\mathcal{H}_X, \mathcal{H}_Y, \mathcal{H}_Z$). μ and ϵ are the magnetic permeability and electric permittivity, respectively, of the medium in which the wave is propagating. $\vec{\mathcal{E}} \rightarrow$ and $\vec{\mathcal{H}} \rightarrow$ are functions of both position P and time t . The symbols \times and \cdot represent a vector cross product and a vector dot product, respectively,

$\nabla = \frac{\partial}{\partial x} \hat{x} + \frac{\partial}{\partial y} \hat{y} + \frac{\partial}{\partial z} \hat{z}$, while $\nabla = \partial \hat{x} + \partial \hat{y} + \partial \hat{z}$, where \hat{x}, \hat{y} and \hat{z} are unit vectors in the x, y , and z directions, respectively.

We assume that the wave is propagating in a dielectric medium. It is important to further specify some properties of that medium. The medium is *linear* if the permittivity satisfies the linearity properties discussed in [Chapter 2](#). The medium is *isotropic* if the permittivity is independent of the direction of polarization of the wave (i.e. the directions of the $\vec{\mathcal{E}} \rightarrow$ and $\vec{\mathcal{H}} \rightarrow$ vectors). The medium is *homogeneous* if the permittivity is constant throughout the region of propagation. The medium is *nondispersive* if the permittivity is independent of wavelength over the wavelength region occupied by the propagating wave. Finally, all media of interest in this book are *nonmagnetic*, which means that the magnetic permeability is always equal to μ_0 , the vacuum permeability.

Applying the $\nabla \times \nabla \times$ operation to the left and right sides of the first equation for $\vec{\mathcal{E}} \rightarrow$, we make use of the vector identity

$$\nabla \times (\nabla \times \vec{\mathcal{E}} \rightarrow) = \nabla (\nabla \cdot \vec{\mathcal{E}} \rightarrow) - \nabla^2 \vec{\mathcal{E}} \rightarrow.$$

$$\nabla \times (\nabla \times \vec{\mathcal{E}}) = \nabla (\nabla \cdot \vec{\mathcal{E}}) - \nabla^2 \vec{\mathcal{E}}.$$

(3-3)

If the propagation medium is linear, isotropic, homogeneous (constant ϵ), and nondispersive, substitution of the two Maxwell's equations for $\mathcal{E} \rightarrow \vec{\mathcal{E}}$ in (3-3) yields

$$\nabla^2 \mathcal{E} \rightarrow -n^2 c^2 \partial^2 \mathcal{E} / \partial t^2 = 0$$

$$\nabla^2 \vec{\mathcal{E}} - \frac{n^2 c^2}{\epsilon_0} \frac{\partial^2 \vec{\mathcal{E}}}{\partial t^2} = 0$$

(3-4)

where n is the *refractive index* of the medium, defined by

$$n = \sqrt{\epsilon/\epsilon_0}$$

$$n = \left(\frac{\epsilon}{\epsilon_0} \right)^{1/2},$$

(3-5)

ϵ_0 is the vacuum permittivity, and c is the speed of propagation in vacuum, given by

$$c = 1/\sqrt{\mu_0 \epsilon_0}$$

$$c = \frac{1}{\sqrt{\mu_0 \epsilon_0}}.$$

(3-6)

The magnetic field satisfies an identical equation,

$$\nabla^2 \mathcal{H} \rightarrow -n^2 c^2 \partial^2 \mathcal{H} / \partial t^2 = 0.$$

$$\nabla^2 \vec{\mathcal{H}} - \frac{n^2 c^2}{\epsilon_0} \frac{\partial^2 \vec{\mathcal{H}}}{\partial t^2} = 0.$$

Since the vector wave equation is obeyed by both $\mathcal{E} \rightarrow \vec{\mathcal{E}}$ and $\mathcal{H} \rightarrow \vec{\mathcal{H}}$, an identical scalar wave equation is obeyed by all components of those vectors. Thus, for example, $\mathcal{E}_X \rightarrow \vec{\mathcal{E}}_X$ obeys the equation

$$\nabla^2 \mathcal{E}_X - n^2 c^2 \partial^2 \mathcal{E}_X / \partial t^2 = 0,$$

$$\nabla^2 \mathcal{E}_X - \frac{n^2 c^2}{\epsilon_0} \frac{\partial^2 \mathcal{E}_X}{\partial t^2} = 0,$$

and similarly for $\mathcal{E}_Y, \mathcal{E}_Z, \mathcal{H}_X, \mathcal{H}_Y, \mathcal{E}_Y, \mathcal{E}_Z, \mathcal{H}_X, \mathcal{H}_Y$, and $\mathcal{H}_Z \mathcal{H}_Z$. Therefore it is possible to summarize the behavior of all components of $\mathcal{E} \rightarrow \vec{\mathcal{E}}$ and $\mathcal{H} \rightarrow \vec{\mathcal{H}}$ through a single scalar wave equation,

$$\nabla^2 u(P,t) - n^2 c^2 \frac{\partial^2 u(P,t)}{\partial t^2} = 0,$$

$$\nabla^2 u(P,t) - \frac{n^2 \partial^2 u(P,t)}{c^2 \partial t^2} = 0,$$

(3-7)

where $u(P,t)$ represents any of the scalar field components, and we have explicitly introduced the dependence of the field u on both position P in space and time t .

From previously we conclude that in a dielectric medium that is linear, isotropic, homogeneous, and nondispersive, all components of the electric and magnetic field behave identically and their behavior is fully described by a single scalar wave equation. How, then, is the scalar theory only an approximation, rather than exact? The answer becomes clear if we consider situations other than propagation in the uniform dielectric medium hypothesized.

For example, if the medium is inhomogeneous with a permittivity $\epsilon(P)$ that depends on position P (but not on time t), it is a simple matter to show (see [Prob. 3-1](#)) that the wave equation satisfied by $\mathcal{E} \rightarrow \vec{\mathcal{E}}$ becomes

$$\nabla^2 \mathcal{E} \rightarrow + 2\nabla(\mathcal{E} \rightarrow \cdot \nabla \ln n) - n^2 c^2 \frac{\partial^2 \mathcal{E} \rightarrow}{\partial t^2} = 0,$$

$$\nabla^2 \vec{\mathcal{E}} + 2\nabla(\vec{\mathcal{E}} \cdot \nabla \ln n) - \frac{n^2 \partial^2 \vec{\mathcal{E}}}{c^2 \partial t^2} = 0,$$

(3-8)

where n and c are again given by [\(3-5\)](#) and [\(3-6\)](#). The new term that has been added to the wave equation will be nonzero for a refractive index that changes over space. More importantly, that term introduces a *coupling* between the various components of the electric field, with the result that $\mathcal{E}_X, \mathcal{E}_Y, \mathcal{E}_Z$ may no longer satisfy the *same* wave equation. This type of coupling is important, for example, when light propagates through a “thick” dielectric diffraction grating.

A similar effect takes place when boundary conditions are imposed on a wave that propagates in a homogeneous medium. At the boundaries, coupling is introduced between $\mathcal{E} \rightarrow \vec{\mathcal{E}}$ and $\mathcal{H} \rightarrow \vec{\mathcal{H}}$, as well as between their various scalar components. As a consequence, even when the propagation medium is homogeneous, the use of a scalar theory entails some degree of error. That error will be small provided the boundary conditions have effect over an area that is a small part of the area through which a wave may be passing. In the case of diffraction of light by an aperture, the $\mathcal{E} \rightarrow \vec{\mathcal{E}}$ and $\mathcal{H} \rightarrow \vec{\mathcal{H}}$ fields are modified only at the edges of the aperture where light interacts with the material of which the edges are composed, and the effects extend over only a

few wavelengths into the aperture itself. Thus if the aperture has an area that is large compared with a wavelength, the coupling effects of the boundary conditions on the $\mathcal{E} \rightarrow \vec{\mathcal{E}}$ and $\mathcal{H} \rightarrow \vec{\mathcal{H}}$ fields will be small. As will be seen, this is equivalent to the requirement that the diffraction angles caused by the aperture are small.

With these discussions as background, we turn away from the vector theory of diffraction to the simpler scalar theory. We close with an observation. Circuit theory is based on the approximation that circuit elements (resistors, capacitors, and inductors) are *small* compared to the wavelength of the fields that appear within them, and for this reason can be treated as lumped elements with simple properties. We need not use Maxwell's equations to analyze such elements under these conditions. In a similar vein, the scalar theory of diffraction introduces substantial simplifications compared with a full vectorial theory. In most applications, the scalar theory is accurate provided that the diffracting structures are *large* compared with the wavelength of light. Thus the approximation implicit in the scalar theory should be no more disturbing than the approximation used in lumped circuit theory. In both cases it is possible to find situations in which the approximation breaks down, but as long as the simpler theories are used only in cases for which they are expected to be valid, the losses of accuracy will be small and the gain of simplicity will be large. However, it is also worth mentioning that polarization can play a role in some experiments for which the scalar theory can be used, the key approximation of the scalar theory being that all polarization components of the field obey the same scalar wave equation and can be treated independently.

3.3 Some Mathematical Preliminaries

Before embarking on a treatment of diffraction itself, we first consider a number of mathematical preliminaries that form the basis of the later diffraction-theory derivations. These initial discussions will also serve to introduce some of the notation used throughout the book.

3.3.1 The Helmholtz Equation

In accord with the previous introduction of the scalar theory, let the light disturbance at position P and time t be represented by the scalar function $u(P, t)$. Attention is now restricted to the case of a purely monochromatic wave, with the generalization to polychromatic waves being deferred to [Section 3.8](#).

For a monochromatic wave, the scalar field may be written explicitly as

$$u(P, t) = A(P) \cos[2\pi\nu t - \phi(P)]$$

$$u(P, t) = A(P) \cos[2\pi\nu t - \phi(P)]$$

(3-9)

where $A(P)$ and $\phi(P)$ are the amplitude and phase, respectively, of the wave at position P , while ν is the optical frequency. A more compact form of (3-9) is found by using complex notation, writing

$$u(P, t) = \operatorname{Re}U(P)\exp(-j2\pi\nu t),$$

$$u(P, t) = \operatorname{Re}[U(P) \exp(-j2\pi\nu t)],$$

(3-10)

where $\operatorname{Re}\{\}$ signifies “real part of,” and $U(P)$ is a complex function of position (sometimes called a *phasor*),

$$U(P) = A(P) \exp[j\phi(P)].$$

$$U(P) = A(P) \exp[j\phi(P)].$$

(3-11)

If the real disturbance $u(P, t)$ is to represent an optical wave, it must satisfy the scalar wave equation

$$\nabla^2 u - n^2 c^2 \frac{\partial^2 u}{\partial t^2} = 0$$

$$\nabla^2 u - \frac{n^2}{c^2} \frac{\partial^2 u}{\partial t^2} = 0$$

(3-12)

at each source-free point. As before, ∇^2 is the Laplacian operator, n represents the refractive index of the dielectric medium within which light is propagating, and c represents the vacuum speed of light. The complex function $U(P)$ serves as an adequate description of the disturbance, since the time dependence is known a priori. If (3-10) is substituted in (3-12), it follows that U must obey the time-independent equation

$$(\nabla^2 + k^2)U = 0.$$

$$(\nabla^2 + k^2)U = 0.$$

(3-13)

Here k is termed the *wave number* and is given by

$$k = 2\pi n v c = 2\pi \lambda,$$

$$k = 2\pi n \frac{v}{c} = \frac{2\pi}{\lambda},$$

and λ is the wavelength in the dielectric medium ($\lambda = c/nv$). The relation (3-13) is known as the *Helmholtz equation*; we may assume in the future that the complex amplitude of any monochromatic optical disturbance propagating in vacuum ($n=1$) or in a homogeneous dielectric medium ($n>1$) must obey such a relation.

3.3.2 Green's Theorem

Calculation of the complex disturbance U at an observation point in space can be accomplished with the help of the mathematical relation known as *Green's theorem*. This theorem, which can be found in most texts on advanced calculus, can be stated as follows:

Let $U(P)$ and $G(P)$ be any two complex-valued functions of position, and let S be a closed surface surrounding a volume V . If U , G , and their first and second partial derivatives are single-valued and continuous within and on S , then we have

$$\iint_V (U \nabla^2 G - G \nabla^2 U) dv = \int_S \int \left(U \frac{\partial G}{\partial n} - G \frac{\partial U}{\partial n} \right) ds$$

$$\iint_V (U \nabla^2 G - G \nabla^2 U) dv = \int_S \int \left(U \frac{\partial G}{\partial n} - G \frac{\partial U}{\partial n} \right) ds$$

(3-14)

where $\partial/\partial n$ signifies a partial derivative in the *outward* normal direction at each point on S .

This theorem is in many respects the prime foundation of scalar diffraction theory. However, only a prudent choice of an auxiliary function G and a closed surface S will allow its direct application to the diffraction problem. We turn now to the former of these problems, considering Kirchhoff's choice of an auxiliary function and the consequent integral theorem that follows.

3.3.3 The Integral Theorem of Helmholtz and Kirchhoff

The Kirchhoff formulation of the diffraction problem is based on a certain integral theorem which expresses the solution of the homogeneous wave equation at an arbitrary point in terms of the values of the solution and its first derivative on an arbitrary closed surface surrounding that point. This theorem had been derived previously in acoustics by H. von Helmholtz.

Let the point of observation be denoted P_0 , and let S denote an arbitrary closed surface surrounding P_0 , as indicated in Fig. 3.5. The problem is to express the optical disturbance at P_0 in terms of its values on the surface S . To solve this problem, we follow Kirchhoff in applying Green's theorem and in choosing as an auxiliary function a unit-amplitude spherical wave expanding about the point P_0 (the so-called *free space* Green's function). Thus the value of Kirchhoff's G at an arbitrary point P_1 is given by²

$$G(P_1) = \exp(jkr_{01})r_{01},$$

$$G(P_1) = \frac{\exp(jkr_{01})}{r_{01}},$$

(3-15)

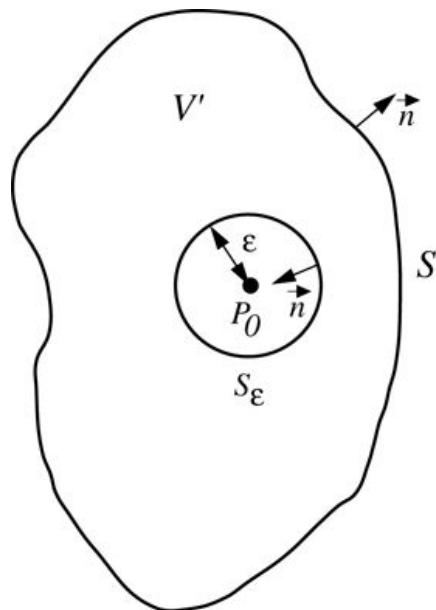


Figure 3.5

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 3.5 Surface of integration.

The illustration shows point P subscript O that is at the center of circle S epsilon of radius epsilon. On the right side of the circle, a downward sloping arrow representing vector n runs inward from the circumference of the circle toward the center. The circle is inside an irregular shape S, which is about ten times larger. In the top right corner of S is an upward sloping arrow representing vector n running outward from the boundary. The area inside the irregular shape and outside the circle is labeled V dash.

where we adopt the notation that r_{01} is the length of the vector \vec{r}_{01} pointing from P_0 to P_1 .

Before proceeding further, a short diversion regarding Green's functions may be in order. Suppose that we wish to solve an inhomogeneous linear differential equation of the form

$$a_2 d^2 U / dx^2 + a_1 dU / dx + a_0 U = V(x)$$

$$a_2 \frac{d^2 U}{dx^2} + a_1 \frac{dU}{dx} + a_0 U = V(x)$$

(3-16)

where $V(x)$ is a driving function and $U(x)$ satisfies a known set of boundary conditions. We have chosen a one-dimensional variable x but the theory is easily generalized to a multidimensional $x \rightarrow \vec{x}$. It can be shown (see [Chapter 1](#) of [269] and [16]) that if $G(x)$ is the solution to the same differential equation (3-16) when $V(x)$ is replaced by the impulsive driving function $\delta(x-x')$ and with the same boundary conditions applying, then the general solution $U(x)$ can be expressed in terms of the specific solution $G(x)$ through a convolution integral

$$U(x) = \int G(x-x') V(x') dx'.$$

$$U(x) = \int G(x - x') V(x') dx'.$$

(3-17)

The function $G(x)$ is known as the *Green's function* of the problem, and is clearly a form of impulse response. Various solutions to the scalar diffraction problem to be discussed in the following sections correspond to results obtained under different assumptions about the Green's function of the problem. The function G appearing in Green's theorem may be regarded either as simply an auxiliary function which we cleverly choose to solve our problem, or it may eventually be related to the impulse response of the problem. Further consideration of the theory of Green's functions is beyond the scope of this treatment.

Returning now to our central discussion, to be legitimately used in Green's theorem, the function G (as well as its first and second partial derivatives) must be continuous within the enclosed volume V . Therefore to exclude the discontinuity at P_0 , a small spherical surface S_ϵ , of radius ϵ , is inserted about the point P_0 . Green's theorem is then applied, the

volume of integration V' being that volume lying between S and $S \in S_\epsilon$, and the surface of integration being the composite surface

$$S' = S + S_\epsilon$$

$$S' = S + S_\epsilon$$

as indicated in [Fig. 3.5](#). Note that the “outward” normal to the composite surface points outward in the conventional sense on S , but inward (towards P_0) on $S \in S_\epsilon$.

Within the volume V' , the disturbance G , being simply an expanding spherical wave, satisfies the Helmholtz equation

$$(\nabla^2 + k^2)G = 0.$$

$$(\nabla^2 + k^2)G = 0.$$

(3-18)

Substituting the two Helmholtz equations [\(3-13\)](#) and [\(3-18\)](#) in the left-hand side of Green’s theorem, we find

$$\iint_{V'} (U \nabla^2 G - G \nabla^2 U) dv = - \iint_{V'} (UGk^2 - GUk^2) dv \equiv 0.$$

$$\iint_{V'} (U \nabla^2 G - G \nabla^2 U) dv = - \iint_{V'} (UGk^2 - GUk^2) dv \equiv 0.$$

Thus the theorem reduces to

$$\iint_{S'} (U \partial G / \partial n - G \partial U / \partial n) ds = 0$$

$$\iint_{S'} (U \frac{\partial G}{\partial n} - G \frac{\partial U}{\partial n}) ds = 0$$

or

$$-\iint_{S \in S'} (U \partial G / \partial n - G \partial U / \partial n) ds = \iint_S (U \partial G / \partial n - G \partial U / \partial n) ds.$$

$$-\iint_{S_\epsilon} (U \frac{\partial G}{\partial n} - G \frac{\partial U}{\partial n}) ds = \iint_S (U \frac{\partial G}{\partial n} - G \frac{\partial U}{\partial n}) ds.$$

(3-19)

Note that, for a general point P_1 on S' , we have

$$G(P_1) = \exp(jkr_{01})$$

$$G(P_1) = \frac{\exp(jkr_{01})}{r_{01}}$$

and

$$\partial G(P_1) \partial n = \cos(n \rightarrow, r \rightarrow 01) jk - 1 r_{01} \exp(jkr_{01}) r_{01}$$

$$\frac{\partial G(P_1)}{\partial n} = \cos(\vec{n}, \vec{r}_{01}) \left(jk - \frac{1}{r_{01}} \right) \frac{\exp(jkr_{01})}{r_{01}}$$

(3-20)

where $\cos(n \rightarrow, r \rightarrow 01) \cos(\vec{n}, \vec{r}_{01})$ represents the cosine of the angle between the outward normal $n \rightarrow \vec{n}$ and the vector $r \rightarrow 01 \vec{r}_{01}$ joining $P_0 P_0$ to $P_1 P_1$. For the particular case of $P_1 P_1$ on $S \in S_\epsilon$, $\cos(n \rightarrow, r \rightarrow 01) = -1$, and these equations become

$$G(P_1) = ejk\epsilon \epsilon \text{ and } \partial G(P_1) \partial n = ejk\epsilon \epsilon 1 \epsilon - jk.$$

$$G(P_1) = \frac{e^{jke}}{\epsilon} \text{ and } \frac{\partial G(P_1)}{\partial n} = \frac{e^{jke}}{\epsilon} \left(\frac{1}{\epsilon} - jk \right).$$

Letting $\epsilon \rightarrow 0$ become arbitrarily small, the continuity of $U U$ (and its derivatives) at $P_0 P_0$ allows us to write

$$\lim_{\epsilon \rightarrow 0} \int_S (U \partial G \partial n - G \partial U \partial n) ds = \lim_{\epsilon \rightarrow 0} 4\pi \epsilon^2 [U(P_0) \exp(jke) \epsilon (1 \epsilon - jk) - \partial U(P_0) \partial n \exp(jke) \epsilon] = 4\pi U(P_0).$$

$$\lim_{\epsilon \rightarrow 0} \int_S \left[\left(U \frac{\partial G}{\partial n} - G \frac{\partial U}{\partial n} \right) ds \right] = \lim_{\epsilon \rightarrow 0} 4\pi \epsilon^2 \left[U(P_0) \frac{\exp(jke)}{\epsilon} \left(\frac{1}{\epsilon} - jk \right) - \frac{\partial U(P_0) \exp(jke)}{\partial n \epsilon} \right] = 4\pi U(P_0).$$

Substitution of this result in (3-19) (taking account of the negative sign) yields

$$U(P_0) = 14\pi \int_S \partial U \partial n \exp(jkr_{01}) r_{01} - U \partial \partial n \exp(jkr_{01}) r_{01} ds.$$

$$U(P_0) = \frac{1}{4\pi} \int_S \left\{ \frac{\partial U}{\partial n} \left[\frac{\exp(jkr_{01})}{r_{01}} \right] - U \frac{\partial}{\partial n} \left[\frac{\exp(jkr_{01})}{r_{01}} \right] \right\} ds.$$

(3-21)

This result is known as the *integral theorem of Helmholtz and Kirchhoff*; it plays an important role in the development of the scalar theory of diffraction, for it allows the field at any point $P_0 P_0$ to be expressed in terms of the “boundary values” of the wave on any closed surface surrounding that point. As we shall now see, such a relation is instrumental in the further development of scalar diffraction equations.

3.4 The Kirchhoff Formulation of Diffraction by a Planar Screen

Consider now the problem of diffraction of light by an aperture in an infinite opaque screen. As illustrated in Fig. 3.6, a wave disturbance is assumed to impinge on the screen and the aperture from the left, and the field at the point P_0 behind the aperture is to be calculated. Again the field is assumed to be monochromatic.

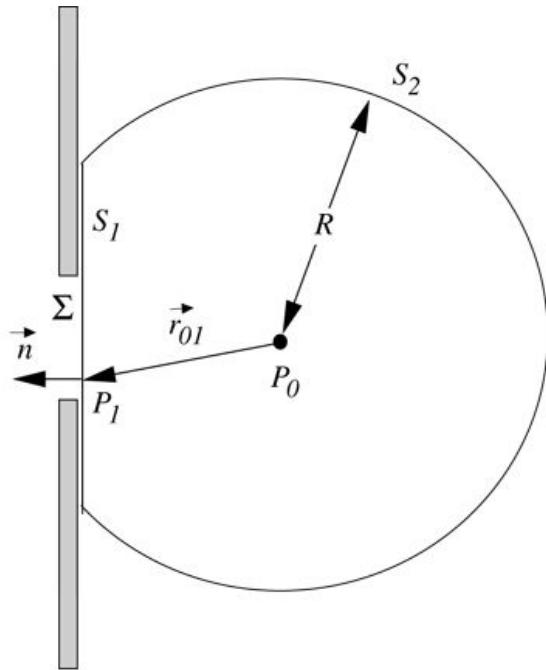


Figure 3.6
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 3.6 Kirchhoff formulation of diffraction by a plane screen.

The illustration shows a cross-sectional view of aperture sigma on a vertical screen, on the right side of which is a circular surface, S_1 , with sigma at its center. A sphere of surface S_2 with a segment removed sits flat on S_1 , which is identical to the flat portion of S_2 , which is centered at P_0 with radius R . An arrow representing vector r_{01} extends in a downward slope from P_0 to P_1 located near the lower end of sigma. On exiting sigma, the vector continues in a leftward direction but horizontally. It is now labeled vector n .

3.4.1 Application of the Integral Theorem

To find the field at the point P_0 , we apply the integral theorem of Helmholtz and Kirchhoff, being careful to choose a surface of integration that will allow the calculation to be performed successfully. Following Kirchhoff, the closed surface S is chosen to consist of two parts, as shown in Fig. 3.6. Let a plane surface, S_1 , lying directly behind the diffracting screen, be

joined and closed by a large spherical cap, S_2 , of radius R and centered at the observation point P_0 . The total closed surface S is simply the sum of S_1 and S_2 . Thus, applying (3-21),

$$U(P_0) = 14\pi \int S_1 + S_2 G \partial U \partial n - U \partial G \partial n ds,$$

$$U(P_0) = \frac{1}{4\pi} \int_{S_1 + S_2} \left(G \frac{\partial U}{\partial n} - U \frac{\partial G}{\partial n} \right) ds,$$

where, as before,

$$G = \exp(jkr_{01})r_{01}.$$

$$G = \frac{\exp(jkr_{01})}{r_{01}}.$$

As R increases, S_2 approaches a large hemispherical shell. It is tempting to reason that, since both U and G will fall off as $1/R$, the integrand will ultimately vanish, yielding a contribution of zero from the surface integral over S_2 . However, the area of integration increases as R^2 , so this argument is incomplete. It is also tempting to assume that, since the disturbances are propagating with finite speed c/n , R will ultimately be so large that the waves have not yet reached S_2 , and the integrand will be zero on that surface. But this argument is incompatible with our assumption of monochromatic disturbances, which must (by definition) have existed for all time. Evidently a more careful investigation is required before the contribution from S_2 can be disposed of.

Examining this problem in more detail, we see that, on S_2 ,

$$G = \exp(jkR)R$$

$$G = \frac{\exp(jkR)}{R}$$

and, from (3-20),

$$\partial G \partial n = jk - 1 R \exp(jkR) R \approx jkG$$

$$\frac{\partial G}{\partial n} = \left(jk - \frac{1}{R} \right) \frac{\exp(jkR)}{R} \approx jkG$$

where the last approximation is valid for large R . The integral in question can thus be reduced to

$$\int S_2 G \partial U \partial n - U(jkG) ds = \int \Omega G \partial U \partial n - jkUR^2 d\omega,$$

$$\int_{S_2} \left[G \frac{\partial U}{\partial n} - U(jkG) \right] ds = \int_{\Omega} G \left(\frac{\partial U}{\partial n} - jkU \right) R^2 d\omega,$$

where Ω is the solid angle subtended by S_2 at P_0 . Now the quantity $|RG|$ is uniformly bounded on S_2 . Therefore the entire integral over S_2 will vanish as R becomes arbitrarily large, provided the disturbance has the property

$$\lim_{R \rightarrow \infty} R \partial U / \partial n - jkU = 0$$

$$\lim_{R \rightarrow \infty} R \left(\frac{\partial U}{\partial n} - jkU \right) = 0$$

(3-22)

uniformly in angle. This requirement is known as the *Sommerfeld radiation condition* [327] and is satisfied if the disturbance U vanishes at least as fast as a diverging spherical wave (see [Prob. 3-2](#)). It guarantees that we are dealing only with *outgoing* waves on S_2 , rather than incoming waves, for which the integral over S_2 might not vanish as $R \rightarrow \infty$. Since only outgoing waves will fall on S_2 in our problem, the integral over S_2 will yield a contribution of precisely zero.

3.4.2 The Kirchhoff Boundary Conditions

Having disposed of the integration over the surface S_2 , it is now possible to express the disturbance at P_0 in terms of the disturbance and its normal derivative over the infinite plane S_1 immediately behind the screen, that is,

$$U(P_0) = 14\pi \int_{S_1} \partial U / \partial n G - U \partial G / \partial n ds.$$

$$U(P_0) = \frac{1}{4\pi} \int_{S_1} \left[\left(\frac{\partial U}{\partial n} G - U \frac{\partial G}{\partial n} \right) ds \right]$$

(3-23)

The screen is opaque, except for the open aperture which will be denoted Σ . It therefore seems intuitively reasonable that the major contribution to the integral (3-23) arises from the points of S_1 located within the aperture Σ , where we would expect the integrand to be largest. Kirchhoff accordingly adopted the following assumptions [194]:

1. Across the surface Σ , the field distribution U and its derivative $\partial U / \partial n$ are exactly the same as they would be in the absence of the screen.
2. Over the portion of S_1 that lies in the geometrical shadow of the screen, the field distribution U and its derivative $\partial U / \partial n$ are identically zero.

These conditions are commonly known as the *Kirchhoff boundary conditions*. The first allows us to specify the disturbance incident on the aperture by neglecting the presence of the screen. The second allows us to neglect all of the surface of integration except that portion lying directly within the aperture itself. Thus (3-23) is reduced to

$$U(P_0) = 14\pi \int \partial U \partial n G - U \partial G \partial n ds.$$

$$U(P_0) = \frac{1}{4\pi} \int_{\Sigma} \left(\frac{\partial U}{\partial n} G - U \frac{\partial G}{\partial n} \right) ds.$$

(3-24)

While the Kirchhoff boundary conditions simplify the results considerably, it is important to realize that neither can be exactly true. The presence of the screen will inevitably perturb the fields on Σ to some degree, for along the rim of the aperture certain boundary conditions must be met that would not be required in the absence of the screen. In addition, the shadow behind the screen is never perfect, for fields will inevitably extend behind the screen for a distance of several wavelengths. However, if the dimensions of the aperture are large compared with a wavelength, these fringing effects can be safely neglected,³ and the two boundary conditions can be used to yield results that agree very well with experiment.

3.4.3 The Fresnel-Kirchhoff Diffraction Formula

A further simplification of the expression for $U(P_0)$ is obtained by noting that the distance r_{01} from the aperture to the observation point is usually many optical wavelengths, and therefore, since $k \gg 1/r_{01}$, (3-20) becomes

$$\partial G(P_1) \partial n = \cos(\vec{n}, \vec{r}_{01}) jk - 1/r_{01} \exp(jkr_{01}) r_{01} \approx jk \cos(\vec{n}, \vec{r}_{01}) \exp(jkr_{01}).$$

$$\begin{aligned} \frac{\partial G(P_1)}{\partial n} &= \cos(\vec{n}, \vec{r}_{01}) \left(jk - \frac{1}{r_{01}} \right) \frac{\exp(jkr_{01})}{r_{01}} \\ &\approx jk \cos(\vec{n}, \vec{r}_{01}) \frac{\exp(jkr_{01})}{r_{01}}. \end{aligned}$$

(3-25)

Substituting this approximation and the expression (3-15) for G in Eq. (3-24), we find

$$U(P_0) = 14\pi \int \exp(jkr_{01}) r_{01} \partial U \partial n - jk U \cos(\vec{n}, \vec{r}_{01}) ds.$$

$$U(P_0) = \frac{1}{4\pi} \int_{\Sigma} \left[\frac{\exp(jkr_{01})}{r_{01}} \left(\frac{\partial U}{\partial n} - jk U \cos(\vec{n}, \vec{r}_{01}) \right) \right] ds.$$

(3-26)

Now suppose that the aperture is illuminated by a single spherical wave,

$$U(P_1) = A \exp(jkr_{21}) r_{21}$$

$$U(P_1) = \frac{A \exp(jkr_{21})}{r_{21}}$$

arising from a point source at P_2 , a distance r_{21} from P_1 (see Fig. 3.7). If r_{21} is many optical wavelengths, then (3-26) can be directly reduced (see Prob. 3-3) to

$$U(P_0) = A j \lambda \int \Sigma \exp[jk(r_{21} + r_{01})] r_{21} r_{01} \cos(\vec{n} \cdot \vec{r}_{01}) - \cos(\vec{n} \cdot \vec{r}_{21}) 2 ds.$$

$$U(P_0) = \frac{A}{j\lambda} \int_{\Sigma} \frac{\exp[jk(r_{21} + r_{01})]}{r_{21} r_{01}} \left[\frac{\cos(\vec{n} \cdot \vec{r}_{01}) - \cos(\vec{n} \cdot \vec{r}_{21})}{2} \right] ds. \quad (3-27)$$

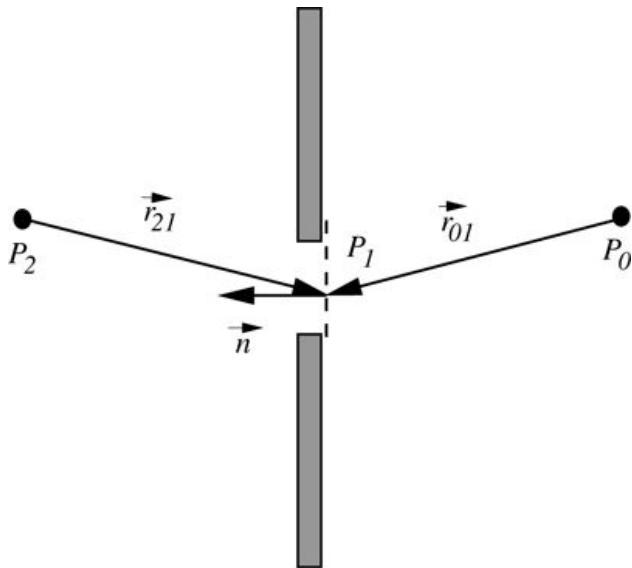


Figure 3.7
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 3.7 Point-source illumination of a plane screen.

The illustration shows a cross-sectional view of an aperture in a screen. A downward sloping ray representing vector \vec{r}_{01} coming from P_0 on the right side is incident at P_1 near the middle of the aperture where the vector turns horizontal to represent vector \vec{n} . On the left side, vector \vec{r}_{21} , originating at P_2 , is a downward sloping ray incident at P_1 .

This result, which holds only for an illumination consisting of an expanding spherical wave, is known as the *Fresnel-Kirchhoff diffraction formula*.

Note that Eq. (3-27) is symmetrical with respect to the illumination point source at P_2 and the observation point at P_0 . Thus a point source at P_0 will produce at P_2 the same effect that a point source of equal intensity placed at P_2 will produce at P_0 . This result is referred to as the *reciprocity theorem of Helmholtz*.

Finally, we point out an interesting interpretation of the diffraction formula (3-27), to which we will return later for a more detailed discussion. Let that equation be rewritten as follows:

$$U(P_0) = \int_{\Sigma} \int U(P_1) \exp(jkr_{01}) r_{01} ds,$$

$$U(P_0) = \int_{\Sigma} \int \tilde{U}(P_1) \frac{\exp(jkr_{01})}{r_{01}} ds,$$

(3-28)

where

$$U(P_1) = 1j\lambda A \exp(jkr_{21}) r_{21} \cos(n \rightarrow, r \rightarrow 01) - \cos(n \rightarrow, r \rightarrow 21) 2.$$

$$\tilde{U}(P_1) = \frac{1}{j\lambda} \left[\frac{A \exp(jkr_{21})}{r_{21}} \right] \left[\frac{\cos(\vec{n}, \vec{r}_{01}) - \cos(\vec{n}, \vec{r}_{21})}{2} \right].$$

(3-29)

Now (3-28) may be interpreted as implying that the field at P_0 arises from an infinity of fictitious "secondary" point sources located within the aperture itself. The secondary sources have certain amplitudes and phases, described by $U(P_1)$, that are related to the illuminating wavefront and the angles of illumination and observation. Assumptions resembling these were made by Fresnel rather arbitrarily in his combination of Huygens' envelope construction and Young's principle of interference. Fresnel *assumed* these properties to hold in order to obtain accurate results. Kirchhoff showed that such properties are a natural consequence of the wave nature of light.

Note that the above derivation has been restricted to the case of an aperture illumination consisting of a single expanding spherical wave. However, as we shall now see, such a limitation can be removed by the Rayleigh-Sommerfeld theory.

3.5 The Rayleigh-Sommerfeld Formulation of Diffraction

The Kirchhoff theory has been found experimentally to yield remarkably accurate results and is widely used in practice. However, there are certain internal inconsistencies in the theory which motivated a search for a more satisfactory mathematical development. The difficulties of the Kirchhoff theory stem from the fact that boundary conditions must be imposed on *both* the field strength and its normal derivative. In particular, it is a well-known theorem of potential theory that if a two-dimensional potential function and its normal derivative vanish *together* along any finite curve segment, then that potential function *must vanish over the entire plane*. Similarly, if a solution of the three-dimensional wave equation and its normal derivative vanish on any finite surface element, the field must vanish in all space. Thus the two Kirchhoff boundary conditions together imply that the field is zero everywhere behind the aperture, a result which contradicts the known physical situation. A further indication of these inconsistencies is the fact that the Fresnel-Kirchhoff diffraction formula can be shown to fail to reproduce the assumed boundary conditions as the observation point approaches the screen or aperture. In view of these contradictions, it is indeed remarkable that the Kirchhoff theory has been found to yield such accurate results in practice.⁴

The inconsistencies of the Kirchhoff theory were removed by Sommerfeld, who eliminated the necessity of imposing boundary values on both the disturbance and its normal derivative simultaneously. This so-called Rayleigh-Sommerfeld theory is the subject of this section.

3.5.1 Choice of Alternative Green's Functions

Consider again [Eq.\(3-23\)](#) for the observed field strength in terms of the incident field and its normal derivative across the entire screen:

$$U(P_0) = \frac{1}{4\pi} \int_{S_1} \int \left(\frac{\partial U}{\partial n} G - U \frac{\partial G}{\partial n} \right) ds.$$

$$U(P_0) = \frac{1}{4\pi} \int_{S_1} \int \left(\frac{\partial U}{\partial n} G - U \frac{\partial G}{\partial n} \right) ds.$$

(3-30)

The conditions for validity of this equation are:

1. The scalar theory holds.
2. Both U and G satisfy the homogeneous scalar wave equation.
3. The Sommerfeld radiation condition is satisfied.

Suppose that the Green's function G of the Kirchhoff theory were modified in such a way that, while the development leading to the above equation remains valid, in addition, either G or $\partial G / \partial n$ vanishes over the entire surface S_1 . In either case the necessity of imposing

boundary conditions on both U and $\partial U / \partial n$ would be removed, and the inconsistencies of the Kirchhoff theory would be eliminated.

Sommerfeld pointed out that Green's functions with the required properties do indeed exist. Suppose G is generated not only by a point source located at P_0 , but also simultaneously by a second point source at a position \tilde{P}_0 which is the mirror image of P_0 on the opposite side of the screen (see Fig. 3.8). Let the source at \tilde{P}_0 be of the same wavelength λ as the source at P_0 , and suppose that the two sources are oscillating with a 180° phase difference. The Green's function in this case is given by

$$G(P_1) = \exp(jkr_{01})r_{01} - \exp(jk\tilde{r}_{01})\tilde{r}_{01}.$$

$$G(P_1) = \frac{\exp(jkr_{01})}{r_{01}} - \frac{\exp(jk\tilde{r}_{01})}{\tilde{r}_{01}}.$$

(3-31)

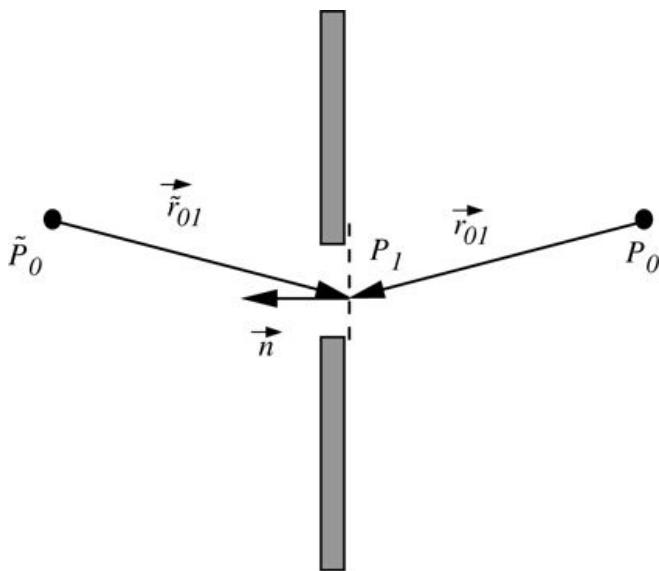


Figure 3.8

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 3.8 Rayleigh-Sommerfeld formulation of diffraction by a plane screen.

An illustration shows a cross-sectional view of an aperture in a screen. A downward sloping ray representing vector r_{01} coming from P_0 on the right side is incident at P_1 near the middle of the aperture where the vector turns horizontal to represent vector n . On the left side, vector $r_{\tilde{0}1}$ originating at \tilde{P}_0 , is a downward sloping ray incident at P_1 .

Clearly such a function vanishes on the plane aperture Σ , and the Kirchhoff boundary conditions may be applied to U alone, leaving the following expression for the observed field:

$$UI(P_0) = -14\pi \int_U \partial G \cdot \partial n ds.$$

$$U_I(P_0) = -\frac{1}{4\pi} \int_{\Sigma} \int U \frac{\partial G_-}{\partial n} ds.$$

(3-32)

We refer to this solution as the *first Rayleigh-Sommerfeld solution*.

To specify this solution further let \tilde{r}_{01} be the distance from P_0 to P_1 . The corresponding normal derivative of G_- is

$$\partial G_- / \partial n(P_1) = \cos(n \rightarrow, r \rightarrow 01) jk - 1 r_{01} \exp(jkr_{01}) r_{01} - \cos(n \rightarrow, \tilde{r} \rightarrow 01) jk - 1 \tilde{r}_{01} \exp(jk\tilde{r}_{01}) \tilde{r}_{01}.$$

$$\begin{aligned} \frac{\partial G_-}{\partial n}(P_1) &= \cos(\vec{n}, \vec{r}_{01})(jk - \frac{1}{r_{01}}) \frac{\exp(jkr_{01})}{r_{01}} \\ &\quad - \cos(\vec{n}, \vec{r}_{01})(jk - \frac{1}{\tilde{r}_{01}}) \frac{\exp(jk\tilde{r}_{01})}{\tilde{r}_{01}}. \end{aligned}$$

(3-33)

Now for P_1 on S_1 , we have

$$r_{01} = \tilde{r}_{01} \cos(n \rightarrow, r \rightarrow 01) = -\cos(n \rightarrow, \tilde{r} \rightarrow 01)$$

$$\begin{aligned} r_{01} &= \tilde{r}_{01} \\ \cos(\vec{n}, \vec{r}_{01}) &= -\cos(\vec{n}, \vec{r}_{01}) \end{aligned}$$

and therefore on that surface

$$\partial G_- / \partial n(P_1) = 2 \cos(n \rightarrow, r \rightarrow 01) jk - 1 r_{01} \exp(jkr_{01}) r_{01}.$$

$$\frac{\partial G_-}{\partial n}(P_1) = 2 \cos(\vec{n}, \vec{r}_{01})(jk - \frac{1}{r_{01}}) \frac{\exp(jkr_{01})}{r_{01}}.$$

(3-34)

Assuming that $U(P_1)$ vanishes in the shadow of the screen and is unperturbed in the open aperture Σ , it follows that the first Sommerfeld solution for the field at P_0 is given by

$$U(P_0) = -12\pi \int_{\Sigma} U(P_1) jk - 1 r_{01} \exp(jkr_{01}) r_{01} \cos(n \rightarrow, r \rightarrow 01) ds,$$

$$U_I(P_0) = -\frac{1}{2\pi} \int_{\Sigma} \int U(P_1) \left(jk - \frac{1}{r_{01}} \right) \frac{\exp(jkr_{01})}{r_{01}} \cos(\vec{n}, \vec{r}_{01}) ds,$$

(3-35)

which can be written as a convolution,

$$UI(x,y,z)=h(x,y,z)*U(x,y,0),$$

$$U_I(x, y, z) = h(x, y, z) * U(x, y, 0),$$

(3-36)

where $h(x,y,z)$ is the impulse response for this solution,

$$h(x,y,z)=12\pi zr^1r-jk\exp(jkr)r,$$

$$h(x, y, z) = \frac{1}{2\pi r} \left(\frac{1}{r} - jk \right) \frac{\exp(jkr)}{r},$$

(3-37)

and we have used the facts that $\cos(n \rightarrow, r \rightarrow 01) = z/r$ and $r = \sqrt{x^2 + y^2 + z^2}$. This result represents the full expression for the first Sommerfeld solution.

Note that the normal derivative of the Green's function G_- is simply twice the normal derivative of the Green's function G used in the Kirchhoff solution, i.e.

$$\partial G_-(P_1) \partial n = 2 \partial G(P_1) \partial n.$$

$$\frac{\partial G_-(P_1)}{\partial n} = 2 \frac{\partial G(P_1)}{\partial n}.$$

An alternative and equally valid Green's function is found by allowing the two point sources to oscillate *in phase*, giving

$$G_+(P_1) = \exp(jkr_{01})r_{01} + \exp(jk\tilde{r}_{01})\tilde{r}_{01}.$$

$$G_+(P_1) = \frac{\exp(jkr_{01})}{r_{01}} + \frac{\exp(jk\tilde{r}_{01})}{\tilde{r}_{01}}.$$

(3-38)

It is readily shown (see [Prob. 3-4](#)) that the *normal derivative* of this function vanishes across the screen and aperture, leading to the *second Rayleigh-Sommerfeld solution*,

$$UII(P_0) = 14\pi \int S_1 \partial U(P_1) \partial n G_+(P_1) ds.$$

$$U_{II}(P_0) = \frac{1}{4\pi} \int_{S_1} \int \frac{\partial U(P_1)}{\partial n} G_+(P_1) ds.$$

(3-39)

It can be shown that, on Σ , G_+ is twice the Kirchhoff Green's function G ,

$$G_+ = 2G.$$

$$G_+ = 2G.$$

This leads to an expression for $U(P_0)$ in terms of the Green's function used by Kirchhoff,
 $UII(P_0)=12\pi\int\Sigma \int\partial U(P_1)\partial n \exp(jkr_{01})r_{01}ds,$

$$U_{II}(P_0) = \frac{1}{2\pi} \int_{\Sigma} \int \frac{\partial U(P_1)}{\partial n} \frac{\exp(jkr_{01})}{r_{01}} ds, \quad (3-40)$$

under the assumption that the normal derivative of $U(P_1)$ is zero outside of the aperture Σ . In general the first Sommerfeld solution is the easiest to apply, since it depends on U rather than $\partial U / \partial n$.

3.5.2 The Rayleigh-Sommerfeld Diffraction Formula

Under the assumption that $r_{01} \gg \lambda$, Eq.(3-35) reduces to⁵

$$UI(P_0)=1j\lambda\int S_1 \int U(P_1)\exp(jkr_{01})r_{01}\cos(n \rightarrow, r \rightarrow 01)ds.$$

$$U_I(P_0) = \frac{1}{j\lambda} \int_{S_1} \int U(P_1) \frac{\exp(jkr_{01})}{r_{01}} \cos(\vec{n}, \vec{r}_{01}) ds. \quad (3-41)$$

The Kirchhoff boundary conditions may now be applied to U alone, yielding the general result

$$UI(P_0)=1j\lambda\int\Sigma \int U(P_1)\exp(jkr_{01})r_{01}\cos(n \rightarrow, r \rightarrow 01)ds.$$

$$U_I(P_0) = \frac{1}{j\lambda} \int_{\Sigma} \int U(P_1) \frac{\exp(jkr_{01})}{r_{01}} \cos(\vec{n}, \vec{r}_{01}) ds. \quad (3-42)$$

Since no boundary conditions need be applied to $\partial U / \partial n$, the inconsistencies of the Kirchhoff theory have been removed.

We now specialize (3-42) and (3-40) to the case of illumination with a diverging spherical wave, allowing direct comparison with Eq.(3-27) of the Kirchhoff theory. The illumination of the aperture in all cases is a spherical wave diverging from a point source at position P_2 (see Fig. 3.7 again):

$$U(P_1)=A\exp(jkr_{21})r_{21}.$$

$$U(P_1) = A \frac{\exp(jkr_{21})}{r_{21}}.$$

Using G- G_- we obtain

$$UI(P_0) = A j \lambda \int_{\Sigma} \exp[jk(r_{21} + r_{01})] r_{21} r_{01} \cos(\vec{n} \cdot \vec{r}_{01}) ds.$$

$$U_I(P_0) = \frac{A}{j\lambda} \int_{\Sigma} \int \frac{\exp[jk(r_{21} + r_{01})]}{r_{21} r_{01}} \cos(\vec{n} \cdot \vec{r}_{01}) ds. \quad (3-43)$$

This result is known as the *Rayleigh-Sommerfeld diffraction formula*.

Using G_+ , and assuming that $r_{21} \gg \lambda$, the corresponding result is

$$UII(P_0) = -A j \lambda \int_{\Sigma} \exp[jk(r_{21} + r_{01})] r_{21} r_{01} \cos(\vec{n} \cdot \vec{r}_{21}) ds$$

$$U_{II}(P_0) = -\frac{A}{j\lambda} \int_{\Sigma} \int \frac{\exp[jk(r_{21} + r_{01})]}{r_{21} r_{01}} \cos(\vec{n} \cdot \vec{r}_{21}) ds \quad (3-44)$$

where the angle between \vec{n} and \vec{r}_{21} is greater than 90° .

3.5.3 Reproduction of Boundary Conditions

Most commonly, the diffraction formulas are stated under the approximation $r_{01} \gg \lambda$, which is satisfied in the vast majority of practical applications. Focusing on the Rayleigh-Sommerfeld case, we know that the complete solution requires an impulse response (see (3-37))

$$h(x, y, z) = 12\pi z r_1 r_j k \exp(jkr) r = h_1(x, y, z) + h_2(x, y, z),$$

$$h(x, y, z) = \frac{1}{2\pi} \frac{z}{r} \left(\frac{1}{r} - jk \right) \frac{\exp(jkr)}{r} = h_1(x, y, z) + h_2(x, y, z),$$

$$(3-45)$$

where $r = \sqrt{x^2 + y^2 + z^2}$, while

$$h_1(x, y, z) = z j \lambda \exp(jkr) r^2$$

$$h_1(x, y, z) = \frac{z}{j\lambda} \frac{\exp(jkr)}{r^2}$$

$$(3-46)$$

and

$$h_2(x, y, z) = z^2 \pi \exp(jkr) r^3.$$

$$h_2(x, y, z) = \frac{z}{2\pi} \frac{\exp(jkr)}{r^3}.$$

(3-47)

When $r \gg \lambda$, h can be approximated by h_1 while when $r \ll \lambda$, h can be approximated by h_2 .

Because of the approximation that leads to h_1 , we can not expect the solution based on this impulse response to reproduce the boundary conditions as $z \rightarrow 0$. When $r \ll \lambda$, the approximate impulse response must take the alternative form h_2 . Using *Mathematica*, it is possible to show that the volume under h_1 approaches zero as $z \rightarrow 0$, and therefore it cannot approach a delta function⁶. On the other hand, the volume under h_2 can be shown to be unity as $z \rightarrow 0$, and the values of h_2 approach 0 for $\rho \neq 0$ and ∞ for $\rho = 0$, which are the properties of a delta function.

We conclude that both h and h_2 reproduce the boundary conditions, but h_1 does not. However, since almost all practical applications involve distances such that $r \gg \lambda$, the approximation $h \approx h_1$ will be made on most occasions in the future.

3.6 Kirchhoff and Rayleigh-Sommerfeld Theories Compared

We briefly summarize the similarities and differences of the Kirchhoff and the Rayleigh-Sommerfeld theories. For the purposes of this section, let G_K represent the Green's function for the Kirchhoff theory, while G_- and G_+ are the Green's functions for the two Rayleigh-Sommerfeld formulations. As pointed out earlier, on the surface Σ , $G_+ = 2G_K$ and $\partial G_- / \partial n = 2\partial G_K / \partial n$. Therefore the general results of interest are as follows. For the Kirchhoff theory (cf. (3-24))

$$U(P_0) = 14\pi \int \partial U \partial n G_K - U \partial G_K \partial n ds,$$

$$U(P_0) = \frac{1}{4\pi} \int_{\Sigma} \left(\frac{\partial U}{\partial n} G_K - U \frac{\partial G_K}{\partial n} \right) ds, \quad (3-48)$$

for the first Rayleigh-Sommerfeld solution (cf. (3-32))

$$U_I(P_0) = -12\pi \int U \partial G_K \partial n ds,$$

$$U_I(P_0) = -\frac{1}{2\pi} \int_{\Sigma} U \frac{\partial G_K}{\partial n} ds, \quad (3-49)$$

and for the second Rayleigh-Sommerfeld solution (cf. (3-40))

$$U_{II}(P_0) = 12\pi \int \partial U \partial n G_K ds.$$

$$U_{II}(P_0) = \frac{1}{2\pi} \int_{\Sigma} \int \frac{\partial U}{\partial n} G_K ds. \quad (3-50)$$

A comparison of the above equations leads us to an interesting and surprising conclusion: *the Kirchhoff solution is the arithmetic average of the two Rayleigh-Sommerfeld solutions!*

Comparing the results of the three approaches for the case of spherical wave illumination and $r_{01} \gg \lambda$, we see that the results derived from the Rayleigh-Sommerfeld theory (i.e. (3-43) and (3-44)) differ from the Fresnel-Kirchhoff diffraction formula, (3-27), only through what is known as the *obliquity factor* ψ , which is the angular dependence introduced by the cosine terms. For all cases we can write

$$U(P_0) = A j \lambda \int \exp[jk(r_{21} + r_{01})] r_{21} r_{01} \psi ds,$$

$$U(P_0) = \frac{A}{j\lambda} \int_{\Sigma} \int \frac{\exp [jk(r_{21} + r_{01})]}{r_{21} r_{01}} \psi ds,$$

(3-51)

where

$\psi = 12[\cos(n \rightarrow, r \rightarrow 01) - \cos(n \rightarrow, r \rightarrow 21)]$ Kirchhoff theory $\cos(n \rightarrow, r \rightarrow 01)$ First Rayleigh-Sommerfeld solution $-\cos(n \rightarrow, r \rightarrow 21)$ Second Rayleigh-Sommerfeld solution.

$$\psi = \begin{cases} \frac{1}{2} [\cos(\vec{n}, \vec{r}_{01}) - \cos(\vec{n}, \vec{r}_{21})] & \text{Kirchhoff theory} \\ \cos(\vec{n}, \vec{r}_{01}) & \text{First Rayleigh-Sommerfeld solution} \\ -\cos(\vec{n}, \vec{r}_{21}) & \text{Second Rayleigh-Sommerfeld solution.} \end{cases}$$

(3-52)

For the special case of an infinitely distant point source producing normally incident plane wave illumination, the obliquity factors become

$\psi = 12[1 + \cos\theta]$ Kirchhoff theory $\cos\theta$ First Rayleigh-Sommerfeld solution 1 Second Rayleigh-Sommerfeld solution,

$$\psi = \begin{cases} \frac{1}{2}[1 + \cos\theta] & \text{Kirchhoff theory} \\ \cos\theta & \text{First Rayleigh-Sommerfeld solution} \\ 1 & \text{Second Rayleigh-Sommerfeld solution,} \end{cases}$$

(3-53)

where θ is the angle between the vectors $n \rightarrow \vec{n}$ and $r \rightarrow 01 \vec{r}_{01}$.

Several authors have compared the two formulations of the diffraction problem. We mention in particular [Wolf and Marchand \[377\]](#), who examined differences between the two theories for circular apertures with observation points at a sufficiently great distance from the aperture to be in the “far field” (the meaning of this term will be explained in the chapter to follow). They found the Kirchhoff solution and the two Rayleigh-Sommerfeld solutions to be essentially the same provided the aperture diameter is much greater than a wavelength. [Heurtley \[168\]](#) examined the predictions of the three solutions for observation points on the axis of a circular aperture for all distances behind the aperture, and found differences between the theories only close to the aperture.

When only small angles are involved in the diffraction problem, it is easy to show that all three solutions are identical. In all three cases the obliquity factors approach unity as the angles become small, and the differences between the results vanish. Note that only small angles will be involved if we are far from the diffracting aperture.

In closing it is worth noting that, in spite of its internal inconsistencies, there is one sense in which the Kirchhoff theory is more general than the Rayleigh-Sommerfeld theory. The latter requires that the diffracting screens be *planar*, while the former does not. However, most of the problems of interest here will involve planar diffracting apertures, so this generality will not be

particularly significant. In fact, we will generally choose to use the first Rayleigh-Sommerfeld solution because of its simplicity.

z

3.7 Further Discussion of the Huygens-Fresnel Principle

The Huygens-Fresnel principle, as predicted by the first Rayleigh-Sommerfeld solution^Z (see (3-42), which assumes $r_{01} \gg \lambda$), can be expressed mathematically as follows:

$$U(P_0) = j\lambda \int \Sigma U(P_1) \exp(jkr_{01}) r_{01} \cos\theta ds,$$

$$U(P_0) = \frac{1}{j\lambda} \int_{\Sigma} U(P_1) \frac{\exp(jkr_{01})}{r_{01}} \cos\theta ds,$$

(3-54)

where θ is the angle between the vectors $n \rightarrow \vec{n}$ and $r \rightarrow \vec{r}$. We give a “quasi-physical” interpretation to this integral. It expresses the observed field $U(P_0)$ as a superposition of diverging spherical waves $\exp(jkr_{01})/r_{01}$ originating from secondary sources located at each and every point P_1 within the aperture Σ . The secondary source at P_1 has the following properties:

1. It has a complex amplitude that is proportional to the amplitude of the excitation $U(P_1)$ at the corresponding point.
2. It has an amplitude that is inversely proportional to λ , or equivalently directly proportional to the optical frequency $v\nu$.
3. It has a phase that *leads* the phase of the incident wave by 90° , as indicated by the factor $1/j$.
4. Each secondary source has a directivity pattern $\cos\theta$.

The first of these properties is entirely reasonable. The wave propagation phenomenon is linear, and the wave passed through the aperture should be proportional to the wave incident upon it.

A reasonable explanation of the second and third properties would be as follows. Wave motion from the aperture to the observation point takes place by virtue of *temporal changes* of the field in the aperture. In the next section we will see more explicitly that the field at P_0 contributed by a secondary source at P_1 depends on the time rate of change of the field at P_1 . Since our basic monochromatic field disturbance is a clockwise rotating phasor of the form $\exp(-j2\pi\nu t)$, the derivative of this function will be proportional to both $v\nu$ and to $-j=1/j$.

The last property, namely the obliquity factor, has no simple “quasi-physical” explanation, but arises in slightly different forms in all the theories of diffraction. It is perhaps expecting too much to find such an explanation. After all, there are no material sources within the aperture; rather, they all lie on the rim of the aperture. Therefore the Huygens-Fresnel principle should be regarded as a relatively simple mathematical construct that allows us to solve diffraction problems without paying attention to the physical details of exactly what is happening at the edges of the aperture.

It is important to realize that the Huygens-Fresnel principle, as expressed by (3-54), is nothing more than a *superposition integral* of the type discussed in [Chapter 2](#). To emphasize this point of view we rewrite (3-54) as

$$U(P_0) = \int \Sigma h(P_0, P_1) U(P_1) ds,$$

$$U(P_0) = \int_{\Sigma} \int h(P_0, P_1) U(P_1) ds,$$

(3-55)

where the impulse response $h(P_0, P_1)$ is given explicitly in the approximate solution by

$$h(P_0, P_1) = 1/j\lambda \exp(jkr_{01}) r_{01} \cos\theta,$$

$$h(P_0, P_1) = \frac{1}{j\lambda} \frac{\exp(jkr_{01})}{r_{01}} \cos\theta,$$

(3-56)

or, in the more complete result in which r_{01} is not assumed to be much larger than λ ,

$$h(P_0, P_1) = 12\pi r_{01} - jk \exp(jkr_{01}) \cos\theta.$$

$$h(P_0, P_1) = \frac{1}{2\pi} \left(\frac{1}{r_{01}} - jk \right) \exp(jkr_{01}) \cos\theta.$$

(3-57)

The occurrence of a superposition integral as a result of our diffraction analysis should not be a complete surprise. The primary ingredient required for such a result was previously seen to be *linearity*, a property that was assumed early in our analysis. When we examine the character of the impulse response $h(P_0, P_1)$ in more detail in [Chapter 4](#), we will find that it is also *space-invariant*, a consequence of the homogeneity assumed for the dielectric medium. The Huygens-Fresnel principle will then be seen to be a *convolution* integral.

3.8 Generalization to Nonmonochromatic Waves

The wave disturbances have previously been assumed to be ideally monochromatic in all cases. Such waves can be closely approximated in practice and are particularly easy to analyze. However, the more general case of a nonmonochromatic disturbance will now be considered briefly; attention is restricted to the predictions of the first Rayleigh-Sommerfeld solution, but similar results can be obtained for the other solutions.

Consider the scalar disturbance $u(P_0, t)$ observed behind an aperture Σ in an opaque screen when a disturbance $u(P_1, t)$ is incident on that aperture. The time functions $u(P_0, t)$ and $u(P_1, t)$ may be expressed in terms of their inverse Fourier transforms:

$$u(P_1, t) = \int_{-\infty}^{\infty} U(P_1, v) \exp(j2\pi vt) dv, \quad u(P_0, t) = \int_{-\infty}^{\infty} U(P_0, v) \exp(j2\pi vt) dv,$$

$$\begin{aligned} u(P_1, t) &= \int_{-\infty}^{\infty} U(P_1, \nu) \exp(j2\pi\nu t) d\nu \\ u(P_0, t) &= \int_{-\infty}^{\infty} U(P_0, \nu) \exp(j2\pi\nu t) d\nu, \end{aligned}$$

(3-58)

where $U(P_0, \nu)$ and $U(P_1, \nu)$ are the Fourier spectra of $u(P_0, t)$ and $u(P_1, t)$, respectively, and ν represents frequency.

Let (3-58) be transformed by the change of variables $v' = -\nu$, yielding

$$u(P_1, t) = \int_{-\infty}^{\infty} U(P_1, -v') \exp(-j2\pi v' t) dv' \quad u(P_0, t) = \int_{-\infty}^{\infty} U(P_0, -v') \exp(-j2\pi v' t) dv'.$$

$$\begin{aligned} u(P_1, t) &= \int_{-\infty}^{\infty} U(P_1, -\nu') \exp(-j2\pi\nu' t) d\nu' \\ u(P_0, t) &= \int_{-\infty}^{\infty} U(P_0, -\nu') \exp(-j2\pi\nu' t) d\nu'. \end{aligned}$$

(3-59)

Now these relations may be regarded as expressing the nonmonochromatic time functions $u(P_1, t)$ and $u(P_0, t)$ as a linear combination of monochromatic time functions of the type represented by (3-10). The monochromatic elementary functions are of various frequencies $v' \nu'$, the complex amplitudes of the disturbance at frequency $v' \nu'$ being simply $U(P_1, -v')$ and $U(P_0, -v')$. By invoking the linearity of the wave-propagation phenomenon, we use the results of the previous section to find the complex amplitude

at P_0 of each monochromatic component of the disturbance, and superimpose these results to yield the general time function $u(P_0, t)$.

To proceed, (3-54) can be directly used to write

$$U(P_0, -v') = -jv' \int_{\Sigma} U(P_1, -v') \exp[j2\pi v' r_{01} / v] \cos(\vec{n}, \vec{r}_{01}) ds,$$

$$U(P_0, -v') = -j \frac{v'}{v} \int_{\Sigma} U(P_1, -v') \frac{\exp(j2\pi v' r_{01} / v)}{r_{01}} \cos(\vec{n}, \vec{r}_{01}) ds,$$

(3-60)

where v' is the speed of propagation of the disturbance in a medium of refractive index n ($v=c/n$), and the relation $v'\lambda=v\nu'\lambda=v$ has been used. Substitution of (3-60) in the second of (3-59) and an interchange of the orders of integration give

$$u(P_0, t) = \int_{\Sigma} \int \cos(n, r_{01}) 2\pi v r_{01} \int_{-\infty}^{\infty} -j2\pi v' U(P_1, -v') \exp(-j2\pi v' t - r_{01} v' ds) dv' ds.$$

$$u(P_0, t) = \int_{\Sigma} \int \frac{\cos(\vec{n}, \vec{r}_{01})}{2\pi v r_{01}} \int_{-\infty}^{\infty} -j2\pi v' U(P_1, -v') \exp\left[-j2\pi v' \left(t - \frac{r_{01}}{v}\right)\right] dv' ds.$$

Finally, the identity

$$\frac{d}{dt} u(P_1, t) = \frac{d}{dt} \int_{-\infty}^{\infty} U(P_1, -v') \exp(-j2\pi v' t) dv' = \int_{-\infty}^{\infty} -j2\pi v' U(P_1, -v') \exp(-j2\pi v' t) dv'$$

$$\begin{aligned} \frac{d}{dt} u(P_1, t) &= \frac{d}{dt} \int_{-\infty}^{\infty} U(P_1, -v') \exp(-j2\pi v' t) dv' \\ &= \int_{-\infty}^{\infty} -j2\pi v' U(P_1, -v') \exp(-j2\pi v' t) dv' \end{aligned}$$

can be used to write

$$u(P_0, t) = \int_{\Sigma} \int \cos(n, r_{01}) 2\pi v r_{01} \frac{d}{dt} u(P_1, t - \frac{r_{01}}{v}) ds.$$

$$u(P_0, t) = \int_{\Sigma} \int \frac{\cos(\vec{n}, \vec{r}_{01})}{2\pi v r_{01}} \frac{d}{dt} u(P_1, t - \frac{r_{01}}{v}) ds.$$

(3-61)

The wave disturbance at point P_0 is seen to be linearly proportional to the *time derivative* of the disturbance at each point P_1 on the aperture. Since it takes time r_{01}/v for the disturbance to propagate from P_1 to P_0 , the observed wave depends on the derivative of the incident wave at the “retarded” time $t - (r_{01}/v)$.

This more general treatment shows that an understanding of diffraction of monochromatic waves can be used directly to synthesize the results for much more general nonmonochromatic

waves. However, the monochromatic results are directly applicable themselves when the optical source has a sufficiently narrow spectrum. See [Prob. 3-6](#) for further elucidation of these points.

3.9 Diffraction at Boundaries

In the statement of the Huygens-Fresnel principle, we found it convenient to regard each point on the aperture as a new source of spherical waves. It was pointed out that such sources are merely mathematical conveniences and have no real physical significance. A more physical point of view, first qualitatively expressed by Thomas Young in 1802, is to regard the observed field as consisting of a superposition of the incident wave transmitted through the aperture unperturbed and a diffracted wave originating at the *rim* of the aperture. The possibility of a new wave originating in the material medium of the rim makes this interpretation a more physical one.

Young's qualitative arguments were given added impetus by Sommerfeld's rigorous electromagnetic solution of the problem of diffraction of a plane wave by a semi-infinite, perfectly conducting screen [326]. This rigorous solution showed that the field in the geometrical shadow of the screen has the form of a cylindrical wave originating on the rim of the screen. In the directly illuminated region behind the plane of the screen the field was found to be a superposition of this cylindrical wave with the directly transmitted wave.

The applicability of a boundary diffraction approach in more general diffraction problems was investigated by [Maggi](#) [238] and [Rubinowicz](#) [301], who showed that the Kirchhoff diffraction formula can indeed be manipulated to yield a form that is equivalent to Young's ideas. More recently, [Miyamoto and Wolf](#) [302] have extended the theory of boundary diffraction. For further discussion of these ideas, the reader should consult the references cited.

Another approach closely related to Young's ideas is the geometrical theory of diffraction developed by [Keller](#) [191]. In this treatment, the field behind a diffracting obstacle is found by the principles of geometrical optics, modified by the inclusion of "diffracted rays" that originate at certain points on the obstacle itself. New rays are assumed to be generated at edges, corners, tips, and surfaces of the obstacle. This theory can often be applied to calculate the fields diffracted by objects that are too complex to be treated by other methods.

3.10 The Angular Spectrum of Plane Waves

It is also possible to formulate scalar diffraction theory in a framework that closely resembles the theory of linear, invariant systems. As we shall see, if the complex field distribution of a monochromatic disturbance is Fourier-analyzed across any plane, the various spatial Fourier components can be identified as plane waves traveling in different directions away from that plane. The field amplitude at any other point (or across any other parallel plane) can be calculated by adding the contributions of these plane waves, taking due account of the phase shifts they have undergone during propagation. For a detailed treatment of this approach to diffraction theory, as well as its applications in the theory of radio-wave propagation, the reader is referred to the work of [Ratcliffe \[294\]](#).

3.10.1 The Angular Spectrum and Its Physical Interpretation

Suppose that, due to some unspecified system of monochromatic sources, a wave is incident on a transverse (x, y) plane traveling with a component of propagation in the positive z direction. Let the complex field across that $z=0$ plane be represented by $U(x, y, 0)$; our ultimate objective is to calculate the resulting field $U(x, y, z)$ that appears across a second, parallel plane a distance z to the right of the first plane.

Across the $z=0$ plane, the function U has a two-dimensional Fourier transform given by

$$A(f_X, f_Y; 0) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U(x, y, 0) \exp[-j2\pi(f_X x + f_Y y)] dx dy.$$

$$A(f_X, f_Y; 0) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U(x, y, 0) \exp[-j2\pi(f_X x + f_Y y)] dx dy. \quad (3-62)$$

As pointed out in [Chapter 2](#), the Fourier transform operation may be regarded as a decomposition of a complicated function into a collection of simpler complex-exponential functions. To emphasize this point of view, we write U as an inverse Fourier transform of its spectrum,

$$U(x, y, 0) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} A(f_X, f_Y; 0) \exp[j2\pi(f_X x + f_Y y)] df_X df_Y.$$

$$U(x, y, 0) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} A(f_X, f_Y; 0) \exp[j2\pi(f_X x + f_Y y)] df_X df_Y. \quad (3-63)$$

To give physical meaning to the functions in the integrand of the above integral, consider the form of a simple plane wave propagating with wave vector \vec{k} , where \vec{k} has

magnitude $2\pi/\lambda$ and has direction cosines (α, β, γ) , as illustrated in Fig. 3.9. Such a plane wave has a complex representation of the form

$$p(x, y, z; t) = \exp[j(\vec{k} \cdot \vec{r} - 2\pi\nu t)]$$

$$p(x, y, z; t) = \exp[j(\vec{k} \cdot \vec{r} - 2\pi\nu t)]$$

(3-64)

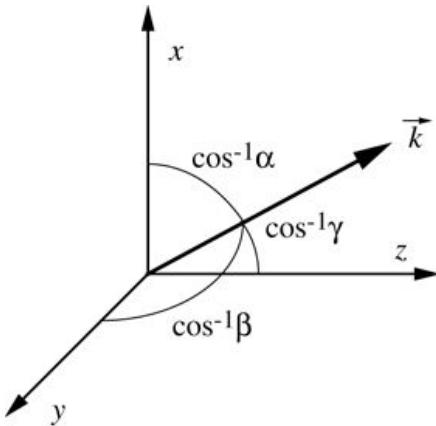


Figure 3.9

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 3.9 The wave vector \vec{k} .

The graph shows axes x, y, and z such that the z y plane is horizontal while the x z and x y planes are vertical. Vector k starts at the origin and is upward sloping in the x z plane such that it forms three angles: with the x axis, inverse cosine alpha; with the y axis, inverse cosine beta; and with the z axis, inverse cosine gamma.

where $\vec{r} = x\hat{x} + y\hat{y} + z\hat{z}$ is a position vector (the $\hat{}$ symbol signifies a unit vector), while $\vec{k} = \frac{2\pi}{\lambda}(\alpha\hat{x} + \beta\hat{y} + \gamma\hat{z})$. Dropping the time dependence, the complex phasor amplitude of the plane wave across a constant z -plane is

$$P(x, y, z) = \exp(jk \cdot \vec{r}) = \exp[j\frac{2\pi}{\lambda}(\alpha x + \beta y + \gamma z)].$$

$$P(x, y, z) = \exp\left(j\vec{k} \cdot \vec{r}\right) = \exp\left(j\frac{2\pi}{\lambda}(\alpha x + \beta y)\right) \exp\left(j\frac{2\pi}{\lambda}\gamma z\right).$$

(3-65)

Note that the direction cosines are interrelated through

$$\gamma^2 = 1 - \alpha^2 - \beta^2.$$

$$\gamma = \sqrt{1 - \alpha^2 - \beta^2}.$$

Thus across the plane $z=0$, a complex-exponential function $\exp[j2\pi(f_X x + f_Y y)]$ may be regarded as representing a plane wave propagating with direction cosines

$$\alpha = \lambda f_X \beta = \lambda f_Y \gamma = 1 - (\lambda f_X)^2 - (\lambda f_Y)^2.$$

$$\alpha = \lambda f_X \beta = \lambda f_Y \gamma = \sqrt{1 - (\lambda f_X)^2 - (\lambda f_Y)^2}.$$

(3-66)

In the Fourier decomposition of U , the complex amplitude of the plane-wave component with spatial frequencies (f_X, f_Y) is simply $A(f_X, f_Y; 0) df_X df_Y$, evaluated at $(f_X = \alpha/\lambda, f_Y = \beta/\lambda)$. For this reason, the function

$$A(\alpha/\lambda, \beta/\lambda; 0) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U(x, y, 0) \exp[-j2\pi(\frac{\alpha}{\lambda}x + \frac{\beta}{\lambda}y)] dx dy$$

(3-67)

is called the *angular spectrum* of the disturbance $U(x, y, 0)$.

3.10.2 Propagation of the Angular Spectrum

Consider now the angular spectrum of the disturbance U across a plane parallel to the (x, y) plane but at a distance z from it. Let the function $A(\alpha/\lambda, \beta/\lambda; z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U(x, y, z) \exp[-j2\pi(\frac{\alpha}{\lambda}x + \frac{\beta}{\lambda}y)] dx dy$ represent the angular spectrum of $U(x, y, z)$; that is,

$$A(\alpha/\lambda, \beta/\lambda; z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U(x, y, z) \exp[-j2\pi(\frac{\alpha}{\lambda}x + \frac{\beta}{\lambda}y)] dx dy.$$

$$A(\alpha/\lambda, \beta/\lambda; z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U(x, y, z) \exp[-j2\pi(\frac{\alpha}{\lambda}x + \frac{\beta}{\lambda}y)] dx dy.$$

(3-68)

Now if the relation between $A(\alpha/\lambda, \beta/\lambda; 0) = A(\alpha/\lambda, \beta/\lambda; z)$ and $A(\alpha/\lambda, \beta/\lambda; z)$ can be found, then the effects of wave propagation on the angular spectrum of the disturbance will be evident.

To find the desired relation, note that U can be written

$$U(x, y, z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} A(\alpha/\lambda, \beta/\lambda; z) \exp[j2\pi(\frac{\alpha}{\lambda}x + \frac{\beta}{\lambda}y)] d\alpha d\beta.$$

$$U(x, y, z) = \int_{-\infty}^{\infty} \int A\left(\frac{\alpha}{\lambda}, \frac{\beta}{\lambda}; z\right) \exp\left[j2\pi\left(\frac{\alpha}{\lambda}x + \frac{\beta}{\lambda}y\right)\right] d\frac{\alpha}{\lambda} d\frac{\beta}{\lambda}. \quad (3-69)$$

In addition, U must satisfy the Helmholtz equation,

$$\nabla^2 U + k^2 U = 0$$

$$\nabla^2 U + k^2 U = 0$$

at all source-free points. Direct application of this requirement to (3-69) shows that A must satisfy the differential equation

$$d^2 dz^2 A \alpha \lambda, \beta \lambda; z + 2\pi \lambda^2 1 - \alpha^2 - \beta^2 A \alpha \lambda, \beta \lambda; z = 0.$$

$$\frac{d^2}{dz^2} A\left(\frac{\alpha}{\lambda}, \frac{\beta}{\lambda}; z\right) + \left(\frac{2\pi}{\lambda}\right)^2 [1 - \alpha^2 - \beta^2] A\left(\frac{\alpha}{\lambda}, \frac{\beta}{\lambda}; z\right) = 0.$$

An elementary solution of this equation can be written in the form

$$A \alpha \lambda, \beta \lambda; z = A \alpha \lambda, \beta \lambda; 0 \exp j 2\pi \lambda 1 - \alpha^2 - \beta^2 z.$$

$$A\left(\frac{\alpha}{\lambda}, \frac{\beta}{\lambda}; z\right) = A\left(\frac{\alpha}{\lambda}, \frac{\beta}{\lambda}; 0\right) \exp\left(j\frac{2\pi}{\lambda} \sqrt{1 - \alpha^2 - \beta^2} z\right).$$

(3-70)

This result demonstrates that when the direction cosines (α, β) satisfy

$$\alpha^2 + \beta^2 < 1,$$

$$\alpha^2 + \beta^2 < 1,$$

(3-71)

as all true direction cosines must, the effect of propagation over distance z is simply a change of the relative phases of the various components of the angular spectrum. Since each plane-wave component propagates at a different angle, each travels a different distance between two parallel planes, and relative phase delays are thus introduced.

However, when (α, β) satisfy

$$\alpha^2 + \beta^2 > 1,$$

$$\alpha^2 + \beta^2 > 1,$$

a different interpretation is required. Note that since $A(\alpha/\lambda, \beta/\lambda; 0)$ is the Fourier transform of a field distribution on which boundary conditions are imposed in the aperture

plane, it is quite possible that this spectrum will contain components that violate (3-71). Under such a condition, α and β are no longer interpretable as direction cosines. Now the square root in (3-70) is imaginary, and that equation can be rewritten

$$A\alpha\lambda,\beta\lambda;z=A\alpha\lambda,\beta\lambda;0\exp(-\mu z)$$

$$A\left(\frac{\alpha}{\lambda}, \frac{\beta}{\lambda}; z\right) = A\left(\frac{\alpha}{\lambda}, \frac{\beta}{\lambda}; 0\right) \exp(-\mu z)$$

(3-72)

where

$$\mu=2\pi\lambda\alpha^2+\beta^2-1.$$

$$\mu = \frac{2\pi}{\lambda} \sqrt{\alpha^2 + \beta^2 - 1}.$$

Since μ is a positive real number, these wave components are rapidly attenuated by the propagation phenomenon. Such components are called *evanescent waves* and are quite analogous to the waves produced in a microwave waveguide driven below its cutoff frequency. As in the case of the waveguide driven below cutoff, these evanescent waves carry no energy away from the aperture.⁸

Finally, we note that the disturbance observed at (x, y, z) can be written in terms of the initial angular spectrum by inverse transforming (3-70), giving

$$U(x, y, z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} A\left(\frac{\alpha}{\lambda}, \frac{\beta}{\lambda}; 0\right) \exp\left(j\frac{2\pi}{\lambda} \sqrt{1 - \alpha^2 - \beta^2} z\right)$$

$$\times \exp\left[j2\pi\left(\frac{\alpha}{\lambda}x + \frac{\beta}{\lambda}y\right)\right] d\frac{\alpha}{\lambda} d\frac{\beta}{\lambda},$$

(3-73)

where the evanescent wave phenomenon will effectively limit the region of integration to the region within which Eq. (3-71) is satisfied⁹. Note that no angular spectrum components beyond the evanescent wave cutoff contribute to $U(x, y, z)$. This fact is the fundamental reason why no conventional imaging system can resolve a periodic structure with a period that is finer than wavelength of the radiation used. It is possible, though, to couple to evanescent wave components of the angular spectrum with very fine structures placed in very close proximity to the diffracting object, and thereby recover information that would otherwise be lost. However, we will focus here on conventional optical instruments, for which the evanescent waves are not recoverable.

3.10.3 Effects of a Diffracting Aperture on the Angular Spectrum

Suppose that an infinite opaque screen containing a diffracting structure is introduced in the plane $z=0$. We now consider the effects of that diffracting screen on the angular spectrum of the disturbance. Define the *amplitude transmittance function* of the aperture as the ratio of the transmitted field amplitude $U_t(x, y; 0)$ to the incident field amplitude $U_i(x, y; 0)$ at each (x, y) in the $z=0$ plane,

$$tA(x, y) = U_t(x, y; 0)/U_i(x, y; 0).$$

$$t_A(x, y) = \frac{U_t(x, y; 0)}{U_i(x, y; 0)}.$$

(3-74)

Then

$$U_t(x, y; 0) = U_i(x, y; 0)tA(x, y)$$

$$U_t(x, y, 0) = U_i(x, y, 0)t_A(x, y)$$

and the convolution theorem can be used to relate the angular spectrum $A_i(\alpha/\lambda, \beta/\lambda)$ of the incident field and the angular spectrum $A_t(\alpha/\lambda, \beta/\lambda)$ of the transmitted field,

$$At\alpha\lambda, \beta\lambda = A_i\alpha\lambda, \beta\lambda * T\alpha\lambda, \beta\lambda,$$

$$A_t\left(\frac{\alpha}{\lambda}, \frac{\beta}{\lambda}\right) = \left[A_i\left(\frac{\alpha}{\lambda}, \frac{\beta}{\lambda}\right) * T\left(\frac{\alpha}{\lambda}, \frac{\beta}{\lambda}\right) \right],$$

(3-75)

where

$$T\alpha\lambda, \beta\lambda = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} t_A(x, y) \exp(-j2\pi\alpha\lambda x + \beta\lambda y) dx dy,$$

$$T\left(\frac{\alpha}{\lambda}, \frac{\beta}{\lambda}\right) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} t_A(x, y) \exp\left[-j2\pi\left(\frac{\alpha}{\lambda}x + \frac{\beta}{\lambda}y\right)\right] dx dy,$$

(3-76)

and $*$ is again the symbol for convolution.

The angular spectrum of the transmitted disturbance is thus seen to be the convolution of the angular spectrum of the incident disturbance with a second angular spectrum that is characteristic of the diffracting structure.

For the case of a unit amplitude plane wave illuminating the diffracting structure normally, the result takes a particularly simple form. In that case

$$A_i\alpha\lambda, \beta\lambda = \delta\alpha\lambda, \beta\lambda$$

$$A_i\left(\frac{\alpha}{\lambda}, \frac{\beta}{\lambda}\right) = \delta\left(\frac{\alpha}{\lambda}, \frac{\beta}{\lambda}\right)$$

and

$$At\alpha\lambda, \beta\lambda = \delta\alpha\lambda, \beta\lambda * T\alpha\lambda, \beta\lambda = T\alpha\lambda, \beta\lambda.$$

$$A_t\left(\frac{\alpha}{\lambda}, \frac{\beta}{\lambda}\right) = \delta\left(\frac{\alpha}{\lambda}, \frac{\beta}{\lambda}\right) * T\left(\frac{\alpha}{\lambda}, \frac{\beta}{\lambda}\right) = T\left(\frac{\alpha}{\lambda}, \frac{\beta}{\lambda}\right).$$

Thus the transmitted angular spectrum is found directly by Fourier transforming the amplitude transmittance function of the aperture.

Note that, if the diffracting structure is an aperture that limits the extent of the field distribution, the result is a broadening of the angular spectrum of the disturbance, from the basic properties of Fourier transforms. The smaller the aperture, the broader the angular spectrum behind the aperture. This effect is entirely analogous to the broadening of the spectrum of an electrical signal as its duration is decreased.

3.10.4 The Propagation Phenomenon as a Linear Spatial Filter

Consider again the propagation of light from the plane $z=0$ to a parallel plane at nonzero distance z . The disturbance $U(x, y, 0)$ incident on the first plane may be considered to be mapped by the propagation phenomenon into a new field distribution $U(x, y, z)$. Such a mapping satisfies our previous definition of a system. We shall, in fact, demonstrate that the propagation phenomenon acts as a linear space-invariant system and is characterized by a relatively simple transfer function.

The linearity of the propagation phenomenon has already been discussed; it is directly implied by the linearity of the wave equation, or alternatively, by the superposition integral (3-55). The space-invariance property is most easily demonstrated by actually deriving a transfer function that describes the effects of propagation; if the mapping has a transfer function, then it must be space-invariant.

To find the transfer function, we return to the angular spectrum point of view. However, rather than writing the angular spectra as functions of the direction cosines (α, β) , it is now more convenient to leave the spectra as functions of spatial frequencies (f_X, f_Y) . The spatial frequencies and the direction cosines are related through (3-66).

Let the spatial Fourier spectrum of $U(x, y, z)$ again be represented by $A(f_X, f_Y; z)$ $A(f_X, f_Y; z)$, while the spectrum of $U(x, y, 0)$ is again written $A(f_X, f_Y; 0)$ $A(f_X, f_Y; 0)$. Thus we may express $U(x, y, z)$ as

$$U(x, y, z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} A(f_X, f_Y; z) \exp[j2\pi(f_X x + f_Y y)] df_X df_Y.$$

$$U(x, y, z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} A(f_X, f_Y; z) \exp[j2\pi(f_X x + f_Y y)] df_X df_Y.$$

But in addition, from (3-73),

$$U(x, y, z) = \iint_{-\infty}^{\infty} A(f_X, f_Y; 0) \times \exp[j2\pi\lambda 1 - (\lambda f_X)^2 - (\lambda f_Y)^2] z \exp[j2\pi f_X x + f_Y y] df_X df_Y.$$

$$\begin{aligned} U(x, y, z) &= \iint_{-\infty}^{\infty} A(f_X, f_Y; 0) \\ &\quad \times \exp\left[j\frac{2\pi}{\lambda}\sqrt{1 - (\lambda f_X)^2 - (\lambda f_Y)^2} z\right] \exp[j2\pi(f_X x + f_Y y)] df_X df_Y. \end{aligned}$$

A comparison of the above two equations shows that

$$A(f_X, f_Y; z) = A(f_X, f_Y; 0) \exp[j2\pi z \lambda 1 - (\lambda f_X)^2 - (\lambda f_Y)^2].$$

$$A(f_X, f_Y; z) = A(f_X, f_Y; 0) \exp\left[j2\pi\frac{z}{\lambda}\sqrt{1 - (\lambda f_X)^2 - (\lambda f_Y)^2}\right].$$

(3-77)

Finally, the transfer function of the wave propagation phenomenon is seen to be

$$H(f_X, f_Y) = \exp\left[j2\pi\frac{z}{\lambda}\sqrt{1 - (\lambda f_X)^2 - (\lambda f_Y)^2}\right].$$

(3-78)

To a good approximation in most problems, the propagation phenomenon may be regarded as a linear, dispersive spatial filter with a finite bandwidth. The transmission of the filter is essentially zero outside a circular region of radius λ^{-1} in the frequency plane. Within that circular bandwidth, the modulus of the transfer function is unity but frequency-dependent phase shifts are introduced. The phase dispersion of the system is most significant at high spatial frequencies and vanishes as both f_X and f_Y approach zero. In addition, for any fixed spatial frequency pair, the phase dispersion increases as the distance of propagation z increases.

In closing we mention the remarkable fact that, despite the apparent differences of their approaches, *the angular spectrum approach and the full first Rayleigh-Sommerfeld solution yield identical predictions of diffracted fields!* This has been proved most elegantly by [Sherman \[315\]](#), who recognized that the Fourier transform of the first Rayleigh-Sommerfeld impulse response (3-32) is the same as the transfer function of the angular spectrum approach, i.e.

$$\mathcal{F}h(x, y) = \mathcal{F}12\pi z r^1 r - jk \exp(jkr) r = \exp[j2\pi z \lambda 1 - (\lambda f_X)^2 - (\lambda f_Y)^2],$$

$$\begin{aligned} \mathcal{F}\{h(x, y)\} &= \mathcal{F}\left\{\frac{1}{2\pi} \frac{z}{r} \left(\frac{1}{r} - jk\right) \frac{\exp(jkr)}{r}\right\} \\ &= \exp\left[j2\pi\frac{z}{\lambda}\sqrt{1 - (\lambda f_X)^2 - (\lambda f_Y)^2}\right], \end{aligned}$$

(3-79)

$$\text{where } r = \sqrt{x^2 + y^2 + z^2}.$$

Problems - Chapter 3

1. 3-1. Show that in an isotropic, nonmagnetic, and inhomogeneous dielectric medium, Maxwell's equations can be combined to yield (3-8).
2. 3-2. Show that a diverging spherical wave satisfies the Sommerfeld radiation condition.
3. 3-3. Show that, if $r_{21} \gg \lambda$, (3-27) can be reduced to (3-28).
4. 3-4. Show that the normal derivative of (3-40) for G_+ vanishes across the screen and aperture.
5. 3-5. Assuming unit-amplitude normally incident plane-wave illumination, find the angular spectrum of
 1. a circular aperture of diameter d .
 2. a circular opaque disk of diameter d .
6. 3-6. Consider a real nonmonochromatic disturbance $u(P, t)$ of center frequency ν^- and bandwidth $\Delta\nu$. Let a related complex-valued disturbance $u_-(P, t)$ be defined as consisting of only the negative-frequency components of $u(P, t)$. Thus

$$u_-(P, t) = \int_{-\infty}^0 U(P, \nu) \exp(j2\pi\nu t) d\nu$$

$$u_-(P, t) = \int_{-\infty}^0 U(P, \nu) \exp(j2\pi\nu t) d\nu$$

where $U(P, \nu)$ is the Fourier spectrum of $u(P, t)$. Assuming the geometry of Fig. 3.6 show that if

$$\Delta\nu \ll 1 \text{ and } \nu \gg r_{01}$$

$$\frac{\Delta\nu}{\nu} \ll 1 \text{ and } \frac{1}{\Delta\nu} \gg \frac{nr_{01}}{\nu}$$

then

$$u_-(P_0, t) = j\lambda \int_{-\infty}^{\infty} u_-(P_1, t) \exp(jk^- r_{01}) r_{01} \cos(\vec{n} \cdot \vec{r}_{01}) ds$$

$$u_-(P_0, t) = \frac{1}{j\lambda} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u_-(P_1, t) \frac{\exp(jk^- r_{01})}{r_{01}} \cos(\vec{n} \cdot \vec{r}_{01}) ds$$

$\lambda = v \sqrt{\nu}$ and $k = 2\pi / \lambda$. In the above equations, n is the refractive index of the medium and v is the speed of propagation.

7.3-7. For a wave that travels only in directions that have small angles with respect to the optical axis, the general form of the phasor complex field may be approximated by

$$U(x, y, z) \approx V(x, y, z) \exp(jkz),$$

$$U(x, y, z) \approx V(x, y, z) \exp(jkz),$$

where $V(x, y, z)$ is a slowly varying function of z .

1. Show that for such a wave the Helmholtz equation can be reduced to

$$\nabla_t^2 V + j2k \frac{\partial V}{\partial z} = 0,$$

$$\nabla_t^2 V + j2k \frac{\partial V}{\partial z} = 0,$$

where $\nabla_t^2 = \partial^2 / \partial x^2 + \partial^2 / \partial y^2$ is the transverse portion of the Laplacian. This equation is known as the *paraxial Helmholtz equation*.

2. Show that an expanding spherical wave,

$$V(x, y, z) = V_1 z \exp(jkx^2 + y^2 / 2z)$$

is a solution to this equation.

4 Fresnel and Fraunhofer Diffraction

In the preceding chapter the results of scalar diffraction theory were presented in their most general forms. Attention is now turned to certain approximations to the general theory, approximations that will allow diffraction pattern calculations to be reduced to comparatively simple mathematical manipulations. These approximations, which are commonly made in many fields that deal with wave propagation, will be referred to as *Fresnel* and *Fraunhofer* approximations. In accordance with our view of the wave propagation phenomenon as a “system,” we shall attempt to find approximations that are valid for a wide class of “input” field distributions.

4.1 Background

In this section we prepare the reader for the calculations to follow. The concept of the *intensity* of a wave field is introduced, and the Huygens-Fresnel principle, from which the approximations are derived, is presented in a form that is especially well suited for approximation.

4.1.1 The Intensity of a Wave Field

In the optical region of the spectrum, a photodetector responds directly to the optical power falling on its surface. Thus for a semiconductor detector, if optical power \mathcal{P} is incident on the photosensitive region, absorption of a photon generates an electron in the conduction band and a hole in the valence band. Under the influence of internal and applied fields, the hole and electron move in opposite directions, leading to a photocurrent i that is the response to the incident absorbed photon. Under most circumstances the photocurrent is linearly proportional to the incident power,

$$i = \mathcal{R}P.$$

$$i = \mathcal{R}\mathcal{P}.$$

(4-1)

The proportionality constant \mathcal{R} is called the *responsivity* of the detector and is given by

$$\mathcal{R} = \eta q e h v,$$

$$\mathcal{R} = \frac{\eta q e q}{h v},$$

(4-2)

where $\eta q e$ is the *quantum efficiency* of the photodetector (the average number of electron-hole pairs released by the absorption of a photon, a quantity that is less than or equal to unity in the absence of internal gain), q is the electronic charge (1.602×10^{-19} coulombs), h is Planck's constant (6.626196×10^{-34} joule-second), and v is the optical frequency.¹

Thus in optics the directly measurable quantity is optical power, and it is important to relate such power to the complex scalar fields $u(P, t)$ and $U(P)$ dealt with in earlier discussions of diffraction theory. To understand this relation requires a return to an electromagnetic description of the problem. We omit the details here, referring the reader to [Sections 5.3 and 5.4 of \[305\]](#), and simply state the major points. Let the medium be isotropic, and the wave monochromatic. Assuming that the wave behaves *locally* as a transverse electromagnetic plane wave (i.e. $\mathcal{E} \rightarrow \vec{\mathcal{E}}$, $\mathcal{H} \rightarrow \vec{\mathcal{H}}$, and $\mathbf{k} \rightarrow \vec{k}$ form a mutually orthogonal triplet), then the electric and magnetic fields can be expressed locally as

$$\mathcal{E} \rightarrow = \operatorname{Re}\{E \rightarrow 0 \exp[-j(2\pi\nu t - k \cdot r)]\} \quad \mathcal{H} \rightarrow = \operatorname{Re}\{H \rightarrow 0 \exp[-j(2\pi\nu t - k \cdot r)]\},$$

$$\begin{aligned}\vec{\mathcal{E}} &= \operatorname{Re} \left\{ \vec{E}_0 \exp \left[-j(2\pi\nu t - \vec{k} \cdot \vec{r}) \right] \right\} \\ \vec{\mathcal{H}} &= \operatorname{Re} \left\{ \vec{H}_0 \exp \left[-j(2\pi\nu t - \vec{k} \cdot \vec{r}) \right] \right\},\end{aligned}\tag{4-3}$$

where $E \rightarrow 0 \vec{E}_0$ and $H \rightarrow 0 \vec{H}_0$ are locally constant and have complex components. The power flows in the direction of the vector $k \rightarrow \vec{k}$ and the power density can be expressed as

$$p = E \rightarrow 0 \cdot E \rightarrow 0^* = 2\eta = E_0 X^2 + E_0 Y^2 + E_0 Z^2,$$

$$p = \frac{\vec{E}_0 \cdot \vec{E}_0^*}{2\eta} = \frac{E_{0X}^2 + E_{0Y}^2 + E_{0Z}^2}{2\eta},$$

(4-4)

where η is the *characteristic impedance* of the medium and is given by

$$\eta = \mu\epsilon.$$

$$\eta = \sqrt{\frac{\mu}{\epsilon}}.$$

In vacuum, η is equal to 377Ω . The total power incident on a surface of area A is the integral of the power density over A , taking into account that the direction of power flow is in the direction of $k \rightarrow \vec{k}$,

$$\text{Power} = \int A p k \rightarrow \cdot n^\wedge |k \rightarrow| dx dy.$$

$$\mathcal{P} = \iint_A p \frac{\vec{k} \cdot \hat{n}}{|\vec{k}|} dx dy.$$

Here $n^\wedge \hat{n}$ is a unit vector pointing normally into the surface of the detector, while $k \rightarrow /|k \rightarrow| \vec{k} / |\vec{k}|$ is a unit vector in the direction of power flow. When $k \rightarrow \vec{k}$ is nearly normal to the surface, the total power is simply the integral of the power density p over the detector area.

The proportionality of power density to the squared magnitude of the $E \rightarrow 0 \vec{E}_0$ vector seen in (4-4) leads us to define the *intensity* of a scalar monochromatic wave at point P as the

squared magnitude of the complex phasor representation $U(P)$ of the disturbance,

$$I(P) = |U(P)|^2.$$

$$I(P) = |U(P)|^2.$$

(4-5)

Note that power density and intensity are not identical, but the latter quantity is directly proportional to the former. For this reason we regard the intensity as the physically measurable attribute of an optical wavefield.

When a wave is not perfectly monochromatic, but is narrow band, a straightforward generalization of the concept of intensity is given by

$$I(P) = \langle |u(P,t)|^2 \rangle,$$

$$I(P) = \langle |u(P,t)|^2 \rangle,$$

(4-6)

where the angle brackets signify an infinite time average. In some cases, the concept of *instantaneous intensity* is useful, defined as

$$I(P,t) = |u(P,t)|^2.$$

$$I(P,t) = |u(P,t)|^2.$$

(4-7)

When calculating a diffraction pattern, we will generally regard the intensity of the pattern as the quantity we are seeking.

4.1.2 The Huygens-Fresnel Principle in Rectangular Coordinates

Before introducing a series of approximations to the Huygens-Fresnel principle, it will be helpful to first state the principle in more explicit form for the case of rectangular coordinates. As shown in [Fig. 4.1](#), the diffracting aperture is assumed to lie in the (ξ, η) plane, and is illuminated in the positive z direction. We will calculate the wavefield across the (x, y) plane, which is parallel to the (ξ, η) plane and at normal distance z from it. The z axis pierces both planes at their origins. According to [\(3-42\)](#), the Huygens-Fresnel principle can be stated (when $r_{01} \gg \lambda$) as

$$U(P_0) = j\lambda \int U(P_1) \exp(jkr_{01}) r_{01} \cos\theta ds,$$

$$U(P_0) = \frac{1}{j\lambda} \int_{\Sigma} \int U(P_1) \frac{\exp(jkr_{01})}{r_{01}} \cos\theta ds,$$

(4-8)

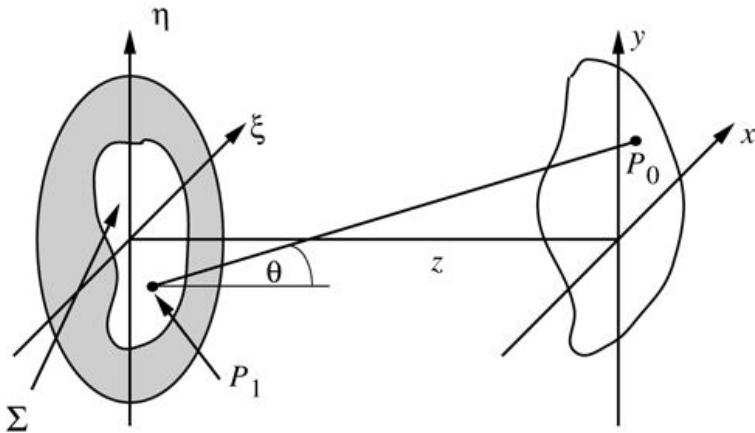


Figure 4.1

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 4.1 Diffraction geometry.

The illustration shows two 3 dimensional planes with a common horizontal axis z ; the one on the left has vertical axis η and third axis ξ , and the one on the right has vertical axis y and third axis x . The origin of the η , ξ , z plane is near the center of aperture Σ . A line segment connects point P_1 below the origin in the ξ η plane and point P_0 above the origin in the x y plane. The line makes angle θ with the horizontal line that runs rightward from P_1 .

where θ is the angle between the outward normal \hat{n} , which points in the negative z direction, and the vector $r_{01} \vec{r}_{01}$, which points from P_0 to P_1 . The term $\cos\theta \cos\theta$ is given exactly by

$$\cos\theta = z r_{01},$$

$$\cos\theta = \frac{z}{r_{01}},$$

and therefore the Huygens-Fresnel principle can be rewritten

$$U(x, y) = z j \lambda \sum \int U(\xi, \eta) \exp(jkr_{01}) r_{01}^2 d\xi d\eta,$$

$$U(x, y) = \frac{z}{j\lambda} \int \int U(\xi, \eta) \frac{\exp(jkr_{01})}{r_{01}^2} d\xi d\eta,$$

(4-9)

where the distance r_{01} is given exactly by

$$r_{01} = \sqrt{z^2 + (x - \xi)^2 + (y - \eta)^2}.$$

$$r_{01} = \sqrt{z^2 + (x - \xi)^2 + (y - \eta)^2}. \quad (4-10)$$

There have been only two approximations in reaching this expression. One is the approximation inherent in the scalar theory. The second is the assumption that the observation distance is many wavelengths from the aperture, $r_{01} \gg \lambda$. We now embark on a series of additional approximations.

4.2 The Fresnel Approximation

To reduce the Huygens-Fresnel principle to a more simple and usable expression, we introduce approximations for the distance r_{01} between P_1 and P_0 . The approximations are based on the binomial expansion of the square root in (4-10). Let b be a number that is less than unity, and consider the expression $1+b\sqrt{1+b}$. The binomial expansion of the square root is given by

$$1+b=1+12b-18b^2+\dots,$$

$$\sqrt{1+b}=1+\frac{1}{2}b-\frac{1}{8}b^2+\dots,$$

(4-11)

where the number of terms needed for a given accuracy depends on the magnitude of b .

To apply the binomial expansion to the problem at hand, factor a $z^{\frac{1}{2}}$ outside the expression for r_{01} , yielding

$$r_{01}=z^{1/2}(x-\xi)^2+(y-\eta)^2.$$

$$r_{01}=z\sqrt{1+\left(\frac{x-\xi}{z}\right)^2+\left(\frac{y-\eta}{z}\right)^2}.$$

(4-12)

Let the quantity b in (4-11) consist of the sum of the second and third terms under the square root in (4-12). Then, retaining only the first two terms of the expansion (4-11), we have

$$r_{01}\approx z^{1/2}(x-\xi)^2+(y-\eta)^2.$$

$$r_{01}\approx z\left[1+\frac{1}{2}\left(\frac{x-\xi}{z}\right)^2+\frac{1}{2}\left(\frac{y-\eta}{z}\right)^2\right].$$

(4-13)

The question now arises as to whether we need to retain all the terms in the approximation (4-13), or whether only the first term might suffice. The answer to this question depends on which of the several occurrences of r_{01} is being approximated. For the r_{01}^{-2} appearing in the denominator of (4-9), the error introduced by dropping all terms but $z^{\frac{1}{2}}$ is generally acceptably small. However, for the r_{01}^{-k} appearing in the exponent, errors are much more critical. First, they are multiplied by a very large number k , a typical value for which might be greater than 10^7 m^{-1} in the visible region of the spectrum (e.g. $\lambda=5\times 10^{-7} \text{ m}$). Second, phase changes of as little as a fraction of a radian can change the value of the exponential

significantly. For this reason we retain both terms of the binomial approximation in the exponent. The resulting expression for the field at (x, y) therefore becomes

$$U(x, y, z) = e^{jkz} \int_{-\infty}^{\infty} \int U(\xi, \eta, 0) \exp[jk2zx - \xi^2 + y - \eta^2] d\xi d\eta,$$

$$U(x, y, z) = \frac{e^{jkz}}{j\lambda z} \int_{-\infty}^{\infty} \int U(\xi, \eta, 0) \exp \left\{ j \frac{k}{2z} [(x - \xi)^2 + (y - \eta)^2] \right\} d\xi d\eta,$$

(4-14)

where we have incorporated the finite limits of the aperture in the definition of $U(\xi, \eta, 0)$, in accord with the usual assumed boundary conditions.

Equation (4-14) is readily seen to be a convolution, expressible in the form

$$U(x, y, z) = \int_{-\infty}^{\infty} \int U(\xi, \eta, 0) h(x - \xi, y - \eta) d\xi d\eta$$

$$U(x, y, z) = \int_{-\infty}^{\infty} \int U(\xi, \eta, 0) h(x - \xi, y - \eta) d\xi d\eta$$

(4-15)

where the convolution kernel is

$$h(x, y) = e^{jkz} \int_{-\infty}^{\infty} \int U(\xi, \eta, 0) e^{-j \frac{k}{2z} (\xi^2 + \eta^2)} d\xi d\eta.$$

$$h(x, y) = \frac{e^{jkz}}{j\lambda z} \exp \left[j \frac{k}{2z} (x^2 + y^2) \right].$$

(4-16)

We will return to this viewpoint a bit later.

Another form of the result (4-14) is found if the term $\exp[jk2zx + y^2 \exp[j \frac{k}{2z} (x^2 + y^2)]]$ is factored outside the integral signs, yielding

$$U(x, y, z) = e^{jkz} \int_{-\infty}^{\infty} \int U(\xi, \eta, 0) e^{jk2z(\xi^2 + \eta^2)} e^{-j2\pi\lambda z(x\xi + y\eta)} d\xi d\eta,$$

$$U(x, y, z) = \frac{e^{jkz}}{j\lambda z} e^{j \frac{k}{2z} (x^2 + y^2)} \int_{-\infty}^{\infty} \int U(\xi, \eta, 0) e^{j \frac{k}{2z} (\xi^2 + \eta^2)} \\ \times e^{-j \frac{2\pi}{\lambda z} (x\xi + y\eta)} d\xi d\eta,$$

(4-17)

which we recognize (aside from multiplicative factors) to be the *Fourier transform* of the product of the complex field just to the right of the aperture and a quadratic-phase exponential.

We refer to both forms of the result, (4-14) and (4-17), as the *Fresnel diffraction integral*. When this approximation is valid, the observer is said to be in the region of Fresnel diffraction.²

4.2.1 Positive vs. Negative Phases

We have seen that it is common practice when using the Fresnel approximation to replace expressions for spherical waves by quadratic-phase exponentials. The question often arises as to whether the sign of the phase should be positive or negative in a given expression. This question is not only pertinent to quadratic-phase exponentials, but also arises when considering exact expressions for spherical waves and when considering plane waves propagating at an angle with respect to the optical axis. We now present the reader with a methodology that will help determine the proper sign of the exponent in all of these cases.

The critical fact to keep in mind is that we have chosen our phasors to rotate in the *clockwise* direction, i.e. their time dependence is of the form $\exp(-j2\pi\nu t)$. For this reason, if we move in space in such a way as to intercept portions of a wavefield that were emitted *later* in time, the phasor will have advanced in the clockwise direction, and therefore the phase must become more *negative*. On the other hand, if we move in space to intercept portions of a wavefield that were emitted *earlier* in time, the phasor will not have had time to rotate as far in the clockwise direction, and therefore the phase must become more *positive*.

If we imagine observing a spherical wave that is diverging from a point on the z^z axis, the observation being in an (x, y) plane that is normal to that axis, then movement away from the origin always results in observation of portions of the wavefront that were emitted earlier in time than that at the origin, since the wave has had to propagate further to reach those points. For that reason the phase must increase in a positive sense as we move away from the origin.

Therefore the expressions $\exp(jkr_{01})$ and $\exp[jk2z(x^2+y^2)]$ represent a diverging spherical wave and a quadratic-phase approximation to such a wave, respectively. By the same token, $\exp(-jkr_{01})$ and $\exp[-jk2z(x^2+y^2)]$ $\exp\left[-j\frac{k}{2z}(x^2+y^2)\right]$ represent a converging spherical wave, again assuming that z^z is positive. Clearly, if z^z is a negative number, then the interpretation must be reversed, since a negative sign is hidden in z^z .

Similar reasoning applies to the expressions for plane waves traveling at an angle with respect to the optical axis. Thus for positive α^α , the expression $\exp(j2\pi\alpha y)$ represents a plane wave with a wave vector in the (y, z) plane. But does the wave vector point with a positive angle with respect to the z^z axis or with a negative angle, keeping in mind that a positive angle is one that has rotated counterclockwise with respect to the z^z axis? If we move in the positive y^y direction, the argument of the exponential increases in a positive sense, and therefore we are moving to a portion of the wave that was emitted earlier in time. This can only be true if the wave vector points with a positive angle with respect to the z^z axis, as illustrated in [Fig. 4.2](#).

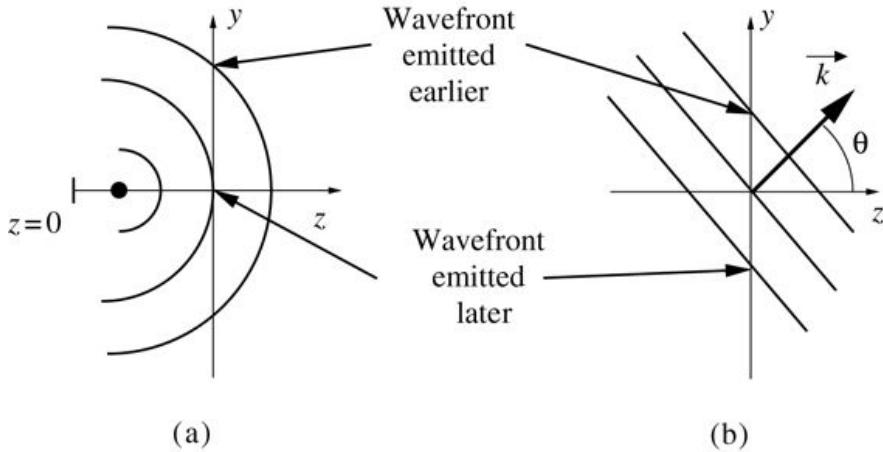


Figure 4.2

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 4.2 Determining the sign of the phases of exponential representations of (a) spherical waves and (b) plane waves.

Graph a shows the z - y plane with three concentric semicircles representing wave fronts. At the origin, the vertical y -axis is tangential to the wave front in the middle labeled "Wave front emitted later". The wave front to the right, that intersects the y -axes at two places, is labeled "Wave front emitted earlier." Graph b also shows the z - y plane with three downward-sloping parallel lines, the one in the middle passing through the origin. The line to its right is labeled "Wave front emitted earlier" and the line to its left is labeled "Wave front emitted later." Beginning at the origin, vector k is upward sloping at an angle θ with the z -axis.

4.2.2 Accuracy of the Fresnel Approximation

Considering the approximation in the exponent, which is the most critical approximation, it can be seen that the *spherical* secondary wavelets of the Huygens-Fresnel principle have been replaced by wavelets with quadratic-phase wavefronts. The accuracy of this approximation is determined by the errors induced when terms higher than first order (linear in b) are dropped in the binomial expansion (4-11). A sufficient condition for accuracy would be that the maximum phase change induced by dropping the $b^2/8$ term be much less than 1 radian. This condition will be met if the distance z satisfies

$$z \gg \pi 4\lambda [(x-\xi)^2 + (y-\eta)^2]_{\max}^{1/2}$$

$$z^3 \gg \frac{\pi}{4\lambda} [(x - \xi)^2 + (y - \eta)^2]_{\max}^2$$

(4-18)

For a circular aperture of size 1 cm, a circular observation region of size 1 cm, and a wavelength of $0.5 \mu m$, this condition would indicate that the distance z must be $\gg 25$ cm for accuracy. However, as the next comment will explain, this sufficient condition is often overly stringent, and accuracy can be expected for much shorter distances.

For the Fresnel approximation to yield accurate results, it is not necessary that the higher-order terms of the expansion be small, only that they not change the value of the Fresnel diffraction integral significantly. Considering the convolution form of the result, (4-14), if the major contribution to the integral comes from points (ξ, η) for which $\xi \approx x$ and $\eta \approx y$, then the particular values of the higher-order terms of the expansion are unimportant.

4.2.3 Finite Integral of the Quadratic-Phase Exponential Function

When calculating the Fresnel diffraction pattern of a finite slit aperture, an integral over finite symmetrical limits of the quadratic-phase exponential function arises. [Figure 4.3](#) shows the magnitude of the integral of a quadratic-phase exponential function,

$$\int_{-X}^X \exp(j\pi x^2) dx = 2C(2X) + j2S(2X)$$

$$\left| \int_{-X}^X \exp(j\pi x^2) dx \right| = |\sqrt{2}C(\sqrt{2}X) + j\sqrt{2}S(\sqrt{2}X)|$$

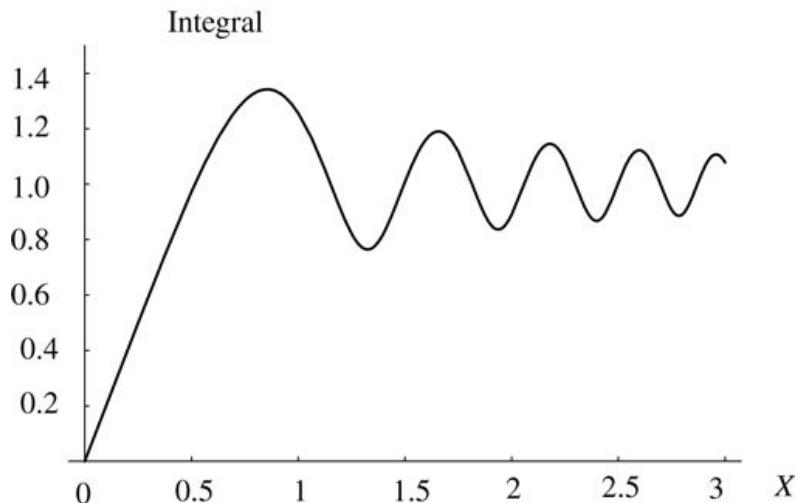


Figure 4.3

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 4.3 Magnitude of the integral of the quadratic-phase exponential function.

The graph plotting 0 to 3 in increments of 0.5 on the horizontal axis and 0 to 1.4 in increments of 0.2 on the vertical axis shows an upward sloping curve beginning at the origin and reaching up to a point marking (0.85, 1.35). The curve then drops and rises alternately in a wavelike pattern of decreasing amplitude.

which has also been expressed in terms of the Fresnel integrals $C(z)$ and $S(z)$ mentioned in [Section 2.2.1](#). As can be seen from the figure, the integral grows toward its asymptotic value of unity with increasing X . Note in particular that the integral first reaches unity when $X=0.5$, and then oscillates about that value with diminishing fluctuations. We conclude that, to a reasonable approximation, the major contributions to a convolution of this function with a second function that is smooth and slowly varying will come from the range

$-2 < X < 2$, due to the fact that outside this range the rapid oscillations of the integrand do not yield a significant addition to the total area.

For the scaled quadratic-phase exponential of (4-14) and (4-16), the corresponding conclusion is that the majority of the contribution to the convolution integral comes from a square in the (ξ, η) plane, with width $4\lambda z$ and centered on the point $(\xi=x, \eta=y)$ ($\xi = x, \eta = y$). This square grows in size as the distance z behind the aperture increases. In effect, when this square lies entirely within the open portion of the aperture, the field observed at distance z is, to a good approximation, what it would be if the aperture were not present. When the square lies entirely behind the obstruction of the aperture, then the observation point lies in a region that is, to a good approximation, dark due to the shadow of the aperture. When the square bridges the open and obstructed parts of the aperture, then the observed field is in the transition region between light and dark. The detailed structure within these regions may be complex, but the general conclusions above are correct. Figure 4.4 illustrates the various regions mentioned. For the case of a one-dimensional rectangular slit, the boundaries between the light region and the transition region, and between the dark region and the transition region, can be shown to be parabolas (see Prob. 4-6). However, the changes between regions do not happen abruptly, as the figure might imply, but rather are more gradual transitions.

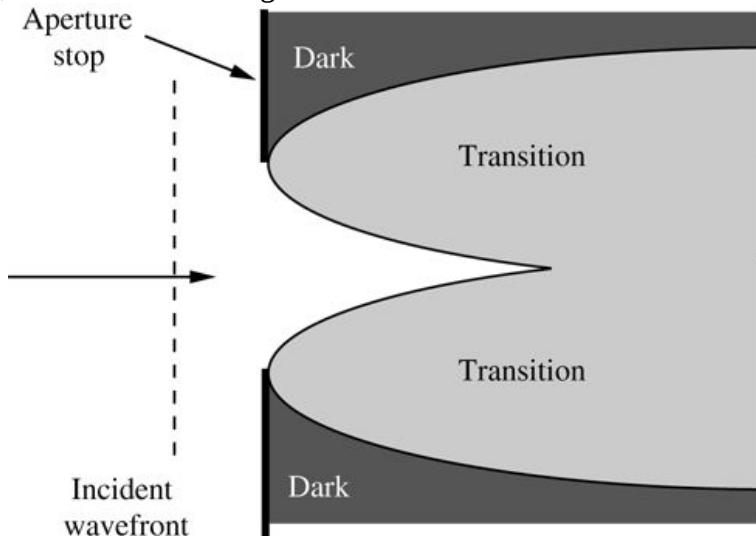


Figure 4.4
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 4.4 Light, dark, and transition regions behind a rectangular slit aperture.

To the left of the screen is a dotted vertical line representing the incident wave front. A horizontal line is directed at the center of the aperture between the aperture stops. On the other side there are two identical parabolas opening to the right such that the vertex of one is on the upper end of the aperture while that of the other is on the lower end of the aperture. There is a slight overlap of the parabolas such that the light entering the aperture progressively narrows to form a horizontal thorn-like shape. The parabolas are gray in color. Below the lower parabola and above the upper parabola it is dark.

Note that if the amplitude transmittance and / or the illumination of the diffracting aperture is not a relatively smooth and slowly varying function, the above conclusions may not hold. For example, if the amplitude of the field transmitted by the aperture has a high-spatial-frequency

sinusoidal component, that component may interact with the high frequencies of the quadratic-phase exponential kernel to produce a nonzero contribution from a location other than the square mentioned above. Thus the restriction of attention to the square of width $4\lambda z$ must be used with some caution. However, the idea is valid when the diffracting apertures do not contain fine structure and when they are illuminated by uniform plane waves.

If the distance z is allowed to approach zero, i.e. the observation point approaches the diffracting aperture, then the two-dimensional quadratic-phase function behaves in the limit like a delta function, producing a field $U(x, y)$ that is identical to the field $U(\xi, \eta)$ in the aperture. In such a case, the predictions of geometrical optics are valid, for such a treatment would predict that the field observed behind the aperture is simply a geometrical projection of the aperture fields onto the plane of observation.

Our discussion above is closely related to the *principle of stationary phase*, a method for finding the asymptotic values of certain integrals. A good discussion of this method can be found in Appendix III of [34]. For other examinations of the accuracy of the Fresnel approximation, see Chapter 9 of [276] and also [329]. The general conclusions of all of these analyses are similar, namely, the accuracy of the Fresnel approximation is usually extremely good to distances that are very close to the aperture.

4.2.4 The Fresnel Approximation and the Angular Spectrum

It is of some interest to understand the implications of the Fresnel approximations from the point of view of the angular spectrum method of analysis. Such understanding can be developed by beginning with (3-78), which expresses the transfer function of propagation through free space,

$$H(f_X, f_Y) = \exp[j2\pi z \lambda^{-1} - (\lambda f_X)^2 - (\lambda f_Y)^2].$$

(4-19)

This result, which is valid subject only to the scalar approximation, can now be compared with the transfer function predicted by the results of the Fresnel analysis. Fourier transforming the Fresnel diffraction impulse response (4-16), we find (with the help of Table 2.1) a transfer function valid for Fresnel diffraction,

$$\begin{aligned} H(f_X, f_Y) &= \mathcal{F}\left\{\frac{e^{jkz}}{j\lambda z} \exp\left[j\frac{\pi}{\lambda z}(x^2 + y^2)\right]\right\} \\ &= e^{jkz} \exp[-j\pi\lambda z(f_X^2 + f_Y^2)]. \end{aligned}$$

(4-20)

Thus in the Fresnel approximation, the general spatial phase dispersion representing propagation is reduced to a *quadratic-phase dispersion*. The factor e^{jkz} on the right of this equation represents a constant phase delay suffered by all plane-wave components traveling between two

parallel planes separated by normal distance z . The second term represents the different phase delays suffered by plane-wave components traveling in different directions.

The expression (4-20) is clearly an approximation to the more general transfer function (4-19). We can obtain the approximate result from the general result by applying a binomial expansion to the exponent of (4-19),

$$1 - \lambda f_X^2 - \lambda f_Y^2 \approx 1 - \lambda f_X^2 - \lambda f_Y^2,$$

$$\sqrt{1 - (\lambda f_X)^2 - (\lambda f_Y)^2} \approx 1 - \frac{(\lambda f_X)^2}{2} - \frac{(\lambda f_Y)^2}{2},$$

(4-21)

which is valid provided $|\lambda f_X| \ll 1$ and $|\lambda f_Y| \ll 1$. Such restrictions on f_X and f_Y are simply restrictions to *small angles* of propagation. In particular, if θ represents the angle of the \vec{k} vector of the plane wave component with frequency pair (f_X, f_Y) with respect to the z axis, then the condition that the third term in the binomial expansion

$$2\pi z \lambda^2 \theta^2 \approx 2\pi z \lambda^2 \theta^2 + \frac{\theta^4}{4}$$

$$2\pi \frac{z}{\lambda} \sqrt{1 - \theta^2} \approx 2\pi \frac{z}{\lambda} \left(1 - \frac{\theta^2}{2} + \frac{\theta^4}{8} \right)$$

(4-22)

be much less than π is

$$\theta^2 \ll 1.$$

$$\frac{\theta^4 z}{4\lambda} \ll 1.$$

(4-23)

This condition is generally more satisfying than (4-18) since it deals directly with diffraction angles rather than restrictions on spatial extents and z .

So we see that, from the perspective of the angular spectrum, the Fresnel approximation is accurate provided only small angles of diffraction are involved, in particular angles that satisfy (4-23). It is for this reason that we often say that the Fresnel approximation and the *paraxial* approximation are equivalent.

4.2.5 Fresnel Diffraction Between Confocal Spherical Surfaces

Until now, attention has been focused on diffraction between two *planes*. An alternative geometry, of more theoretical than practical interest but nonetheless quite instructive, is diffraction between two confocal spherical surfaces (see, for example, [30], [31]). As shown in Fig. 4.5, two spheres are said to be confocal if the center of each lies on the surface of the other. In our case, the two

spheres are tangent to the planes previously used, with the points of tangency being the points where the z -axis pierces those planes. The distance r_{01} in our previous diffraction analysis is now the distance between the two spherical caps shown.

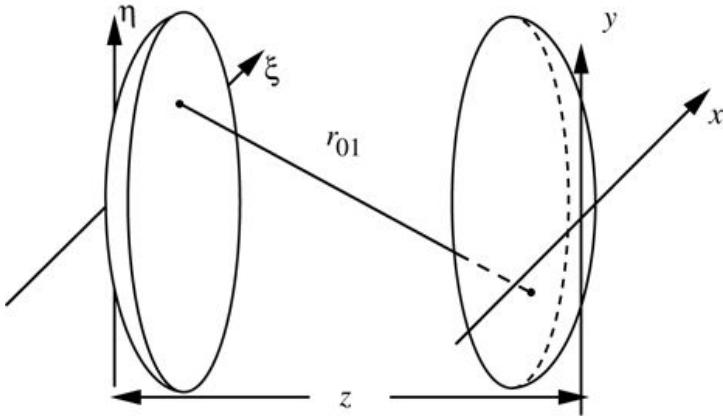


Figure 4.5

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 4.5 Confocal spherical surfaces.

The illustration shows two vertically placed shallow spherical discs facing each other along the horizontal axis z , the one on the left is along the vertical axis η and third axis ξ and the one on the right is along the vertical axis y and third axis x . A downward sloping straight line r_{01} connects the inner surfaces of the discs, from the upper half of the left disc to the lower half of the right disc.

A proper analysis would write equations for the left-hand spherical surface and for the right-hand spherical surface, and then use those equations to find the distance r_{01} between the two spherical caps. In the process it would be helpful to simplify certain square roots by using the first two terms of their binomial expansions (i.e. to make *paraxial* approximations to the spherical surfaces). The result of such an analysis is the following simple expression for r_{01} , valid if the extent of the spherical caps about the z -axis is small compared with their radii:

$$r_{01} \approx z - x\xi / z - y\eta / z.$$

$$r_{01} \approx z - x\xi / z - y\eta / z.$$

The Fresnel diffraction equation now becomes

$$U(x, y, z) = e^{jkz} j \lambda z \int_{-\infty}^{\infty} \int U(\xi, \eta, 0) e^{-j2\pi z(x\xi + y\eta)} d\xi d\eta,$$

$$U(x, y, z) = \frac{e^{jkz}}{j\lambda z} \int_{-\infty}^{\infty} \int U(\xi, \eta, 0) e^{-j\frac{2\pi}{\lambda z}(x\xi + y\eta)} d\xi d\eta,$$

(4-24)

which, aside from constant multipliers and scale factors, expresses the field observed on the right-hand spherical cap as the *Fourier transform* of the field on the left-hand spherical cap.

Comparison of this result with the previous Fourier-transform version of the Fresnel diffraction integral, (4-17), shows that the quadratic-phase factors in (x, y) and (ξ, η) have been eliminated by moving from the two planes to the two spherical caps. The two quadratic-phase factors in the earlier expression are in fact simply paraxial representations of spherical phase surfaces, and it is therefore reasonable that moving to the spheres has eliminated them.

One subtle point worth mention is that, when we analyze diffraction between two spherical caps, it is not really valid to use the Rayleigh-Sommerfeld result as the basis for the calculation, for that result was explicitly valid only for diffraction by a planar aperture. However, the Kirchhoff analysis remains valid, and its predictions are the same as those of the Rayleigh-Sommerfeld approach provided paraxial conditions hold.

4.2.6 Fresnel Diffraction in Terms of Ray Transfer Matrices

In [Appendix B](#), the geometrical-optics concept of a ray transfer matrix is introduced. In this section, we wish to generalize the Huygens-Fresnel principle for light propagating through a complex optical system described by an ABCD $ABCD$ matrix, with the system possibly containing many elements. For alternative discussions of this material, see [76] and [316]. To summarize the relations we need here, we consider only meridional rays, i.e. rays confined to the (y, z) plane, where z is defined along the optical axis. In addition, we consider only paraxial rays. With reference to [Fig. 4.6](#), consider an arbitrary linear optical system, but with no internal aperture stops. Let the system be illuminated by an expanding spherical wave with radius of curvature R_1 , producing a converging spherical wave at the output with radius of curvature R_2 . Let y_1 represent the y -coordinate where an input ray enters the optical system, and let $\hat{\theta}_1$ represent the “reduced angle” of that ray at the input plane, where the reduced angle is simply the angle divided by the refractive index to the left of the input plane. Likewise, let y_2 represent the y -coordinate where the ray exits the optical system, and let $\hat{\theta}_2$ represent the reduced angle with which the ray exits. Then the relation between the quantities defined in the input plane and those in the output plane can be written

$$y_2 \hat{\theta}_2 = ABCD y_1 \hat{\theta}_1,$$

$$\begin{bmatrix} y_2 \\ \hat{\theta}_2 \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} y_1 \\ \hat{\theta}_1 \end{bmatrix},$$

(4-25)

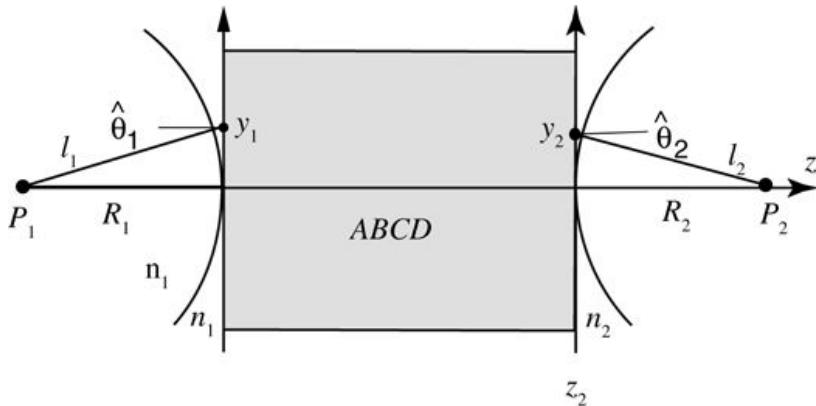


Figure 4.6

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 4.6 Input and output of an optical system with an arbitrary $ABCD$ matrix.

The illustration shows a rightward horizontal axis z extending from P_1 on the left extreme to P_2 in the right extreme. Axis z is intersected by two vertical axes, z_1 on the left and z_2 on the right, separated by a square matrix $A\ B\ C\ D$. Space to the left of z_1 is labeled n_1 and that to the right of z_2 is labeled n_2 . An arc centered at P_1 and of radius R_1 is tangential to z_1 and lies outside $ABCD$ such that the arc's point of intersection with z_1 is on axis z . A line segment measuring l_1 extends from P_1 to y_1 located on z_1 above the z axis. This line segment makes angle θ_1 with the horizontal line from y_1 to the left. Similarly, an arc centered at P_2 and of radius R_2 is tangential to z_2 and lies outside $ABCD$ such that the arc's point of intersection with z_2 is on axis z . A line segment measuring l_2 extends from P_2 to y_2 located on z_2 above the z axis. This line segment makes angle θ_2 with the horizontal line from y_2 to the right.

where the matrix relating the two vectors is called the ray-transfer matrix, or the $ABCD$ matrix of the system.

Fermat's principle implies that for a perfect imaging system, optical paths from P_1 to P_2 must all be equal. Therefore we can calculate that total optical path length L_{tot} by considering only the path down the z -axis of the system. That path length is given by

$$L_{\text{tot}} = n_1 R_1 - n_2 R_2 + \sum_i n_i \Delta z_i,$$

$$L_{\text{tot}} = n_1 R_1 - n_2 R_2 + \sum_i n_i \Delta z_i,$$

(4-26)

where the sum over Δz_i represents the sum of the individual path segments along the z -axis as the light passes to and through the various optical components contained between z_1 and z_2 . The convention that the radius of curvature is positive for a diverging spherical wave and negative for a converging spherical wave has been used. Now to calculate the optical path between the point y_1 and the point y_2 , we must subtract from L_{tot} the product n_1 times the

distance l_1 from P_1 to y_1 and n_2 times the distance l_2 from y_2 to P_2 . Using paraxial approximations we have

$$l_1 \approx n_1 R_1 + n_1 y_1^2 / 2R_1$$

$$l_2 \approx -n_2 R_2 - n_2 y_2^2 / 2R_2.$$

(4-27)

Thus the optical path length from y_1 to y_2 is given by

$$L_{12} = \sum_i n_i \Delta z_i - n_1 y_1^2 / 2R_1 - n_2 y_2^2 / 2R_2.$$

$$L_{12} = \sum_i n_i \Delta z_i - \frac{n_1 y_1^2}{2R_1} - \frac{n_2 y_2^2}{2R_2}.$$

(4-28)

Returning to the ray-transfer equation (4-25), we note that

$$y_2 = Ay_1 + B\theta^1, \quad \theta^2 = Cy_1 + D\theta^1.$$

$$\begin{aligned} y_2 &= Ay_1 + B\hat{\theta}_1 \\ \hat{\theta}_2 &= Cy_1 + D\hat{\theta}_1. \end{aligned}$$

(4-29)

We can solve this pair of equations for θ^2 and θ^1 , expressing the results in terms of y_1 , y_2 and the ray-transfer matrix elements:

$$\theta^1 = y_2 - Ay_1, \quad \theta^2 = Cy_1 + D\theta^1 = (BC - DA)y_1 + Dy_2.$$

$$\begin{aligned} \hat{\theta}_1 &= \frac{y_2 - Ay_1}{B} \\ \hat{\theta}_2 &= Cy_1 + D\hat{\theta}_1 = \frac{(BC - DA)y_1 + Dy_2}{B} = \frac{Dy_2 - y_1}{B}, \end{aligned}$$

(4-30)

where we have used the fact that the determinant of the ray-transfer matrix, $DA - BC$ is always unity ([316], p. 584). Now for small angles, replacing $\tan \theta^1$ by $\hat{\theta}_1$, we use the above results to conclude that

$$R1n1=y1\theta^1=By1y2-Ay1R2n2=y2\theta^2=By2Dy2-y1.$$

$$\begin{aligned}\frac{R_1}{n_1} &= \frac{y_1}{\hat{\theta}_1} = \frac{By_1}{y_2 - Ay_1} \\ \frac{R_2}{n_2} &= \frac{y_2}{\hat{\theta}_2} = \frac{By_2}{Dy_2 - y_1}.\end{aligned}$$

(4-31)

Substituting these results into (4-28), it follows that the total path length from y_1 to y_2 can be written

$$L_{12}=L_0-12BAy_{12}-2y_1y_2+Dy_{22},$$

$$L_{12}=L_0-\frac{1}{2B}[Ay_1^2-2y_1y_2+Dy_2^2],$$

(4-32)

$$L_0=\sum_i n_i \Delta z_i$$

where $L_0=\sum_i n_i \Delta z_i$. Now recall that the paraxial Huygens-Fresnel principle for propagation over a distance z in free space is given by (4-16), while in one dimension it can be written

$$U(y_2)=\int_{-\infty}^{\infty} U(y_1)h(y_2, y_1) dy_1$$

$$U(y_2)=\int_{-\infty}^{\infty} U(y_1)h(y_2, y_1) dy_1$$

(4-33)

where

$$h(y_2, y_1)=e^{jkz}j\lambda z \exp[jk\lambda z(y_2-y_1)]$$

$$h(y_2, y_1)=\frac{e^{jkz}}{\sqrt{j\lambda z}} \exp\left[\frac{j\pi}{\lambda z}(y_2-y_1)^2\right]$$

(4-34)

Our results above show that when light is propagating through a system with an arbitrary ray-transfer matrix, rather than simply free space, the corresponding impulse response can be written

$$h(y_2, y_1)=e^{jk_0 L_0} B \lambda_0 \exp[jk_0 B \lambda_0 A y_{12} - 2 y_1 y_2 + D y_2^2],$$

$$h(y_2, y_1)=\frac{e^{jk_0 L_0}}{\sqrt{jB\lambda_0}} \exp\left[j\frac{\pi}{B\lambda_0}(Ay_1^2-2y_1y_2+Dy_2^2)\right],$$

(4-35)

where k_0 and λ_0 are the vacuum values of wavenumber and wavelength, and the factor B under the square root sign is needed to preserve energy. It follows that

$$U_2(y_2) = e^{jk_0 L_0} \int_{-\infty}^{\infty} U_1(y_1) \exp\left[\frac{j\pi}{B\lambda_0}(Ay_1^2 - 2y_1 y_2 + Dy_2^2)\right] dy_1,$$

$$U_2(y_2) = \frac{e^{jk_0 L_0}}{\sqrt{jB\lambda_0}} \int_{-\infty}^{\infty} U_1(y_1) \exp\left[\frac{j\pi}{B\lambda_0}(Ay_1^2 - 2y_1 y_2 + Dy_2^2)\right] dy_1, \quad (4-36)$$

where U_1 is the complex field at the input and U_2 is the complex field at the output.

While we have used a point source in Fig. 4.6 to illuminate the system, the resulting spherical wave incident on the system should be regarded as a probe that helps us determine the general complex-valued impulse response between the planes containing y_1 and y_2 . The result is therefore not limited to imaging systems, but rather can be applied to any paraxial system that has no limiting apertures. If limiting apertures exist, the ABCD system up to that aperture should be determined and the corresponding impulse response calculated, followed by a space-limitation of the result by the aperture, followed by calculation of the next ABCD system and its impulse response, etc.

The relations derived above can easily be extended to two-dimensional systems that are not astigmatic. For astigmatic systems, a 4×4 ray-transfer matrix is needed, but we will not consider that situation here. The reader is reminded that we have assumed that the optical system does not contain any apertures or pupil stops. If a stop exists, then the formalism must be applied from the input to the pupil stop, and then again from the pupil stop to the output.

4.3 The Fraunhofer Approximation

Before presenting several examples of diffraction pattern calculations, we consider another more stringent approximation which, when valid, greatly simplifies the calculations. It was seen in (4-17) that, in the region of Fresnel diffraction, the observed field strength $U(x, y)$ can be found from a Fourier transform of the product of the aperture distribution $U(\xi, \eta)$ and a quadratic phase function $\exp[j(k/2z)(\xi^2 + \eta^2)]$. If in addition to the Fresnel approximation the stronger (Fraunhofer) approximation

$$z \gg k(\xi^2 + \eta^2)_{\max}$$

$$z \gg \frac{k(\xi^2 + \eta^2)_{\max}}{2}$$

(4-37)

is satisfied, then the quadratic-phase factor under the integral sign in Eq.(4-17) is approximately unity over the entire aperture, and the observed field strength can be found (up to a multiplicative phase factor in (x, y)) directly from a *Fourier transform* of the aperture distribution itself. Thus in the region of *Fraunhofer diffraction* (or equivalently, in the *far field*),

$$U(x, y, z) = e^{jkze^{jk2z}(x^2 + y^2)} \int_{-\infty}^{\infty} \int U(\xi, \eta, 0) \exp[-j\frac{2\pi}{\lambda z}(x\xi + y\eta)] d\xi d\eta.$$

$$U(x, y, z) = \frac{e^{jkz}}{j\lambda z} e^{j\frac{k}{2z}(x^2 + y^2)} \int_{-\infty}^{\infty} \int U(\xi, \eta, 0) \exp\left[-j\frac{2\pi}{\lambda z}(x\xi + y\eta)\right] d\xi d\eta.$$

(4-38)

Aside from multiplicative factors preceding the integral, this expression is simply the Fourier transform of the aperture distribution, evaluated at spatial frequencies

$$f_X = x / \lambda z, f_Y = y / \lambda z.$$

$$f_X = x / \lambda z \\ f_Y = y / \lambda z.$$

(4-39)

At optical frequencies, the conditions required for validity of the Fraunhofer approximation can be severe ones. For example, at a wavelength of $0.6 \mu m$ (red light) and an aperture width of 2.5 cm (1 inch), the observation distance z must satisfy

$$z \gg 1,600 \text{ meters.}$$

$$z \gg 1,600 \text{ meters.}$$

An alternative, slightly less stringent condition, known as the “antenna designer’s formula,” states that for an aperture of linear dimension D , the Fraunhofer approximation will be valid provided

$$z > 2D^2\lambda$$

$$z > \frac{2D^2}{\lambda}$$

(4-40)

where the inequality is now $>$ rather than \gg . However, for this example the distance z is still required to be larger than 2,000 meters. Nonetheless, the required conditions are met in a number of important problems. In addition, Fraunhofer diffraction patterns can be observed at distances much closer than implied by (4-37) provided the aperture is illuminated by a spherical wave converging toward the observer (see [Prob. 4-18](#)), or if a positive lens is properly situated between the observer and the aperture (see [Chapter 6](#)).

Finally, it should be noted that there exists no transfer function that can be associated with Fraunhofer diffraction, for dropping the quadratic-phase exponential in (4-17)) has destroyed the space invariance of the diffraction equation (cf. [Prob. 2-12](#)). The secondary wavelets with quadratic-phase surfaces (as implied by the Fresnel approximation) no longer shift laterally in the (x, y) plane with the particular (ξ, η) point under consideration. Rather, when the location of the secondary source shifts, the corresponding quadratic-phase surface tilts in the (x, y) plane by an amount that depends on the location of the secondary source. Nonetheless, it should not be forgotten that since Fraunhofer diffraction is only a special case of Fresnel diffraction, the transfer function (4-20) remains valid throughout both the Fresnel and the Fraunhofer regimes. That is, it is always possible to calculate diffracted fields in the Fraunhofer region by retaining the full accuracy of the Fresnel approximation.

4.4 Examples of Fraunhofer Diffraction Patterns

We consider next several examples of Fraunhofer diffraction patterns. For additional examples the reader may consult the problems (see [Probs. 4-9 through 4-12](#)).

The results of the preceding section can be applied directly to find the complex field distribution across the Fraunhofer diffraction pattern of any given aperture. However, often of ultimate interest, for reasons discussed at the beginning of this chapter, is the intensity rather than the complex field strength. The final descriptions of the specific diffraction patterns considered here will therefore be distributions of intensity.

4.4.1 Rectangular Aperture

Consider first a rectangular aperture with an amplitude transmittance given by

$$t_A(\xi, \eta) = \text{rect}\left(\frac{\xi}{\ell_X}\right) \text{rect}\left(\frac{\eta}{\ell_Y}\right)$$

$$t_A(\xi, \eta) = \text{rect}\left(\frac{\xi}{\ell_X}\right) \text{rect}\left(\frac{\eta}{\ell_Y}\right)$$

The constants ℓ_X and ℓ_Y are the widths of the aperture in the ξ and η directions. If the aperture is illuminated by a unit-amplitude, normally incident, monochromatic plane wave, then the field distribution across the aperture is equal to the transmittance function t_A . Thus using [Eq.\(4-38\)](#), the Fraunhofer diffraction pattern is seen to be

$$U(x, y, z) = e^{jkz} e^{j\frac{k}{2z}(x^2 + y^2)} \mathcal{F}\{U(\xi, \eta, 0)\} \Big|_{f_x=x/\lambda z, f_y=y/\lambda z}$$

$$U(x, y, z) = \frac{e^{jkz} e^{j\frac{k}{2z}(x^2 + y^2)}}{j\lambda z} \mathcal{F}\{U(\xi, \eta, 0)\} \Big|_{f_x=x/\lambda z, f_y=y/\lambda z}$$

Noting that $\mathcal{F}\{U(\xi, \eta, 0)\} = A \text{sinc}(\ell_X f_x) \text{sinc}(\ell_Y f_y)$, where A is the area of the aperture ($A = \ell_X \ell_Y$), we find

$$U(x, y, z) = e^{jkz} e^{j\frac{k}{2z}(x^2 + y^2)} A \text{sinc}\left(\frac{\ell_X x}{\lambda z}\right) \text{sinc}\left(\frac{\ell_Y y}{\lambda z}\right)$$

$$U(x, y, z) = \frac{e^{jkz} e^{j\frac{k}{2z}(x^2 + y^2)}}{j\lambda z} A \text{sinc}\left(\frac{\ell_X x}{\lambda z}\right) \text{sinc}\left(\frac{\ell_Y y}{\lambda z}\right)$$

and

$$I(x, y, z) = A^2 \lambda^2 z^2 \text{sinc}^2\left(\frac{\ell_X x}{\lambda z}\right) \text{sinc}^2\left(\frac{\ell_Y y}{\lambda z}\right)$$

$$I(x, y, z) = \frac{A^2}{\lambda^2 z^2} \text{sinc}^2\left(\frac{\ell_X x}{\lambda z}\right) \text{sinc}^2\left(\frac{\ell_Y y}{\lambda z}\right)$$

(4-41)

[Figure 4.7](#) shows a cross section of the Fraunhofer intensity pattern along the x^X axis. Note that the width of the main lobe (i.e. the distance between the first two zeros) is

$$\Delta x = 2\lambda z \ell X.$$

$$\Delta x = 2 \frac{\lambda z}{\ell_X}.$$

(4-42)

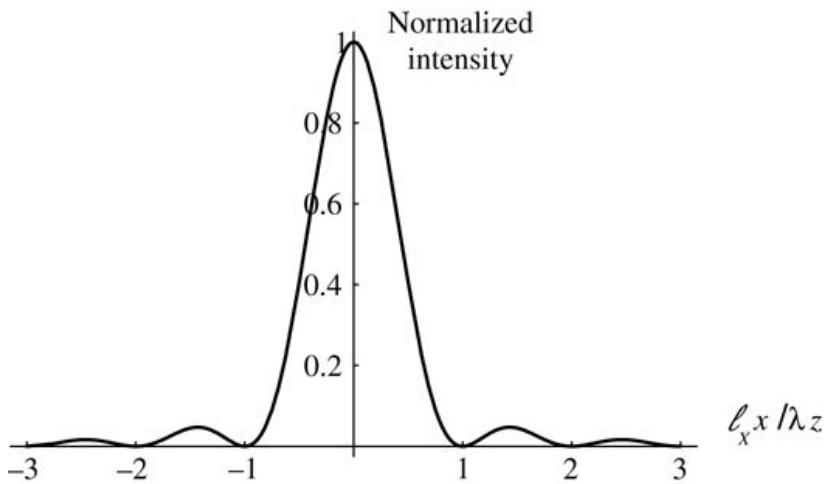


Figure 4.7

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 4.7 Cross section of the Fraunhofer diffraction pattern of a rectangular aperture.

A symmetric graph with its horizontal axis (ℓ subscript x times x divided by λz) ranging from minus 3 to +3 and vertical axis ranging from 0 to 1 shows a curve that is wavelike from minus 3 to minus 1 and between 3 and 1. At minus 1 it rises sharply to the +1 mark on the vertical axis and then falls sharply to + 1 on the horizontal axis.

[Figure 4.8](#) shows a photograph of the diffraction pattern produced by a rectangular aperture with a width ratio of $\ell X / \ell Y = 2$.

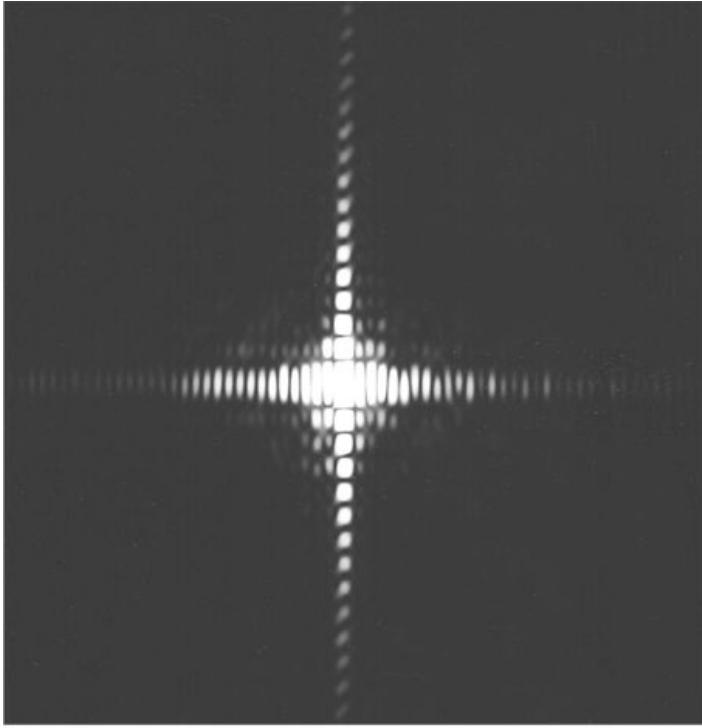


Figure 4.8

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 4.8 The Fraunhofer diffraction pattern of a rectangular aperture ($\ell_X/\ell_Y=2$).

4.4.2 Circular Aperture

Consider a diffracting aperture that is circular rather than rectangular, and let the diameter of the aperture be ℓ . Thus if q is a radius coordinate in the plane of the aperture, then

$$t_A(q) = \text{circ}\left(2\frac{q}{\ell}\right).$$

$$t_A(q) = \text{circ}\left(2\frac{q}{\ell}\right).$$

The circular symmetry of the problem suggests that the Fourier transform of (4-38) be rewritten as a Fourier-Bessel transform. Thus if r is the radius coordinate in the observation plane, we have

$$U(r, z) = e^{jkz} j \lambda z \exp(jkr) 2z \mathcal{B} U(q, 0) \rho = r/\lambda z,$$

$$U(r, z) = \frac{e^{jkz}}{j\lambda z} \exp\left(j\frac{kr^2}{2z}\right) \mathcal{B}\{U(q, 0)\} \Big|_{\rho=r/\lambda z},$$

(4-43)

where $q = \sqrt{\xi^2 + \eta^2}$ represents radius in the aperture plane, and $\rho = \sqrt{f_x^2 + f_y^2}$ represents radius in the spatial frequency domain. For unit-amplitude, normally

incident plane-wave illumination, the field transmitted by the aperture is equal to the amplitude transmittance; in addition,

$$\mathcal{B} \text{circ} 2q\ell = A 2J_1(\pi\ell\rho) \pi\ell\rho,$$

$$\mathcal{B} \left\{ \text{circ} \left(2 \frac{q}{\ell} \right) \right\} = A \left[2 \frac{J_1(\pi\ell\rho)}{\pi\ell\rho} \right],$$

where $A = \pi\ell/22$. The amplitude distribution in the Fraunhofer diffraction pattern is seen to be

$$U(r, z) = e^{jkr} \exp \left(j \frac{kr^2}{2z} \right) \frac{A}{j\lambda z} \left[2 \frac{J_1(k\ell r / 2z)}{k\ell r / 2z} \right],$$

$$U(r, z) = e^{jkr} \exp \left(j \frac{kr^2}{2z} \right) \frac{A}{j\lambda z} \left[2 \frac{J_1(k\ell r / 2z)}{k\ell r / 2z} \right],$$

and the intensity distribution can be written

$$I(r, z) = \left(\frac{A}{\lambda z} \right)^2 \left[2 \frac{J_1(k\ell r / 2z)}{k\ell r / 2z} \right]^2.$$

$$I(r, z) = \left(\frac{A}{\lambda z} \right)^2 \left[2 \frac{J_1(k\ell r / 2z)}{k\ell r / 2z} \right]^2.$$

(4-44)

This intensity distribution is referred to as the *Airy pattern*, after G.B. Airy who first derived it. Table 4.1 shows the values of the Airy pattern at successive maxima and minima, from which it can be seen that the half-width of the central lobe, measured along the x or y axis, is given by

$$d = 1.22\lambda z\ell.$$

$$d = 1.22 \frac{\lambda z}{\ell}.$$

(4-45)

Table 4.1: Locations of maxima and minima of the Airy pattern.

x	$\mathcal{J}_1(\pi x)$	$\mathcal{J}_1(\pi x) \pi x^2 \left[2 \frac{J_1(\pi x)}{\pi x} \right]^2$	max, min
0	1		max
1.220	0		min
1.635	0.0175		max
2.233	0		min
2.679	0.0042		max
3.238	0		min
3.699	0.0016		max

[Figure 4.9](#) shows a cross section of the Airy pattern, while [Fig. 4.10](#) is a photograph of the Fraunhofer diffraction pattern of a circular aperture.

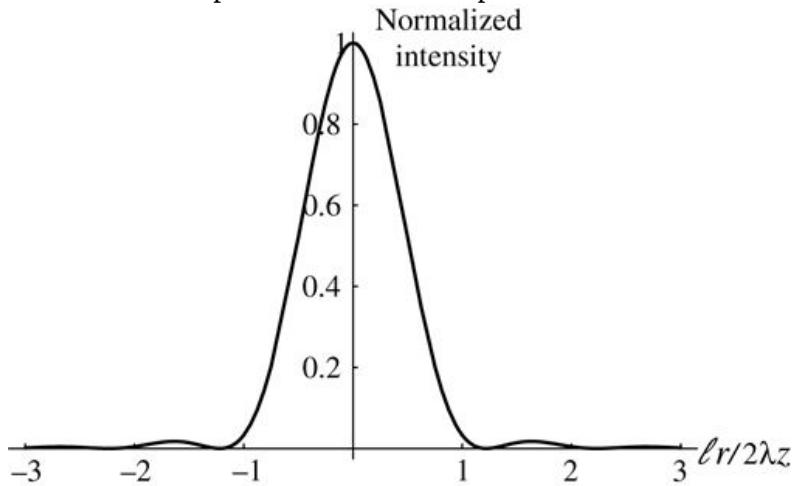


Figure 4.9

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 4.9 Cross section of the Fraunhofer diffraction pattern of a circular aperture.

A graph with its horizontal axis (labeled l times r divided by $2 \lambda z$) ranging from minus 3 to +3 and vertical axis ranging from 0 to 1 shows a curve that runs very close to the horizontal axis from minus 3 to minus 1 and between 3 and 1. At minus 1 it rises sharply to the +1 mark on the vertical axis and then falls sharply to +1 on the horizontal axis. The entire curve is thus symmetric about the vertical axis.

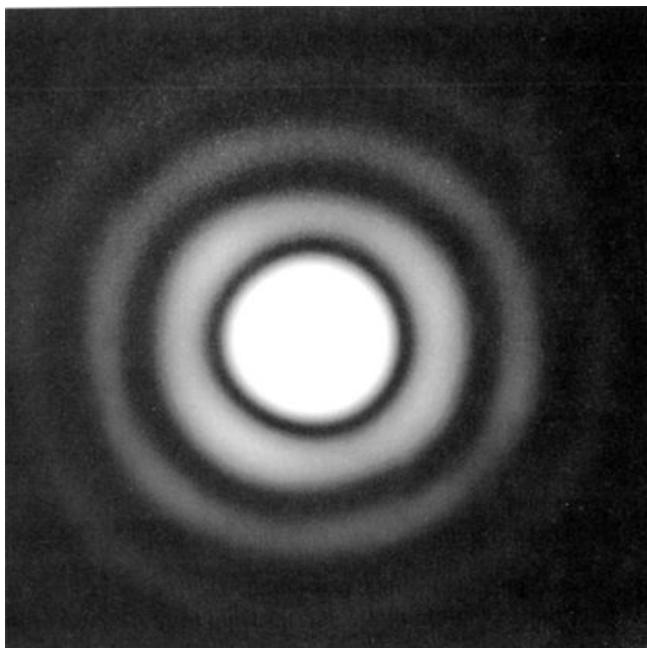


Figure 4.10

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 4.10 Fraunhofer diffraction pattern of a circular aperture.

4.4.3 Thin Sinusoidal Amplitude Grating

In the previous examples, diffraction was assumed to be caused by clear apertures in infinite opaque screens. In practice, diffracting objects can be far more complex. In accord with our earlier definition (3-74), the amplitude transmittance $t_A(\xi, \eta)$ of a screen is defined as the ratio of the complex amplitude of the field immediately behind the screen to the complex amplitude incident on the screen. Until now, our examples have involved only transmittance functions of the form

$$t_A(\xi, \eta) = 1 \text{ in the aperture} = 0 \text{ outside the aperture.}$$

$$t_A(\xi, \eta) = \begin{cases} 1 & \text{in the aperture} \\ 0 & \text{outside the aperture.} \end{cases}$$

It is possible, however, to introduce a prescribed amplitude transmittance function within a given aperture. Spatial attenuation can be introduced with, for example, an absorbing photographic transparency, thus allowing real values of t_A between zero and unity to be realized. Spatial patterns of phase shift can be introduced by means of transparent plates of varying thickness, thus extending the realizable values of t_A to all points within or on the unit circle in the complex plane.

As an example of this more general type of diffracting screen, consider a *thin sinusoidal amplitude grating* defined by the amplitude transmittance function

$$t_A(\xi, \eta) = 1 + m \cos(2\pi f_0 \xi) \operatorname{rect}\left(\frac{\xi}{\ell}\right) \operatorname{rect}\left(\frac{\eta}{\ell}\right)$$

$$t_A(\xi, \eta) = \left[\frac{1}{2} + \frac{m}{2} \cos(2\pi f_0 \xi) \right] \operatorname{rect}\left(\frac{\xi}{\ell}\right) \operatorname{rect}\left(\frac{\eta}{\ell}\right)$$

(4-46)

where for simplicity we have assumed that the grating structure is bounded by a square aperture of width ℓ . The parameter m represents the peak-to-valley change of amplitude transmittance across the screen, and f_0 is the spatial frequency of the grating. The term *thin* in this context means that the structure can indeed be represented by a simple amplitude transmittance. Structures that are not sufficiently thin can not be so represented, a point we shall return to in a later chapter. [Figure 4.11](#) shows a cross section of the grating amplitude transmittance function.

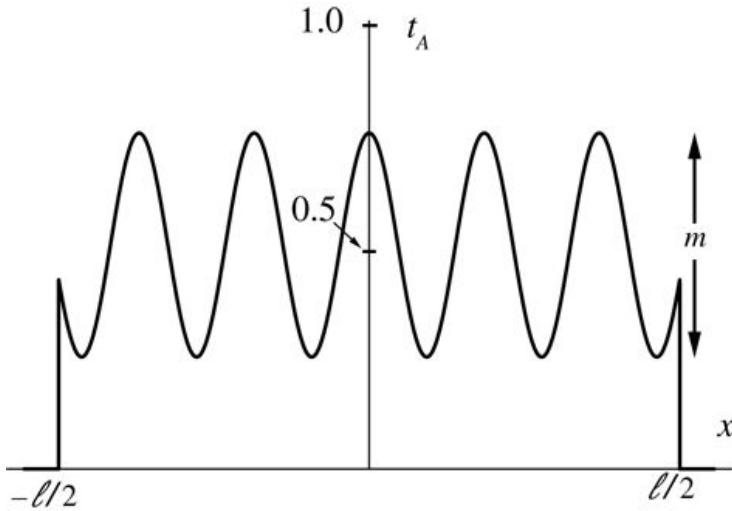


Figure 4.11

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 4.11 Amplitude transmittance function of the sinusoidal amplitude grating.

On the left side of the graph, the curve runs a short distance over the horizontal axis from an unmarked point to minus $l/2$, where it rises vertically to a height that is a little less than 0.5. It then slopes downward for approximately a third of the initial vertical distance. The curve then rises and falls in a uniform pattern of 4 troughs and 5 crests of uniform wavelength and amplitude in such that the curve is symmetrical about the vertical axis, which coincides with the central crest. The vertical height of the wave is marked m . The vertical axis ranges from 0 to 1.0, labeled t subscript A, 0.5 being the center of the wave's vertical height.

If the screen is normally illuminated by a unit-amplitude plane wave, the field distribution transmitted by the aperture is equal simply to t_A . To find the Fraunhofer diffraction pattern, we first Fourier transform that field distribution. Noting that

$$\mathcal{F}\{1 + m \cos(2\pi f_0 \xi)\} = 1 + m \delta(f_x + f_0, f_y) + m \delta(f_x - f_0, f_y)$$

$$\begin{aligned} \mathcal{F}\left\{\frac{1}{2} + \frac{m}{2} \cos(2\pi f_0 \xi)\right\} &= \frac{1}{2} \delta(f_x, f_y) \\ &\quad + \frac{m}{4} \delta(f_x + f_0, f_y) + \frac{m}{4} \delta(f_x - f_0, f_y) \end{aligned} \tag{4-47}$$

and

$$\mathcal{F}\text{rect}\xi \text{rect}\eta = A \text{sinc}(\ell f_x) \text{sinc}(\ell f_y),$$

$$\mathcal{F}\left\{\text{rect}\left(\frac{\xi}{\ell}\right) \text{rect}\left(\frac{\eta}{\ell}\right)\right\} = A \text{sinc}(\ell f_x) \text{sinc}(\ell f_y),$$

the convolution theorem can be used to write

$$\mathcal{F}U(\xi, \eta, 0) = A^2 \text{sinc}(\ell f_y) \text{sinc}(\ell f_x) + m^2 \text{sinc}(\ell f_x + f_0) + m^2 \text{sinc}(\ell f_x - f_0),$$

$$\mathcal{F}\{U(\xi, \eta, 0)\} = \frac{A}{2} \operatorname{sinc}(\ell f_Y) \left\{ \operatorname{sinc}(\ell f_X) + \frac{m}{2} \operatorname{sinc}[\ell(f_X + f_0)] + \frac{m}{2} \operatorname{sinc}[\ell(f_X - f_0)] \right\},$$

where A signifies the area of the aperture bounding the grating. The Fraunhofer diffraction pattern can now be written

$$U(x, y, z) = A j 2 \lambda z e j k z e^{j \frac{k}{2z}(x^2+y^2)} \operatorname{sinc}\left(\frac{\ell y}{\lambda z}\right) \left\{ \operatorname{sinc}\left(\frac{\ell x}{\lambda z}\right) + \frac{m}{2} \operatorname{sinc}\left[\frac{\ell}{\lambda z}(x + f_0 \lambda z)\right] + \frac{m}{2} \operatorname{sinc}\left[\frac{\ell}{\lambda z}(x - f_0 \lambda z)\right] \right\}. \quad (4-48)$$

$$U(x, y, z) = \frac{A}{j2\lambda z} e^{jkz} e^{j\frac{k}{2z}(x^2+y^2)} \operatorname{sinc}\left(\frac{\ell y}{\lambda z}\right) \left\{ \operatorname{sinc}\left(\frac{\ell x}{\lambda z}\right) + \frac{m}{2} \operatorname{sinc}\left[\frac{\ell}{\lambda z}(x + f_0 \lambda z)\right] + \frac{m}{2} \operatorname{sinc}\left[\frac{\ell}{\lambda z}(x - f_0 \lambda z)\right] \right\}.$$

(4-48)

Finally, the corresponding intensity distribution is found by taking the squared magnitude of (4-48). Note that if there are many grating periods within the aperture, then $f_0 \gg 1/\ell$, and there will be negligible overlap of the three sinc functions, allowing the intensity to be calculated as the sum of the squared magnitudes of the three terms in (4-48). The intensity is then given by

$$I(x, y, z) \approx A 2 \lambda z 2 \operatorname{sinc}^2 \ell y \lambda z \operatorname{sinc}^2 \ell x \lambda z + m 24 \operatorname{sinc}^2 \ell \lambda z (x + f_0 \lambda z) + m 24 \operatorname{sinc}^2 \ell \lambda z (x - f_0 \lambda z).$$

$$I(x, y, z) \approx \left[\frac{A}{2\lambda z} \right]^2 \operatorname{sinc}^2\left(\frac{\ell y}{\lambda z}\right) \left\{ \operatorname{sinc}^2\left(\frac{\ell x}{\lambda z}\right) + \frac{m^2}{4} \operatorname{sinc}^2\left[\frac{\ell}{\lambda z}(x + f_0 \lambda z)\right] + \frac{m^2}{4} \operatorname{sinc}^2\left[\frac{\ell}{\lambda z}(x - f_0 \lambda z)\right] \right\}.$$

(4-49)

This intensity pattern is illustrated in Fig. 4.12. Note that some of the incident light is absorbed by the grating, and in addition the sinusoidal transmittance variation across the aperture has deflected some of the energy out of the central diffraction pattern into two additional side patterns. The central diffraction pattern is called the *zero order* of the diffraction pattern, while the two sidelobes are called the *first orders*. The spatial separation of the first orders from the zero order is $f_0 \lambda z$, while the half-width (to the first zero) of the main lobe of all orders is $\lambda z / \ell$.

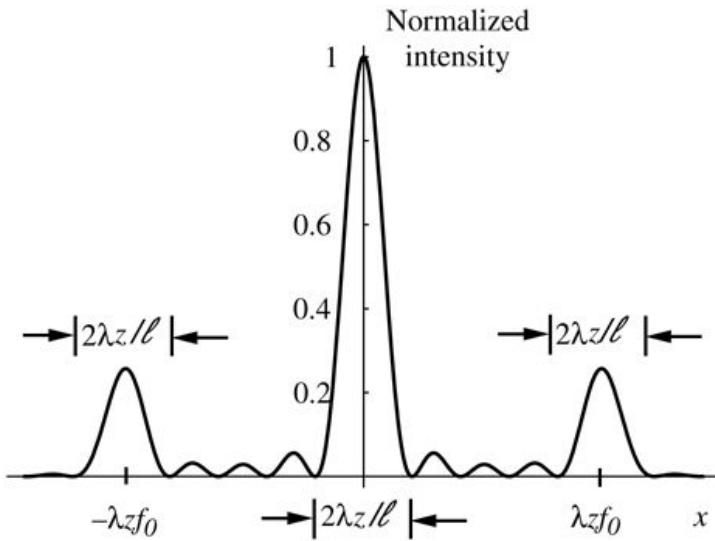


Figure 4.12

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 4.12 Fraunhofer diffraction pattern for a thin sinusoidal amplitude grating.

On the left of the vertical axis, a rightward curve runs a short distance over the horizontal axis and then rises in an upward slope to mark a crest at around 0.27 units above minus lambda z f subscript 0 on the horizontal axis. The curve then descends symmetrically to reach the horizontal axis. The curve continues the pattern to describe three crests of much shorter heights. Close to origin it rises sharply to crest at 1 on the vertical axis and then fall to the other side and continue in a way that mirrors the curve on the left of the vertical axis. The crests on the extremes and the one in the center are each 2 lambda z / l units wide.

Another quantity of some practical interest in both holography and optical information processing is the *diffraction efficiency* of the grating. The diffraction efficiency is defined as the fraction of the incident optical power that appears in a single diffraction order (usually the +1 order) of the grating. The diffraction efficiency for the grating of interest can be deduced from (4-47). The fraction of power appearing in each diffraction order can be found by squaring the coefficients of the delta functions in this representation, for it is the delta functions that determine the power in each order, not the sinc functions that simply spread these impulses. From this equation we conclude that the diffraction efficiencies $\eta_0, \eta_{+1}, \eta_{-1}$ associated with the three diffraction orders are given by

$$\eta_0 = 0.25 \quad \eta_{+1} = m^2 / 16 \quad \eta_{-1} = m^2 / 16.$$

$$\begin{aligned}\eta_0 &= 0.25 \\ \eta_{+1} &= m^2 / 16 \\ \eta_{-1} &= m^2 / 16.\end{aligned}$$

(4-50)

Thus a single first diffraction order carries at most $1/16 = 6.25\%$ of the incident power, a rather small fraction. If the efficiencies of the three orders are added up, it will be seen

that only $1/4 + m^2/8$ of the total is accounted for. The rest is lost through absorption by the grating.

For a further discussion of gratings and their orders, see [Appendix D](#).

4.4.4 Thin Sinusoidal Phase Grating

As a final example of Fraunhofer diffraction calculations, consider a *thin sinusoidal phase grating* defined by the amplitude transmittance function

$$t_A(\xi, \eta) = \exp[jm2\sin(2\pi f_0 \xi)] \operatorname{rect}\left(\frac{\xi}{\ell}\right) \operatorname{rect}\left(\frac{\eta}{\ell}\right)$$

(4-51)

where, by proper choice of phase reference, we have dropped a factor representing the average phase delay through the grating. The parameter m represents the peak-to-valley excursion of the phase delay.

If the grating is illuminated by a unit-amplitude, normally incident plane wave, then the field distribution immediately behind the screen is given precisely by [Eq.\(4-51\)](#). The analysis is simplified by use of the identity

$$\exp[jm2\sin(2\pi f_0 \xi)] = \sum_{q=-\infty}^{\infty} J_q(m) \exp(j2\pi q f_0 \xi)$$

$$\exp[j\frac{m}{2}\sin(2\pi f_0 \xi)] = \sum_{q=-\infty}^{\infty} J_q\left(\frac{m}{2}\right) \exp(j2\pi q f_0 \xi)$$

where J_q is a Bessel function of the first kind, order q . Thus

$$\mathcal{F}\{\exp[jm2\sin(2\pi f_0 \xi)]\} = \sum_{q=-\infty}^{\infty} J_q(m) \delta(f_X - qf_0, f_Y)$$

$$\mathcal{F}\left\{\exp[j\frac{m}{2}\sin(2\pi f_0 \xi)]\right\} = \sum_{q=-\infty}^{\infty} J_q\left(\frac{m}{2}\right) \delta(f_X - qf_0, f_Y)$$

(4-52)

and

$$\begin{aligned} \mathcal{F}\{U(\xi, \eta, 0)\} &= \mathcal{F}\{t_A(\xi, \eta)\} \\ &= [A \operatorname{sinc}(\ell f_X) \operatorname{sinc}(\ell f_Y)] * [\sum_{q=-\infty}^{\infty} J_q(m) \delta(f_X - qf_0, f_Y)] \\ &= \sum_{q=-\infty}^{\infty} A J_q(m) \operatorname{sinc}[\ell(f_X - qf_0)] \operatorname{sinc}(\ell f_Y). \end{aligned}$$

$$\begin{aligned} \mathcal{F}\{U(\xi, \eta, 0)\} &= \mathcal{F}\{t_A(\xi, \eta)\} \\ &= [A \operatorname{sinc}(\ell f_X) \operatorname{sinc}(\ell f_Y)] * \left[\sum_{q=-\infty}^{\infty} J_q\left(\frac{m}{2}\right) \delta(f_X - qf_0, f_Y) \right] \\ &= \sum_{q=-\infty}^{\infty} A J_q\left(\frac{m}{2}\right) \operatorname{sinc}[\ell(f_X - qf_0)] \operatorname{sinc}(\ell f_Y). \end{aligned}$$

Thus the field strength in the Fraunhofer diffraction pattern can be written

$$U(x, y, z) = A j \lambda z e^{j k z} e^{j k 2z(x^2 + y^2)} \times \sum_{q=-\infty}^{\infty} J_q(m) \operatorname{sinc}[\ell \lambda z(x - q f_0 \lambda z)] \operatorname{sinc}(\ell y \lambda z).$$

$$U(x, y, z) = \frac{A}{j \lambda z} e^{j k z} e^{j \frac{k}{2z}(x^2 + y^2)} \times \sum_{q=-\infty}^{\infty} J_q\left(\frac{m}{2}\right) \operatorname{sinc}\left[\frac{\ell}{\lambda z}(x - q f_0 \lambda z)\right] \operatorname{sinc}\left(\frac{\ell y}{\lambda z}\right).$$

(4-53)

If we again assume that there are many periods of the grating within the bounding aperture ($f_0 \gg 1/\ell$), there is negligible overlap of the various diffracted terms, and the corresponding intensity pattern becomes

$$I(x, y, z) \approx A \lambda z^2 \sum_{q=-\infty}^{\infty} J_q^2\left(\frac{m}{2}\right) \operatorname{sinc}^2\left[\frac{\ell}{\lambda z}(x - q f_0 \lambda z)\right] \operatorname{sinc}^2\left(\frac{\ell y}{\lambda z}\right).$$

$$I(x, y, z) \approx \left(\frac{A}{\lambda z}\right)^2 \sum_{q=-\infty}^{\infty} J_q^2\left(\frac{m}{2}\right) \operatorname{sinc}^2\left[\frac{\ell}{\lambda z}(x - q f_0 \lambda z)\right] \operatorname{sinc}^2\left(\frac{\ell y}{\lambda z}\right).$$

(4-54)

The introduction of the sinusoidal phase grating has thus deflected energy out of the zero order into a multitude of higher orders. The peak intensity of the q^{th} order is given by $A J_q m / \lambda z^2 [A J_q(m/2)^2 / \lambda z]^2$, while the displacement of that order from the center of the diffraction pattern is $q f_0 \lambda z$. [Figure 4.13](#) shows a cross section of the intensity pattern when the peak-to-peak phase delay m is 8 radians. Note that the strengths of the various orders are symmetric about the zero order.

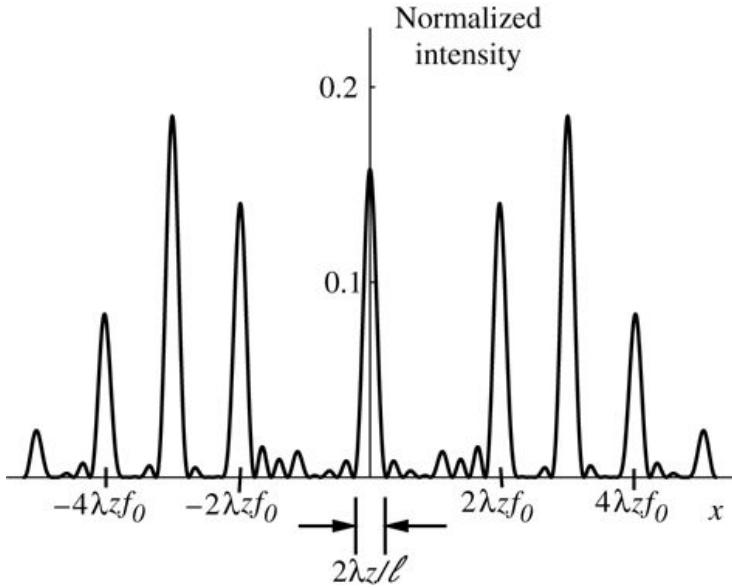


Figure 4.13

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 4.13 Fraunhofer diffraction pattern for a thin sinusoidal phase grating. The ± 1 orders have nearly vanished in this example. Note that in this example there is some overlap of the diffraction orders.

On both sides of the vertical axis, a curve makes short irregular wavelike path rising from and staying near the horizontal axis. In places it has sharp prominent crests whose locations and heights are as follows. 0.025 units at minus 5 lambda z f subscript 0 and + 5 lambda z f subscript 0; 0.08 units at minus 4 lambda z f subscript 0 and + 4 lambda z f subscript 0; 0.18 units at minus 3 lambda z f subscript 0 and + 3 lambda z f subscript 0; 0.14 units at minus 2 lambda z f subscript 0 and + 2 lambda z f subscript 0; 0.16 units at the origin

The diffraction efficiency of the thin sinusoidal phase grating can be found by determining the squared magnitude of the coefficients in (4-52). Thus the diffraction efficiency of the q^{th} order of this grating is

$$\eta_q = J_q^2(m/2).$$

$$\eta_q = J_q^2(m/2).$$

(4-55)

Figure 4.14 shows a plot of η_q vs. $m/2$ for various values of q . Note that whenever $m/2$ is a root of J_0 , the central order vanishes entirely! The largest possible diffraction efficiency into one of the $+1$ and -1 diffraction orders is the maximum value of J_{12}^2 . This maximum is 33.8%, far greater than for the case of a thin sinusoidal amplitude grating. No power is absorbed by this grating, and therefore the sum of the powers appearing in all orders remains constant and equal to the incident power within the aperture as m is changed.

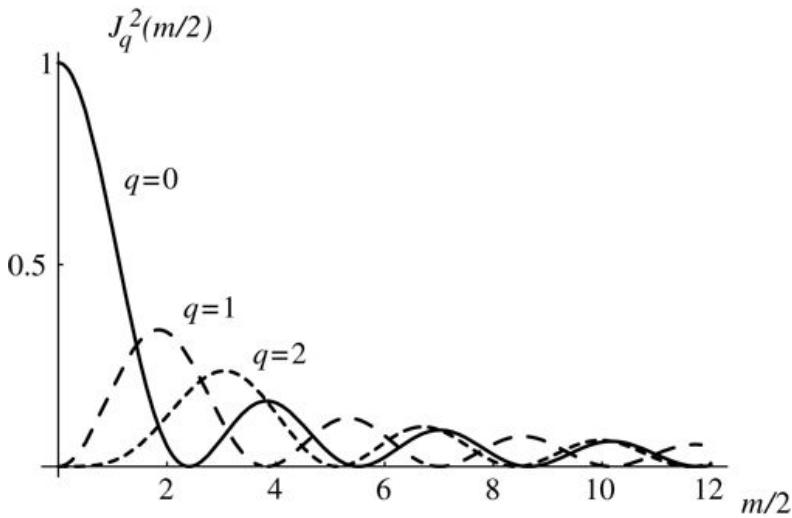


Figure 4.14

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 4.14 Diffraction efficiency $J_q^2(m/2)$ vs. $m/2$ for three values of q .

The graph plotting $m/2$ values from 0 to 12 on the horizontal axis and diffraction efficiency from 0 to 1 on the vertical axis shows three curves. The approximate data are as follows. The curve for $q = 0$ begins at the 1 mark on the vertical axis and drops sharply rightward to the mark of 2.5 on the horizontal axis. Thereafter it rises thrice between 0 and 12, describing crests at diffraction efficiency values of 0.2, 0.1, and 0.75. The curve for $q = 1$ begins at the origin and follows a wavelike pattern of decreasing amplitude that places the troughs on the horizontal axis, rising four times between 0 and 12 to describe crests at diffraction efficiency values of 0.35, 0.125, 0.06, and 0.05. The curve for $q = 2$ is like the curve for $q = 1$, but rising thrice between 0 and 12 to describe crests at diffraction efficiency values of 0.25, 0.1, and 0.06.

4.4.5 General Method for Calculating Diffraction Efficiency of Gratings

Before closing this section on Fraunhofer diffraction, we describe a very general method for finding the diffraction efficiencies of the various diffraction orders of an arbitrary thin periodic grating. Let the grating amplitude transmittance profile be represented by the periodic function $P(x)$ (assumed to be uniform in the y direction). This function may be complex-valued or real-valued. Ignore the bounding aperture or any broad window function that may exist on the grating, for the window function influences only the shape of the orders and not the power carried by them. Now expand the function $P(x)$ in a complex Fourier series,

$$P(x) = \sum_{q=-\infty}^{\infty} c_q e^{j2\pi qx/L},$$

$$P(x) = \sum_{q=-\infty}^{\infty} c_q e^{j2\pi qx/L},$$

(4-56)

where L is the period of the grating and the Fourier series coefficients are

$$c_q = \frac{1}{L} \int_{-L/2}^{L/2} P(x) e^{-j2\pi q x/L} dx.$$

$$c_q = \frac{1}{L} \int_{-L/2}^{L/2} P(x) e^{-j2\pi q x/L} dx. \quad (4-57)$$

Then the diffraction efficiency of the q^{th} order will be

$$\eta_q = |c_q|^2.$$

$$\eta_q = |c_q|^2. \quad (4-58)$$

This formalism allows one to easily calculate diffraction efficiencies of the orders of an arbitrary thin grating.

4.5 Examples of Fresnel Diffraction Calculations

In a previous section, several different methods for calculating Fresnel diffraction patterns have been introduced. For the beginner, it is difficult to know when one method will be easier than another, and therefore in this section two examples are presented that provide some insight in this regard. The first example, Fresnel diffraction by a square aperture, illustrates the application of the classical space-domain approach based on the convolution representation of the diffraction calculation. The second example, Talbot imaging, illustrates a case in which a frequency-domain approach has a large advantage.

4.5.1 Fresnel Diffraction by a Square Aperture

Suppose that a square aperture of width ℓ is normally illuminated by a monochromatic plane wave of unit amplitude. The distribution of complex field immediately behind the aperture is

$$U(\xi, \eta, 0) = \text{rect}\xi\ell \text{rect}\eta\ell.$$

$$U(\xi, \eta, 0) = \text{rect}\left(\frac{\xi}{\ell}\right) \text{rect}\left(\frac{\eta}{\ell}\right).$$

The convolution form of the Fresnel diffraction equation is most convenient for this problem, yielding

$$U(x, y, z) = e^{jkz} j \lambda z \int_{-\ell/2}^{\ell/2} \int_{-\ell/2}^{\ell/2} \exp\left\{j\frac{\pi}{\lambda z}[(x - \xi)^2 + (y - \eta)^2]\right\} d\xi d\eta.$$

$$U(x, y, z) = \frac{e^{jkz}}{j\lambda z} \int_{-\ell/2}^{\ell/2} \int_{-\ell/2}^{\ell/2} \exp\left\{j\frac{\pi}{\lambda z}[(x - \xi)^2 + (y - \eta)^2]\right\} d\xi d\eta.$$

This expression can be separated into the product of two one-dimensional integrals,

$$U(x, y, z) = e^{jkz} j \mathcal{J}_X(x) \mathcal{J}_Y(y)$$

$$U(x, y, z) = \frac{e^{jkz}}{j} \mathcal{J}_X(x) \mathcal{J}_Y(y)$$

(4-59)

where

$$\mathcal{J}_X(x) = \lambda z \int_{-\ell/2}^{\ell/2} \exp\left\{j\frac{\pi}{\lambda z}[(x - \xi)^2]\right\} d\xi$$

$$\mathcal{J}_Y(y) = \lambda z \int_{-\ell/2}^{\ell/2} \exp\left\{j\frac{\pi}{\lambda z}[(y - \eta)^2]\right\} d\eta.$$

$$\begin{aligned}\mathcal{J}_X(x) &= \frac{1}{\sqrt{\lambda z}} \int_{-\ell/2}^{\ell/2} \exp \left[j \frac{\pi}{\lambda z} (\xi - x)^2 \right] d\xi \\ \mathcal{J}_Y(y) &= \frac{1}{\sqrt{\lambda z}} \int_{-\ell/2}^{\ell/2} \exp \left[j \frac{\pi}{\lambda z} (\eta - y)^2 \right] d\eta.\end{aligned}\quad (4-60)$$

These integrals are identical in form, so we can focus our attention on one of them, $\mathcal{J}_X(x)$.

To reduce this integral to an expression that is related to the Fresnel integrals mentioned on several previous occasions, make the following change of variables: $\xi' = \xi / \sqrt{\lambda z}$, yielding

$$\mathcal{J}_X(x) = \int -N F N F \exp[j \pi \xi' - N F x \ell / 2] d\xi'$$

$$\mathcal{J}_X(x) = \int_{-\sqrt{N_F}}^{\sqrt{N_F}} \exp \left[j \pi \left(\xi' - \sqrt{N_F} \frac{x}{\ell} \right)^2 \right] d\xi'\quad (4-61)$$

where the quantity N_F is an important dimensionless description of the geometry called the *Fresnel number*, and is given by

$$N_F = (\ell / 2) / 2 \lambda z.$$

$$N_F = \frac{(\ell / 2)^2}{\lambda z}.$$

(4-62)

If the integrand of this integral is expressed in terms of sin and cos, rather than an exponential, the results of integration can be expressed in terms of *Fresnel integrals* defined by

$$C(z) = \int_0^z \cos(\pi t^2 / 2) dt, \quad S(z) = \int_0^z \sin(\pi t^2 / 2) dt,$$

$$\begin{aligned}C(z) &= \int_0^z \cos(\pi t^2 / 2) dt \\ S(z) &= \int_0^z \sin(\pi t^2 / 2) dt,\end{aligned}\quad (4-63)$$

allowing us to write

$$\mathcal{J}_X(x) = 12 C^2 N_F^2 (1 - 2x/\ell + C^2 N_F^2 (1 + 2x/\ell + j 2 S^2 N_F^2 (1 - 2x/\ell + S^2 N_F^2 (1 + 2x/\ell.$$

$$I_X(x) = \frac{1}{\sqrt{2}}[C(\sqrt{2N_F}(1 - 2x/\ell)) + C(\sqrt{2N_F}(1 + 2x/\ell))] \\ + \frac{j}{\sqrt{2}}[S(\sqrt{2N_F}(1 - 2x/\ell)) + S(\sqrt{2N_F}(1 + 2x/\ell))].$$

(4-64)

The Fresnel integrals are tabulated functions and are available in many mathematical computer programs such as *Mathematica* and *MatLab*.³ It is therefore a straightforward matter to calculate the above intensity distribution. Note that, for fixed ℓ and λ , as z increases the Fresnel number N_F decreases and the normalized space coordinate $2N_F z / \ell$ enlarges the true physical width of the diffraction pattern. [Figure 4.15](#) shows a series of graphs of the normalized intensity distribution along the x axis ($y=0$) for various normalized distances from the aperture, as represented by different Fresnel numbers.

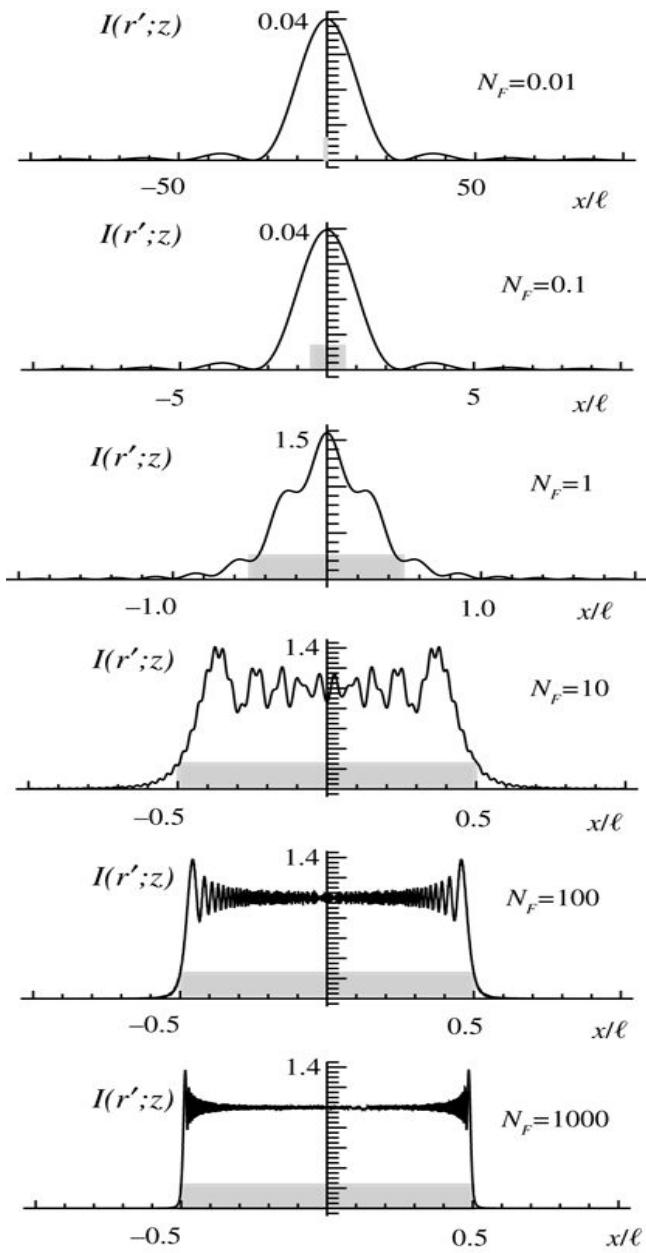


Figure 4.15
Goodman, Introduction to Fourier Optics, 4e,
 © 2017 W. H. Freeman and Company

Figure 4.15 Normal Fresnel diffraction patterns at different distances from a square aperture. The width of the diffraction pattern increases as the Fresnel number N_F shrinks. The width of the original rectangular aperture is indicated by the width of the shaded area.

1. The curve in the first graph, for $N_F = 0.01$, runs almost overlapping the negative half of the horizontal axis, rising slightly at minus 50 and then returning to the horizontal axis at minus 25, where it rises steeply in a rightward slope up to the 0.04 mark on the vertical axis, which mirrors the path so far on to the other side. The width of the original rectangular aperture is indicated by a gray band slightly wider than the axis extending from the origin to the 0.006 mark on the vertical axis.
2. The curve in the second graph, for $N_F = 0.1$, is similar to the

previous one with two differences. The curve's slight rise and drop are at minus 5 and minus 2.5 and the much wider gray band extends from the origin to the 0.007 mark on the vertical axis and from minus 6 to +6 on the horizontal axis. 3. 3. The curve in the third graph, for N subscript F = 1, runs almost overlapping the negative half of the horizontal axis, rising slightly at minus 1.0 and then forming a crest without fully returning to the horizontal axis near minus 0.8, where it similarly rises further to form a gentle crest. Thereafter the curve rises steeply in a rightward slope to a height of 1 unit on the vertical axis where it crests slightly and then in another steep stretch reaches the 1.5 mark on the vertical axis, which mirrors the path so far on to the other side. The gray band extends from the origin to the 0.3 mark on the vertical axis and from minus 0.45 to +0.45 on the horizontal axis. 4. 4. The curve in the fourth graph, for N subscript F = 10, runs almost overlapping the negative half of the horizontal axis, rising at minus 0.5 in a steep step-like path going rightward to a height corresponding to the 1.4 mark on the vertical axis. The curve then extends rightward dropping and rising in an irregular fashion within the 1.2 and 0.8 range of the vertical axis, which mirrors the path so far on to the other side. The gray band extends from the origin to the 0.25 mark on the vertical axis and from minus 0.5 to +0.5 on the horizontal axis. 5. 5. The curve in the fifth graph, for N subscript F = 100, rises almost vertically at minus 0.5 going slightly rightward up to a height corresponding to the 1.4 mark on the vertical axis. The curve then extends horizontally rightward, dropping and rising in a regular wavelike pattern between the levels 0.8 and 1.2 on the vertical axis. The pattern grows denser and shorter as it approaches the 1.0 mark on the vertical axis, which mirrors the path so far on to the other side. The gray band extends from the origin to the 0.25 mark on the vertical axis and from minus 0.5 to +0.5 on the horizontal axis. 6. 6. The curve in the sixth graph, for N subscript F = 1000, is the same as the previous one with a few differences. The line that rises at minus 0.5 is more vertical, almost perpendicular, and the horizontal wavelike pattern is much denser and grows shorter by minus 0.3, thus the waves almost appear like a thick line between minus 0.3 and +0.3.

Attention is called to the fact that, as the observation plane approaches the plane of the aperture (N_F becomes large), the Fresnel kernel approaches the product of a delta function and a factor $e^{jkz} e^{jky}$, and the shape of the diffraction pattern approaches the shape of the aperture itself. In fact, the limit of this process is the geometrical optics prediction of the complex field,

$$U(x, y, z) = e^{jkz} U(x, y, 0) = e^{jkz} \text{rect}(x/\ell) \text{rect}(y/\ell)$$

$$U(x, y, z) = e^{jkz} U(x, y, 0) = e^{jkz} \text{rect}\left(\frac{x}{\ell}\right) \text{rect}\left(\frac{y}{\ell}\right)$$

where, to avoid confusion, we have explicitly included the z coordinate in the argument of the complex field U .

Note also that, as the distance z becomes large (N_F grows small), the diffraction pattern becomes much wider than the size of the aperture, and comparatively smooth in its structure. In this limit the diffraction pattern is approaching the Fraunhofer limit discussed earlier.

4.5.2 Fresnel Diffraction by a Circular Aperture

The diffraction pattern produced by a circularly symmetric aperture is itself circularly symmetric. The structure of the diffraction pattern can therefore be described by either a two-dimensional plot of intensity or a simple radial profile of the intensity distribution along any axis passing through the origin. To find the radial profile of the diffraction pattern, we can use the Fourier-Bessel

transform, writing the field $U(x, y, z)$ in the diffraction pattern in terms of its radial profile $R(r; z)$, which in turn can be expressed in terms of the radial profile of the aperture field $R(\rho; 0)$ according to

$$R(r; z) = 2\pi e^{j\pi r^2 / (\lambda z)} \int_0^\infty \rho R(\rho; 0) \exp(j\pi\rho^2 / (\lambda z)) J_0(2\pi\rho r / (\lambda z)) d\rho,$$

$$R(r; z) = \frac{2\pi e^{j\pi r^2 / (\lambda z)}}{\lambda z} \int_0^\infty \rho R(\rho; 0) \exp(j\pi\rho^2 / (\lambda z)) J_0(2\pi\rho r / (\lambda z)) d\rho,$$

(4-65)

where $R(\rho; 0)$ is the radial profile of the circularly symmetric aperture function, ρ is radius in the aperture plane, and r is radius in the diffraction pattern plane. Changing variables of integration to $\rho' = \rho / \lambda z$, replacing $r / \lambda z$ by $r' r'$ and dropping the complex exponential in front of the integral since we are interested in intensity yields the equation

$$R(r'; z) = 2\pi \int_0^\infty \rho' R(\rho'; 0) \exp(j\pi\rho'^2) J_0(2\pi\rho' r' / (\lambda z)) d\rho'.$$

$$R(r'; z) = 2\pi \int_0^\infty \rho' R(\rho'; 0) \exp(j\pi\rho'^2) J_0(2\pi\rho' r' / (\lambda z)) d\rho'.$$

(4-66)

In the case of a circular aperture, $R(\rho'; 0) = \text{rect}(\rho'/\text{NF})$, where $\text{NF} = (\ell/2)^2 / (\lambda z)$ is the Fresnel number, ℓ is the diameter of the aperture, and the radial profile in the diffraction pattern can be written

$$R(r'; z) = 2\pi \int_0^{\sqrt{\text{NF}}} \rho' \exp(j\pi\rho'^2) J_0(2\pi\rho' r' / (\lambda z)) d\rho'.$$

$$R(r'; z) = 2\pi \int_0^{\sqrt{\text{NF}}} \rho' \exp(j\pi\rho'^2) J_0(2\pi\rho' r' / (\lambda z)) d\rho'.$$

(4-67)

Note that the variable $r' r'$ can also be expressed as $\text{NF}\ell/2\sqrt{\text{NF}}(\frac{r}{\ell/2})$.

Unfortunately no closed form solution exists for this integral, but it can be calculated by numerical integration. [Figure 4.16](#) shows the diffraction pattern intensities $|R(r'; z)|^2$ for $\text{NF}=0.01, 0.1, 1, 10, 100$, and 1000 . The radial variable has been extended to negative values in order to show the symmetry of the patterns about the origin. The zero of intensity on the optical axis for $\text{NF}=10, 100$ and 1000 is the complement of the Poisson spot mentioned in [Chapter 2](#), as predicted by Babinet's principle⁴.

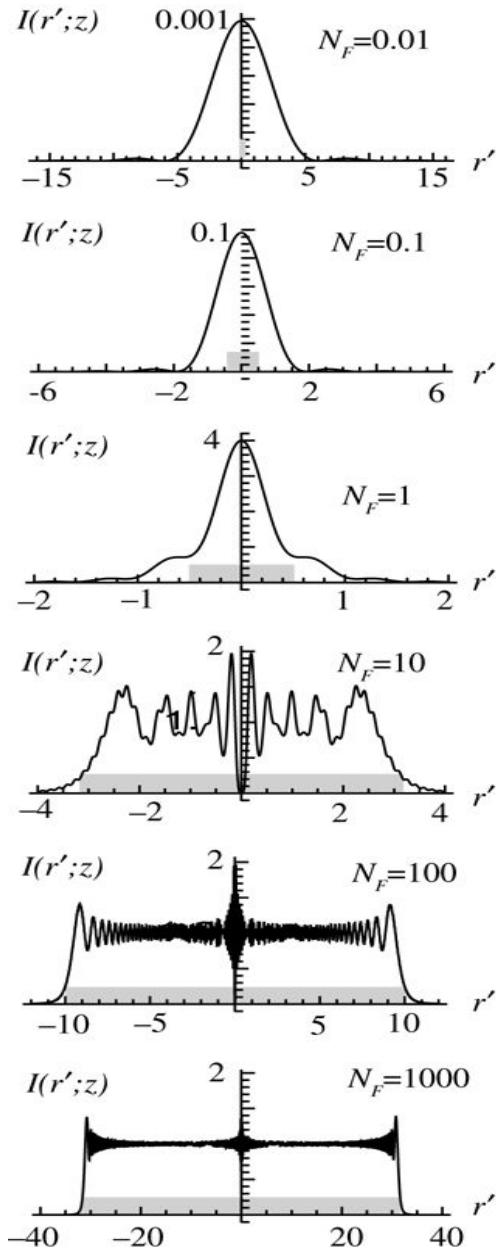


Figure 4.16
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 4.16 Fresnel diffraction patterns at different distances from a circular aperture. The width of the diffraction pattern increases as the Fresnel number N_F shrinks. The diameter of the original circular aperture is indicated by the width of the shaded area.

The patterns for $N_F = 10, 100$ and 1000 have a zero of intensity at their center.

1. The curve in the first graph, for $N_F = 0.01$, runs almost overlapping the negative half of the horizontal axis, rising sharply at minus 5 in a rightward slope up to the 0.001 mark on the vertical axis, which mirrors the path so far on to the other side. The width of the original rectangular aperture is indicated by a gray band slightly wider than the axis extending from the origin to the 0.00015 mark on the vertical axis.
2. The curve in the second graph, for $N_F = 0.1$, is similar to the previous one with some differences. The curve rises sharply at minus 2 in

a rightward slope that reaches the 0.1 mark on the vertical axis. The gray band extends from the origin to the 0.015 mark on the vertical axis and from minus 0.5 to +0.5 on the horizontal axis. 3. 3. The curve in the third graph, for N subscript $F = 1$, runs almost overlapping the negative half of the horizontal axis, rising slightly at minus 1.5 and then rising steeply at minus 1 to run a short horizontal distance before rising even more steeply at around minus 0.5 in a rightward slope to reach the 4 mark on the vertical axis, which mirrors the path so far on to the other side. The gray band extends from the origin to the 0.5 mark on the vertical axis and from minus 0.5 to +0.5 on the horizontal axis. 4.4. The curve in the fourth graph, for N subscript $F = 10$, rises steeply at minus 4 in a rightward slope to a height corresponding approximately to the 1.5 mark on the vertical axis. The curve then extends rightward dropping and rising in an irregular fashion within the 0.5 and 1.5 range of the vertical axis, finally rising to 2 and then falling sharply to the origin. The vertical axis mirrors the path so far on to the other side. The gray band extends from the origin to the 0.3 mark on the vertical axis and from minus 3.2 to +3.2 on the horizontal axis. 5. 5. The curve in the fifth graph, for N subscript $F = 100$, rises almost vertically at minus 10 going slightly rightward up to a height corresponding to the 1.4 mark on the vertical axis. The curve then extends horizontally rightward, dropping and rising in a regular wavelike pattern between the levels 0.8 and 1.2 on the vertical axis. The pattern grows denser and shorter, and then midway begins to gradually become broader as it approaches the 1.0 mark on the vertical axis; around the point (1, minus 1), the wavelike pattern greatly increases its sweep to finally extend from 0.5 to 1.5 on reaching the vertical axis, which mirrors the path so far on to the other side. The gray band extends from the origin to the 2.5 mark on the vertical axis and from minus 10 to +10 on the horizontal axis. 6. 6. The curve in the sixth graph, for N subscript $F = 1000$, is the same as the previous one with a few differences. The line rises at minus 33 and is more vertical, almost perpendicular and the horizontal wavelike pattern is much denser and grows shorter by minus 20, thus the waves almost appear like a thick line between minus 20 and +20, widening a bit as they approach the vertical axis.

4.5.3 Fresnel Diffraction by a Sinusoidal Amplitude Grating-Talbot Images

Our final example of a diffraction calculation considers again the case of a thin sinusoidal amplitude grating, but this time within the region of Fresnel diffraction rather than Fraunhofer diffraction. For simplicity we initially neglect the finite extent of the grating and concentrate on the effects of diffraction and propagation on the periodic structure of the fields transmitted by the grating.

The geometry is illustrated in [Fig. 4.17](#). The grating is modeled as a transmitting structure with amplitude transmittance

$$tA(\xi, \eta) = 121 + m\cos(2\pi\xi/L)$$

$$t_A(\xi, \eta) = \frac{1}{2}[1 + m\cos(2\pi\xi/L)]$$

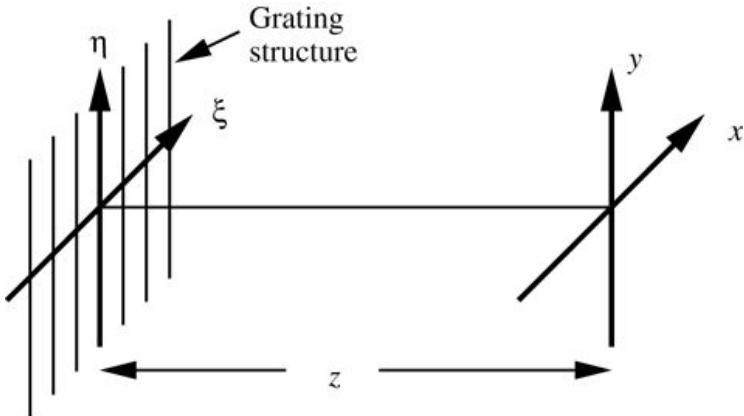


Figure 4.17

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 4.17 Geometry for diffraction calculation.

The illustration shows two 3 dimensional graphs with common horizontal axis z . On the left extreme of the z axis is a vertical axis η intersected by the third axis ξ . Parallel lines perpendicular to the ξ axis are labeled “grating structure.” On the right extreme of the z axis is the y axis parallel to the η axis and the x axis parallel to the ξ axis.

with period L and with the grating lines running parallel to the η axis. The field and intensity will be calculated some distance z to the right of the grating. The structure is assumed to be illuminated by a unit-amplitude, normally incident plane wave, so the field immediately behind the grating is equal to the amplitude transmittance written above.

There are several possible approaches to calculating the fields behind the grating. We could use the convolution form of the Fresnel diffraction equation, i.e. [Eq. \(4-14\)](#), or the “Fresnel transform” form of [\(4-17\)](#). Alternatively, we could use the Fresnel transfer function approach represented by [\(4-20\)](#), and reproduced here as

$$H(f_X, f_Y) = \exp[-j\pi\lambda z(f_X^2 + f_Y^2)],$$

$$H(f_X, f_Y) = \exp[-j\pi\lambda z(f_X^2 + f_Y^2)],$$

(4-68)

where we have omitted a constant term $\exp(jkz)$. In this problem, and indeed in any problem that deals with a purely periodic structure, the transfer function approach will yield the simplest calculations, and we adopt that approach here.

The solution begins by first finding the spatial frequency spectrum of the field transmitted by the structure. To that end we Fourier transform the amplitude transmittance above, yielding

$$\mathcal{F}t_A(\xi, \eta) = 12\delta(f_X, f_Y) + m4\delta(f_X - \frac{1}{L}, f_Y) + m4\delta(f_X + \frac{1}{L}, f_Y).$$

$$\mathcal{F}\{t_A(\xi, \eta)\} = \frac{1}{2} \delta(f_X, f_Y) + \frac{m}{4} \delta\left(f_X - \frac{1}{L}, f_Y\right) + \frac{m}{4} \delta\left(f_X + \frac{1}{L}, f_Y\right).$$

(4-69)

Now the above transfer function has value unity at the origin, and when evaluated at frequencies $(f_X, f_Y) = (\pm \frac{1}{L}, 0)$ yields

$$H_{\pm 1L,0} = \exp - j\pi\lambda z L^2.$$

$$H\left(\pm \frac{1}{L}, 0\right) = \exp\left(-j\frac{\pi\lambda z}{L^2}\right).$$

(4-70)

Thus after propagation over distance z behind the grating, the Fourier transform of the field becomes

$$\mathcal{F}U(x,y,z) = 12\delta(f_X, f_Y) + m4e^{-j\pi\lambda z L^2} \delta(f_X - \frac{1}{L}, f_Y) + m4e^{-j\pi\lambda z L^2} \delta(f_X + \frac{1}{L}, f_Y).$$

$$\mathcal{F}\{U(x, y, z)\} = \frac{1}{2} \delta(f_X, f_Y) + \frac{m}{4} e^{-j\frac{\pi\lambda z}{L^2}} \delta\left(f_X - \frac{1}{L}, f_Y\right) + \frac{m}{4} e^{-j\frac{\pi\lambda z}{L^2}} \delta\left(f_X + \frac{1}{L}, f_Y\right).$$

Inverse transforming this spectrum we find the field at distance z from the grating to be given by

$$U(x, y, z) = 12 + m4e^{-j\pi\lambda z L^2} e^{j\frac{2\pi x}{L}} + m4e^{-j\pi\lambda z L^2} e^{-j\frac{2\pi x}{L}},$$

which can be simplified to

$$U(x, y, z) = 12 + me^{-j\pi\lambda z L^2} \cos(2\pi x L).$$

$$U(x, y, z) = \frac{1}{2} \left[1 + me^{-j\frac{\pi\lambda z}{L^2}} \cos\left(\frac{2\pi x}{L}\right) \right].$$

(4-71)

Finally, the intensity distribution is given by

$$I(x, y, z) = 141 + 2m \cos(\pi\lambda z L^2) \cos(2\pi x L) + m^2 \cos^2(2\pi x L).$$

$$I(x, y, z) = \frac{1}{4} \left[1 + 2m \cos\left(\frac{\pi\lambda z}{L^2}\right) \cos\left(\frac{2\pi x}{L}\right) + m^2 \cos^2\left(\frac{2\pi x}{L}\right) \right].$$

(4-72)

We now consider three special cases of this result that have interesting interpretations.

- Suppose that the distance z behind the grating satisfies $\pi\lambda z / L^2 = 2n\pi$ or $z = 2nL^2 / \lambda$, where n is an integer. Then the intensity observed at this distance behind the grating is

$$I(x, y, 2nL^2 / \lambda) = 141 + m \cos(2\pi x L)$$

$$I\left(x, y, \frac{2nL^2}{\lambda}\right) = \frac{1}{4} \left[1 + m \cos\left(\frac{2\pi x}{L}\right)\right]^2$$

which can be interpreted as a *perfect image* of the grating. That is, it is an exact replica of the intensity that would be observed just behind the grating. A multiplicity of such images appears behind the grating, without the help of lenses! Such images are called *Talbot images* (after the scientist who first observed them), or simply *self-images*. A good discussion of such images is found in [340].

2. Suppose that the observation distance satisfies $\pi\lambda z/L^2 = (2n+1)\pi$, or $z = (2n+1)L^2/\lambda$. Then

$$I(x, y, (2n+1)L^2/\lambda) = 141 - m \cos 2\pi x L^2.$$

$$I\left(x, y, \frac{(2n+1)L^2}{\lambda}\right) = \frac{1}{4} \left[1 - m \cos\left(\frac{2\pi x}{L}\right)\right]^2.$$

This distribution is also an image of the grating, but this time with a 180° spatial phase shift, or equivalently with a *contrast reversal*. This, too, is called a Talbot image.

3. Finally, consider distances satisfying $\pi\lambda z/L^2 = (2n-1)\pi/2$, or $z = n-12L^2/\lambda$. Then $\cos(\pi\lambda z/L^2) = 0$, and

$$I(x, y, n-12L^2/\lambda) = 141 + m^2 \cos 22\pi x L^2 = 141 + m^2 22 + m^2 22 \cos 4\pi x L^2.$$

$$\begin{aligned} I\left(x, y, \frac{(n-\frac{1}{2})L^2}{\lambda}\right) &= \frac{1}{4} \left[1 + m^2 \cos^2\left(\frac{2\pi x}{L}\right)\right] \\ &= \frac{1}{4} \left[\left(1 + \frac{m^2}{2}\right) + \frac{m^2}{2} \cos\left(\frac{4\pi x}{L}\right)\right]. \end{aligned}$$

This image has twice the frequency of the original grating and has reduced contrast. Such an image is called a *Talbot subimage*. Note that if $m \ll 1$, then the periodic image will effectively vanish at the subimage planes.

[Figure 4.18](#) shows the locations of the various types of images behind the original grating.

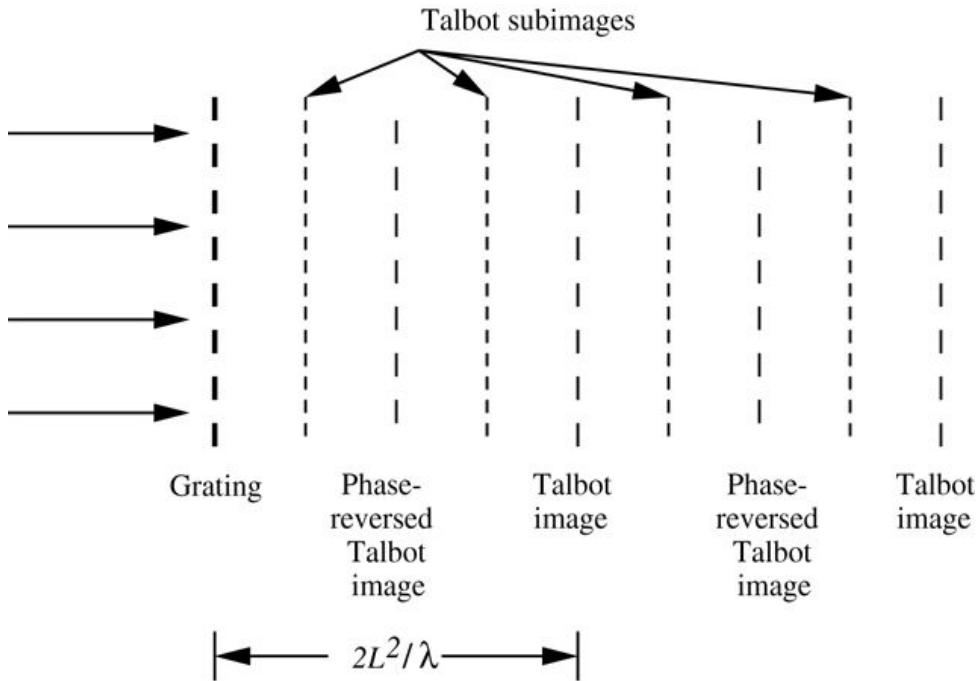


Figure 4.18

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 4.18 Locations of Talbot image planes behind the grating.

The illustration shows on the left extreme rightward rays entering a vertical grating. To the right of the grating is a series of vertical equally spaced parallel dotted lines representing images. Beginning with the one closest to the grating, the images are as follows. Talbot sub image, phase-reversed Talbot image, Talbot sub image, and Talbot image. The width covered by these four is labeled $2 \times L^2 / \lambda$. The pattern continues to the right.

We have initially neglected the finite extent of the grating, and now introduce it to determine its effects. Let the grating be bounded by a square aperture of width ℓ . Figure 4.19 shows the $0, +1 -1$ orders propagating away from the grating when its width is ℓ , under the assumption that the distances are sufficiently short that diffraction spreading of the orders have not yet taken place. The distance z_0 is the location beyond which the orders no longer overlap and interference between them no longer takes place, thus eliminating the Talbot images. With a paraxial approximation, the distance z_0 is given by

$$z_0 = \ell L \lambda,$$

$$z_0 = \frac{\ell L}{\lambda},$$

(4-73)

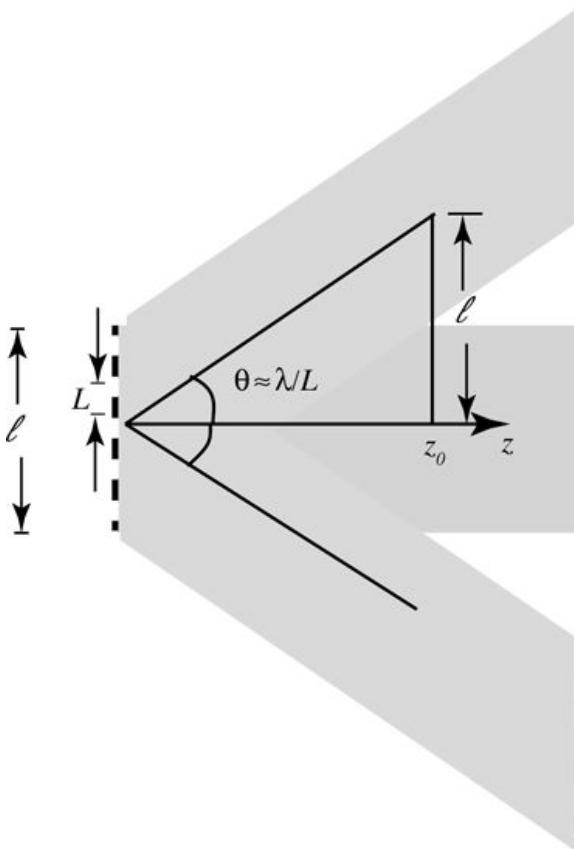


Figure 4.19

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 4.19 Overlap of the $-1, 0, +1$ diffraction orders deep in the Fresnel zone.

The illustration shows a rightward horizontal axis z originating at a grating of width lowercase ℓ at period uppercase L . Beginning at the left extreme of the horizontal axis at the aperture, an upward sloping line makes angle theta with the horizontal axis, where theta approximately equals lambda over L . From a point on the sloping line, a perpendicular is dropped at z subscript 0 on the horizontal axis, the height of the perpendicular is marked lowercase l .

and therefore to observe Talbot images we require $z < z_0 = \ell L / \lambda$. We can also state this requirement equivalently as

$$NF > 0.25 \ell L,$$

$$N_F > 0.25 \frac{\ell}{L},$$

(4-74)

where again N_F is the Fresnel number, $(\ell/2)^2 / (\lambda z)$.

The width W of the diffraction orders is approximately

$$W \approx \ell + \lambda z / \ell,$$

$$W \approx \ell + \lambda z / \ell,$$

(4-75)

where the first term is the grating width and the second term is the diffraction spreading beyond that width. Thus for the diffraction orders to still be the same width as the grating requires $\lambda z / \ell \ll \ell$, or equivalently

$$NF > 0.25.$$

$$N_F > 0.25.$$

(4-76)

Thus, since ℓ/L is always greater than one, the dominant requirement is that of (4-74), i.e. $z < z_0$, in which case the overlap of orders is as depicted in Fig. 4.19.

The Talbot image phenomenon is much more general than just the particular case analyzed here. It can be shown to be present for *any* periodic structure (see Prob. 4-20).

4.6 Beam Optics

In this section we follow the development of [305], where more details can be found. In many cases in practice, it is possible to represent the phasor field by a function with a slowly varying complex envelope $V(x, y, z)$ times a rapidly changing phase factor $\exp(jkz)$, i.e.

$$\begin{aligned} U(x, y, z) &= V(x, y, z) \exp(jkz), \\ U(x, y, z) &= V(x, y, z) \exp(jkz), \end{aligned} \tag{4-77}$$

the assumption being that $V(x, y, z)$ changes negligibly over distances as small as λ . If a solution of such a form is substituted into the Helmholtz equation, and the slowly varying

assumption is applied ($\partial^2 \partial z^2 V \ll j2k \partial V \partial z$) $\left(\frac{\partial^2}{\partial z^2} V \ll j2k \frac{\partial V}{\partial z} \right)$, we obtain a differential equation that $V(x, y, z)$ must satisfy:

$$\begin{aligned} \nabla_t^2 V + j2k \frac{\partial V}{\partial z} &= 0, \\ \nabla_t^2 V + j2k \frac{\partial V}{\partial z} &= 0. \end{aligned} \tag{4-78}$$

Here the symbol ∇_t^2 represents the transverse Laplacian, $\nabla_t^2 = \partial^2 / \partial x^2 + \partial^2 / \partial y^2$. This equation is known as the *paraxial Helmholtz equation*.

It is a simple matter to show (cf. [Prob. 3-7](#)) that one type of solution to such an equation is a paraxial spherical wave,

$$\begin{aligned} V(x, y, z) &= V_1 z \exp(jkx^2 + ky^2 - kz^2), \\ V(x, y, z) &= \frac{V_1}{z} \exp \left[jk \frac{x^2 + y^2}{2z} \right], \end{aligned} \tag{4-79}$$

where V_1 is a constant. However, as we shall see, a more interesting solution is the *Gaussian beam*, to be described now.

4.6.1 Gaussian Beams

An alternative solution to the paraxial Helmholtz equation is one of the form

$$V(x, y, z) = V_1 q(z) \exp[jkx^2 + y^2/2q(z)],$$

$$V(x, y, z) = \frac{V_1}{q(z)} \exp \left[jk \frac{x^2 + y^2}{2q(z)} \right],$$

(4-80)

where $q(z)$ is a *complex-valued* constant of the form

$$q(z) = z - jz_0.$$

$$q(z) = z - jz_0.$$

(4-81)

$q(z)$ is known as the q -parameter and z_0 as the Rayleigh range.

If we define two real-valued functions $R(z)$ and $W(z)$ such that

$$q(z) = 1/R(z) + j\lambda\pi W(z),$$

$$\frac{1}{q(z)} = \frac{1}{R(z)} + j\frac{\lambda}{\pi W^2(z)},$$

(4-82)

then with an appropriate substitution, we find the complex envelope of the field to be

$$V(x, y, z) = V_1 R(z) + j\lambda\pi W(z) \exp[-r^2/2W^2(z)] \exp[jk(r^2/2R(z))],$$

$$V(x, y, z) = V_1 \left(\frac{1}{R(z)} + j\frac{\lambda}{\pi W^2(z)} \right) \exp \left[-\frac{r^2}{W^2(z)} \right] \exp \left[jk \frac{r^2}{2R(z)} \right],$$

(4-83)

where $r = \sqrt{x^2 + y^2}$. As can be seen from this expression, $W(z)$ is the $1/e$ half-width of a Gaussian amplitude profile, while $R(z)$ is the radius of curvature of the wavefront. With appropriate definitions to be listed and considerable algebra, the phasor complex field can now be written

$$U(x, y, z) = V_0 W(z) \exp[-r^2/2W^2(z)] \exp[jkz + jkr^2/2R(z) - j\psi(z)],$$

$$U(x, y, z) = V_0 \frac{W_0}{W(z)} \exp \left[-\frac{r^2}{W^2(z)} \right] \exp \left[jkz + jk \frac{r^2}{2R(z)} - j\psi(z) \right],$$

(4-84)

where

$$V_0 = jV_1/z_0, \quad (4-85)$$

$$W(z) = W_0 \sqrt{1 + \left(\frac{z}{z_0}\right)^2}, \quad (4-86)$$

$$W_0 = \sqrt{\frac{\lambda z_0}{\pi}}, \quad (4-87)$$

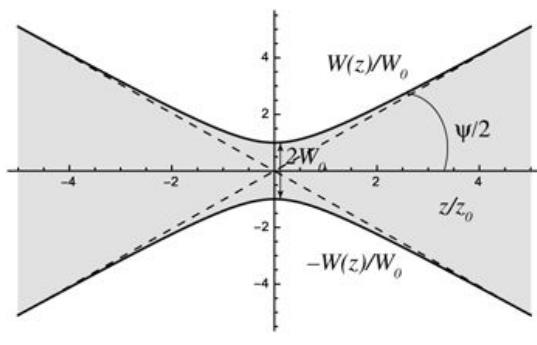
$$R(z) = z \left[1 + \left(\frac{z_0}{z}\right)^2 \right], \quad (4-88)$$

$$\psi(z) = \tan^{-1} \frac{z}{z_0}. \quad (4-89)$$

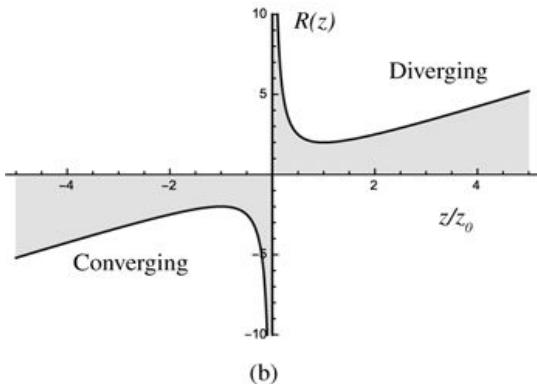
[Figure 4.20](#) shows plots of (a) $\pm W(z)/W_0 \pm W(z)/W_0$, (b) $R(z)$ and (c) $\psi(z)$, all as a function of z/z_0 . Note in particular that W_0 represents the *radius of the beam waist at focus*, and that waist occurs at $z/z_0=0$. The radius of curvature is negative to the left of focus, implying a converging wave, and positive to the right of focus corresponding to a diverging wave. At focus the radius of curvature becomes infinite, implying that in the plane of focus the wavefront is planar and perpendicular to the z -axis. Lastly, the phase $\psi(z)$, which is known as the *Gouy phase*, undergoes a sign change as the wave passes through focus. The phase starts at $z/z_0=-\infty$ with value $-\pi/2$ and approaches value $+\pi/2$ on the far right as the wave diverges towards $z/z_0=\infty$. Thus the total phase shift approaches π . The distance within which the area of the cross-section of the beam is no more than twice its cross-section at focus is defined as the *depth of focus* of the beam and is given by twice the Rayleigh range, i.e. $2z_0$. The full angle of divergence of the beam is

$$\Psi = 4\pi \lambda / 2W_0.$$

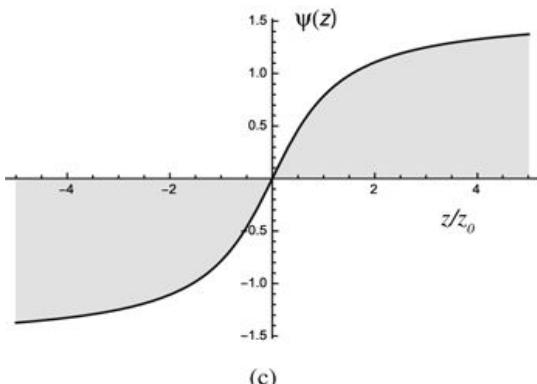
$$(4-90)$$



(a)



(b)



(c)

Figure 4.20

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 4.20 (a) Normalized beam width, (b) radius of curvature, and (c) Gouy phase as a function of z/z_0 .

Graph a for normalized beam width shows a horizontal and a vertical axis, each ranging from minus 5 to +5. Two dotted lines intersect at the origin, one passing from the first quadrant to the third and the other, from the second quadrant to the fourth, such that both the axes mirror the lines on to the adjoining quadrant. The U-shaped graph for $W(z)/W_0$ overlaps the dotted lines in the first and second quadrants but veers away from the lines as it approaches the origin to form a blunt vertex near the 1 mark on the vertical axis. An arc from the curve to the horizontal axis is marked $\psi/2$. The graph for $-W(z)/W_0$ is a mirror image of the graph for $W(z)/W_0$. Graph b for normalized radius

of curvature shows a horizontal axis ranging from minus 5 to +5 and a vertical axis $R(z)$ ranging from minus 10 and +10. The diverging curve begins in the first quadrant near the +10 mark on the vertical axis and follows a steep downward slope up in a rightward direction to a point around (1, 2), thereafter it takes a gentle upward sloping path. The area between the curve and the horizontal axis is shaded. The curve is mirrored through the origin into the third quadrant to mark the converging curve, which begins near the minus 10 mark on the vertical axis, rises leftward steeply to a point around (minus 1, minus 2) and then slopes gently in the leftward direction. The area between the curve and the horizontal axis is shaded. Graph C for normalized Gouy phase as a function of z/z subscript 0 shows a horizontal axis ranging from minus 5 to +5 and a vertical axis $\psi(z)$ ranging from minus 1.5 to + 1.5. The curve begins in the first quadrant at (5, 1.4) and slopes gently downward in the leftward direction to reach the origin. It continues into the third quadrant mirroring the first quadrant segment through the origin.

Finally, the intensity profile of the beam at all z positions is Gaussian in shape,

$$I(x,y,z)=I_0 W(z) 2 \exp -2(x^2+y^2) W^2(z),$$

$$I(x, y, z) = I_0 \left[\frac{W_0}{W(z)} \right]^2 \exp \left[-\frac{2(x^2 + y^2)}{W^2(z)} \right].$$

(4-91)

$$\text{where } I_0 = |V_0|^2.$$

The Gaussian beam is a mode of a spherical resonator that has spherical end mirrors with radii of curvature that match the radii of curvature of the Gaussian beam at each mirror location.

4.6.2 Hermite-Gaussian Beams

A beam closely related to the Gaussian beam is the Hermite-Gaussian beam. Such a beam shares the wavefronts and divergence of a Gaussian beam, but has a different intensity distribution. In particular, Hermite-Gaussian beams are useful in describing a wider set of modes of an optical resonator than just the simple zero-order Gaussian beam.

The modes of an optical resonator are complex beam profiles that reproduce themselves after reflection from two end mirrors, those mirrors generally being spherical in shape. For a wave to be a mode of the resonator, it must reproduce itself after round-trip passage through the resonator, where in particular both the amplitude and the phase must be reproduced.

If we assume that the Gaussian mode is modified by a separable set of multiplicative amplitude profiles, one in x and one in y , and require that the resulting wave satisfies the paraxial Helmholtz equation, separation of variables yields a set of three ordinary differential equations, one for each of these three variables (x, y, z) . The modal amplitude modulations are solutions to an eigenvalue problem, and are found to be Hermite polynomials that modify the amplitude of the beam (see [Section 3.3](#) of [305] for details). The Hermite polynomials are defined by the recursion relation

$$H_{l+1}(u) = 2u H_l(u) - 2l H_{l-1}(u)$$

$$H_{l+1}(u) = 2u H_l(u) - 2l H_{l-1}(u)$$

(4-92)

with

$$H_0(u)=1, H_1(u)=2u.$$

$$H_0(u) = 1 \quad H_1(u) = 2u.$$

(4-93)

The Hermite-Gaussian functions are defined by

$$G_l(u)=H_l(u)\exp(-u^2/2).$$

$$G_l(u) = H_l(u) \exp\left(-u^2/2\right).$$

(4-94)

The Hermite-Gaussian mode amplitudes can then be written

$$U_{l,m}(x,y,z)=A_{l,m}W_0W(z)G_l(x)G_m(y)\exp(jkz+jkrR(z)-j(l+m+1)\psi(z),$$

$$\begin{aligned} U_{l,m}(x, y, z) &= A_{l,m} \left[\frac{W_0}{W(z)} \right] G_l\left(\frac{\sqrt{2}x}{W(z)}\right) G_m\left(\frac{\sqrt{2}y}{W(z)}\right) \\ &\times \exp\left[jkz + jk\frac{r^2}{2R(z)} - j(l+m+1)\psi(z)\right], \end{aligned}$$

(4-95)

where $A_{l,m}$ is a constant and again $r=x^2+y^2$. Note that the Guoy phase for the (l,m) th mode now varies between $-(l+m+1)\pi/2$ and $+(l+m+1)\pi/2$, while the term $j\theta$ indicates a helical tilt of the wavefront as the wave travels in the z direction. The intensity associated with the (l,m) th mode is of course the squared magnitude of $U_{l,m}(x,y,z)$,

$$I_{l,m}(x,y,z)=A_{l,m}^2W_0^2W(z)^2G_l^2(x)G_m^2(y)\exp(-2jkz-2jkrR(z)+2j(l+m+1)\psi(z)).$$

$$I_{l,m}(x, y, z) = |A_{l,m}|^2 \left[\frac{W_0}{W(z)} \right]^2 G_l^2\left(\frac{\sqrt{2}x}{W(z)}\right) G_m^2\left(\frac{\sqrt{2}y}{W(z)}\right)$$

(4-96)

[Figure 4.21](#) shows density plots of the intensity distributions for a few lower-order modes. Note that the widths of the modes broaden as the index increases, and the mode complexity increases with increases of indices. The Hermite-Gaussian functions form a complete basis set for expanding other forms of wave-amplitude profile that satisfy the paraxial Helmholtz equation.

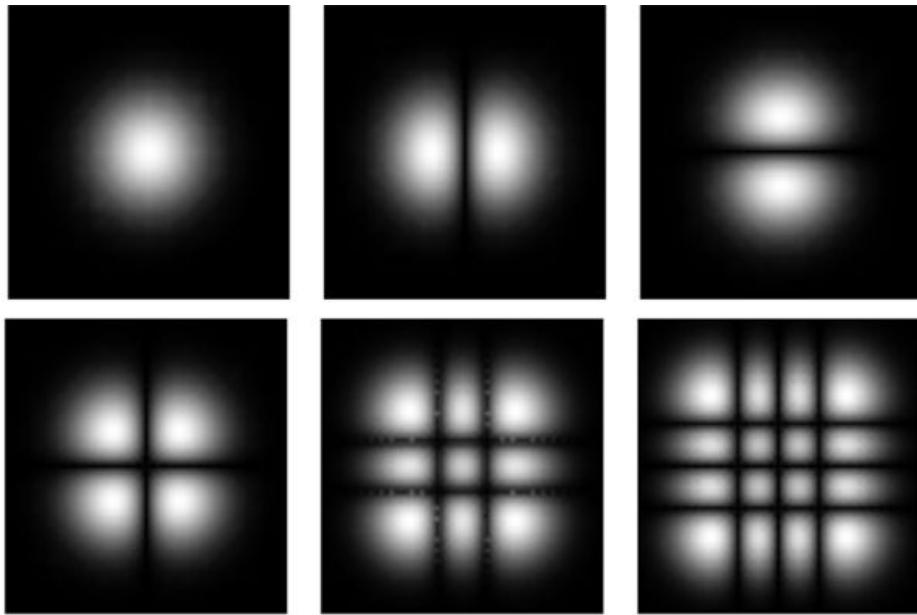


Figure 4.21

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 4.21 Mode intensities for Hermite-Gaussian modes numbered (left to right, top row) (0, 0), (1, 0) (0, 1) and (bottom row) (1, 1), (2, 2) (3, 3).

The six images show patterns of brightness, each in a square area of darkness. In the top row, the first image is of a single bright circular spot in the center. The second image shows the spot vertically split at the center by a dark strip. The third image shows the spot horizontally split at the center by a dark strip. In the bottom row, the first image shows the spot split in quarters through the center both horizontally and vertically. The second image shows two horizontal dark strips intersecting two vertical ones, thus dividing the bright spot into 9 parts including a square at the center. The third image shows three horizontal dark strips intersecting three vertical ones, thus dividing the bright spot into 16 parts including a square at the center split into four squares.

4.6.3 Laguerre-Gaussian Beams

If the paraxial Helmholtz equation is rewritten in cylindrical coordinates (r, θ, z) , another set of eigensolutions can be found. Separation of variables in r and θ yields a set of beams with complex amplitudes given by

$$U_{l,p}(r, \theta, z) = A_{l,p} W_0 W(z) 2r W(z) |l| L_p^{|l|} |l|! r^{2|l|} \exp(-r^2/2W(z)^2) \times \exp(jkz + jkr^2/2W(z)^2) \exp(jl\theta - j(|l|+2p+1)\psi(z)).$$

$$U_{l,p}(r, \theta, z) = A_{l,p} \left[\frac{W_0}{W(z)} \prod_{n=1}^{|l|} \frac{\sqrt{2r}}{W(z)} \right]^{l!} L_p^{|l|} \left(\frac{2r^2}{W^2(z)} \right) \exp \left[-\frac{r^2}{W^2(z)} \right] \times \exp \left[jkz + jk \frac{r^2}{2R(z)} + jl\theta - j(|l|+2p+1)\psi(z) \right].$$

(4-97)

Here $L_p^{||l|}$ is the *generalized Laguerre polynomial* (also known as an *associated Laguerre polynomial*) of order (l, p) , while the parameters $W_0, W(z)$ and $R(z)$ are identical with those defined previously for the Gaussian beam. The parameter $p \geq 0$ is called the radial index while the parameter $|l|$ is the azimuthal index. The Gouy phase is seen to vary from $-(|l|+2p+1)(\pi/2) - (|l| + 2p + 1)(\pi/2)$ to $+(|l|+2p+1)(\pi/2) + (|l| + 2p + 1)(\pi/2)$. The modes are usually normalized by constants that equalize their total power [316],

$$A_{l,p} = 2p(1+80l)\pi(p+|l|),$$

$$A_{l,p} = \sqrt{\frac{2p}{(1+\delta_{0l})\pi(p+|l|)}},$$

(4-98)

where δ_{0l} is a Kronecker delta function. The intensity of the mode,

$$I_{l,p}(r, \theta, z) = A_{l,p}^2 W_0^2 W(z)^2 r^2 W^2(z)^2 \exp(-2r^2/W^2(z)),$$

$$I_{l,p}(r, \theta, z) = |A_{l,p}|^2 \left[\frac{W_0}{W(z)} \right]^2 \left[\frac{2r^2}{W^2(z)} \right]^{||l|} \left[L_p^{||l|} \left(\frac{2r^2}{W^2(z)} \right) \right]^2 \exp \left(-\frac{2r^2}{W^2(z)} \right),$$

(4-99)

is seen to depend only on r^r ; thus the intensities of all modes of the Laguerre-Gaussian solution are circularly symmetric.

An interesting property of these modes arises from the phase term that is proportional to the azimuthal angle θ^θ . The presence of this term implies that, when $l > 0$, the wavefront of the beam has a helical twist, the amount of the twist increasing with the azimuthal index $|l|$. Such a wave can be shown to carry *orbital angular momentum* (see [305], p. 454) and to impart a torque to any object it strikes. When phase circulates around a zero of intensity, we have what is known as an “optical vortex,” or a “phase vortex.” Such modes also play a role in the synthesis of point-spread functions that rotate as a function of defocus (see [Section 8.2](#)).

[Figure 4.22](#) shows several intensity profiles and phase profiles of low-order Laguerre-Gaussian modes. As the mode indices increase the modes become broader, as does also the quantity $W(z)$.

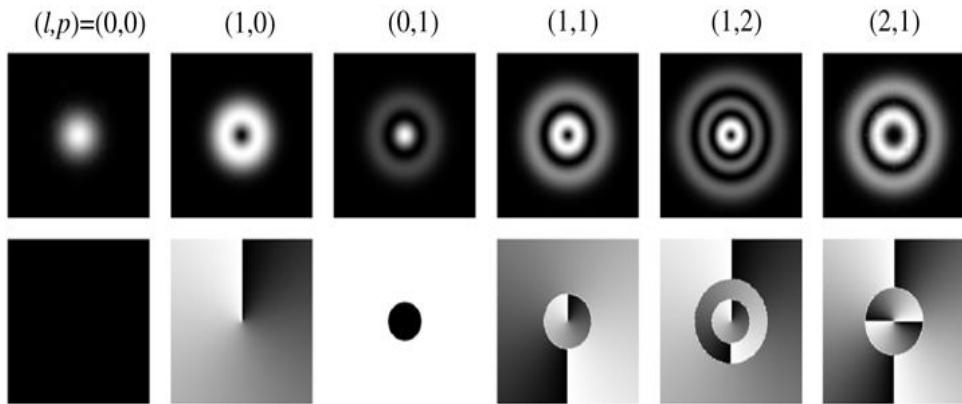


Figure 4.22

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 4.22 Modal intensity and phase distributions for Laguerre-Gaussian beams with various mode numbers. The top row shows the modal intensities, and the bottom row shows the corresponding modal phase distributions, with black representing 0 radians and white 2π radians. The abrupt transitions of the phase from white to black are a result of phase wrapping into the primary interval $(0, 2\pi)$.

“LAT: The pair for $(l, p) = (0, 0)$: Image 1 shows a small bright spot at the center of darkness. Image 2 shows a patch of darkness. The pair for $(l, p) = (1, 0)$: Image 1 shows a large bright spot at the center of darkness. At the center of the bright spot is a small dark spot. Image 2 shows a vertical line from the center to the top end. To its left, in the top left corner, is total brightness and as we move in the counterclockwise direction the intensity reduces. Thus eventually the top right corner is total darkness.

The pair for $(l, p) = (0, 1)$: Image 1 shows a small bright spot at the center surrounded by a ring of darkness, around which is a ring of relatively less darkness. Image 2 shows a spot of darkness in a background of brightness,

The pair for $(l, p) = (1, 1)$: Image 1 shows a small dark spot at the center surrounded by a ring of brightness followed by a ring of darkness, around which is a ring of relatively less darkness. Image 2 shows a circle in the center that goes from intense brightness in the top left quarter to total darkness in the top right quarter as we move counterclockwise from a vertical radius. The background goes from intense brightness in the bottom right quarter to total darkness in the bottom left quarter as we move counterclockwise from a perpendicular at the center of the lower side.

The pair for $(l, p) = (1, 2)$: Image 1 is the same as that in the pair for $(l, p) = (1, 1)$ with an additional ring of darkness followed by one of relative darkness. Image 2 shows a ring in the center that goes from intense brightness in the bottom right quarter to total darkness in the bottom left quarter as we move counterclockwise from a vertical line at the ring's lowest point. The background, which is seen around and inside the ring, goes from intense brightness in the top left quarter to total darkness in the top right quarter as we move counterclockwise from a perpendicular at the center of the upper half.

The pair for $(l, p) = (2, 1)$: Image 1 is the same as that in the pair for $(l, p) = (1, 1)$ with the rings being wider. Image 2 shows a background that is vertically split in halves. The left half goes from intense brightness at the top to total darkness at the bottom; the same is shown in the right half but in the opposite direction. A circle is set in the center. The lower semicircle goes from intense brightness to total darkness as we move counterclockwise; the same is shown in the upper semicircle.”

4.6.4 Bessel Beams

Bessel beams are idealized solutions to the full Helmholtz equation (no paraxial approximation) that have the remarkable property that they do not diverge with propagation in the z direction. To derive the form of a Bessel beam, we start by considering an idealized angular spectrum of the scalar field in plane $z=0$, one that consists of a circular annulus of equal-strength Fourier components, where the annulus is sufficiently thin that it can be represented by a delta function in radius in the Fourier plane. That is, we assume an angular spectrum for the beam of the form (cf. [Section 3.10.1](#))

$$\begin{aligned} A(f_X, f_Y; 0) &= A_0 \delta(\rho - \rho_0), \\ A(f_X, f_Y; 0) &= A_0 \delta(\rho - \rho_0), \end{aligned} \quad (4-100)$$

where $f_X = \alpha/\lambda$, $f_Y = \beta/\lambda$, $\rho = \sqrt{f_X^2 + f_Y^2}$, A_0 is a constant, and ρ_0 is a fixed radial frequency with value less than $1/\lambda$. The spatial distribution of field that corresponds to such a spectrum is found by performing an inverse Fourier-Bessel transform,

$$\begin{aligned} U(x, y, 0) &= 2\pi A_0 \int_0^\infty \rho \delta(\rho - \rho_0) J_0(2\pi\rho r) d\rho = 2\pi\rho_0 A_0 J_0(2\pi\rho_0 r), \\ U(x, y, 0) &= 2\pi A_0 \int_0^\infty \rho \delta(\rho - \rho_0) J_0(2\pi\rho r) d\rho \\ &= 2\pi\rho_0 A_0 J_0(2\pi\rho_0 r), \end{aligned} \quad (4-101)$$

where $r = \sqrt{x^2 + y^2}$.

Now we ask what happens to this field distribution as the wave propagates in the z direction? To answer the question we return to the frequency domain. We know that free-space propagation can be represented by a transfer function

$$H(f_X, f_Y) = \exp[j2\pi z \lambda^{-1} - (\lambda f_X)^2 - (\lambda f_Y)^2].$$

$$H(f_X, f_Y) = \exp[j2\pi \frac{z}{\lambda} \sqrt{1 - (\lambda f_X)^2 - (\lambda f_Y)^2}].$$

(4-102)

If we multiply this transfer function times the angular spectrum of [Eq. \(4-100\)](#) we obtain a new angular spectrum,

$$A(f_X, f_Y; z) = A_0 \delta(\rho - \rho_0) \exp[j2\pi \frac{z}{\lambda} \sqrt{1 - \lambda^2 \rho_0^2}].$$

$$A(f_X, f_Y; z) = A_0 \delta(\rho - \rho_0) \exp[j2\pi \frac{z}{\lambda} \sqrt{1 - \lambda^2 \rho_0^2}].$$

(4-103)

Since ρ_0 is a constant, we see that the effect of propagation is simply to apply a z -dependent phase factor to the entire spectral ring, the phase factor being identical for all spatial frequencies in that ring. The field at distance z is thus seen to be

$$U(x,y,z)=2\pi\rho_0A_0J_0(2\pi\rho_0r)\exp j2\pi z\lambda 1-\lambda 2\rho_0 2.$$

$$U(x, y, z) = 2\pi\rho_0 A_0 J_0(2\pi\rho_0 r) \exp \left[j2\pi \frac{z}{\lambda} \sqrt{1 - \lambda^2 \rho_0^2} \right].$$

(4-104)

The intensity of the field,

$$I(x,y,z)=(2\pi\rho_0A_0)2J_02(2\pi\rho_0r),$$

$$I(x, y, z) = (2\pi\rho_0 A_0)^2 J_0^2(2\pi\rho_0 r),$$

(4-105)

is seen to be independent of z , and therefore the beam propagates without divergence.

The Bessel beam is idealized in several respects. First, it is simple to show that it contains infinite energy. Second, it extends over the entire infinite plane; truncation destroys the lack of divergence with propagation. Finally, the infinitely thin (and infinitely tall) angular spectrum ring can never be realized exactly in practice. A ring of finite width does not exhibit the ideal property of zero divergence, although with proper design, the divergence can be made small. See [145] for more details.

Problems - Chapter 4

1. 4-1. Consider the quadratic-phase exponential $1j\lambda z \exp(j\pi\lambda z(x^2+y^2))$.
1. Show that the volume (with respect to x and y) under this function is unity.
 2. Show that the two-dimensional quadratic-phase sinusoidal part of this function contributes all of the volume and the two-dimensional quadratic-phase cosinusoidal part contributes none of the volume.
- (Hint: Make use of [Table 2.1](#).)
2. 4-2. Consider a spherical wave expanding about the point $(0, 0, -z_1)$ in a Cartesian coordinate system. The wavelength of the light is λ , and $z_1 > 0$.
1. Express the phase distribution of the spherical wave across an (x, y) plane located normal to the z axis at coordinate $z=0$.
 2. Using a paraxial approximation, express the phase distribution of the quadratic-phase wavefront that approximates this spherical wavefront.
 3. Find an exact expression for the phase by which the spherical wavefront *lags* or *leads* the quadratic-phase wavefront. Does it lag or lead?
3. 4-3. Consider a spherical wave converging toward the point $(0, 0, +z_1)$ in a Cartesian coordinate system. The wavelength of the light is λ and $z_1 > 0$.
1. Express the phase distribution of the spherical wave across an (x, y) plane located normal to the z axis at coordinate $z=1$.
 2. Using a paraxial approximation, express the phase distribution of the quadratic-phase wavefront that approximates this spherical wavefront.
 3. Find an exact expression for the phase by which the spherical wavefront *lags* or *leads* the quadratic-phase wavefront. Does it lag or lead?
4. 4-4. Recalling that the Fresnel number of an aperture of width ℓ is defined by $N_F = (\ell/2)^2 / (\lambda z)$,
1. Show that the antenna designer's formula for the condition for Fraunhofer diffraction can be expressed as $N_F < 1/8$.
 2. Show that a condition $k^2 z \xi^2 + \eta^2 \max < \pi^2 (\xi^2 + \eta^2)_{\max} < \frac{\pi^2}{2}$ for Fraunhofer diffraction, where ξ and η are rectangular coordinates in the plane of the aperture, can be expressed as $N_F < 1/2$.
5. 4-5. Fresnel propagation over a sequence of successive distances z_1, z_2, \dots, z_n must be equivalent to Fresnel propagation over the single distance

$z = z_1 + z_2 + \dots + z_n$. Find a simple proof that this is the case. Repeat for propagation without the Fresnel approximation.

6. 4-6. Show that the top “transition region” shown in Fig. 4.4 is bounded by the parabola $(\ell/2 - x)^2 = 4\lambda z$ and the bottom transition region by $(\ell/2 + x)^2 = 4\lambda z$, where the aperture is ℓ wide, the origin of the coordinates is at the center of the aperture, z is the distance from the plane of the aperture, and x is the vertical coordinate throughout the figure.
7. 4-7. Using the ray-transfer matrix for free-space propagation (Eq. (B-10)), show that the general one-dimensional Huygens-Fresnel impulse response of Eq.(4-35) reduces to that of Eq.(4-34) when propagation is through free space.
8. 4-8. A spherical wave is converging toward a point $(0, 0, z_1)$ to the right of a circular aperture of radius R , centered on $(0, 0, 0)$. The wavelength of the light is λ . Consider the field observed at an arbitrary point (axial distance z) to the right of the aperture. Show that the wavefront error made in a paraxial approximation of the illuminating spherical wave and the error incurred by using a quadratic-phase approximation in the Fresnel diffraction equation partially cancel one another. Under what condition does complete cancellation occur?
9. 4-9. Assuming unit-amplitude, normally incident plane-wave illumination:

1. Find the intensity distribution in the Fraunhofer diffraction pattern of the double-slit aperture shown in Fig. P4.9.
2. Sketch normalized cross sections of this pattern that appear along the x and y axes in the observation plane, assuming $X/\lambda z = 10 \text{ m}^{-1}$, $Y/\lambda z = 1 \text{ m}^{-1}$, $\Delta/\lambda z = 3/2 \text{ m}^{-1}$, z being the observation distance and λ the wavelength.

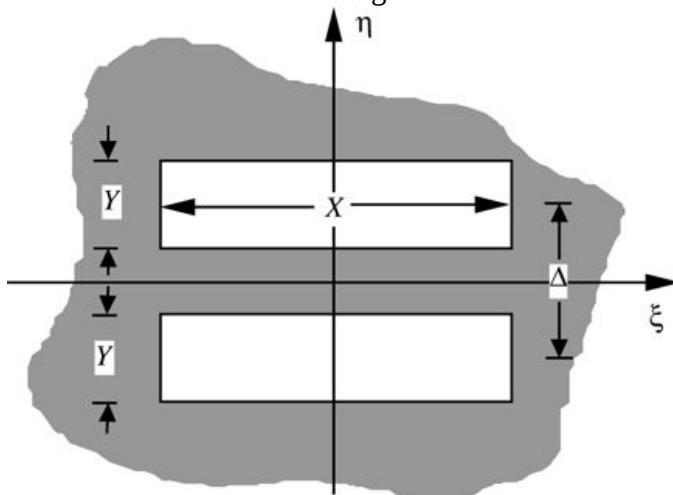


Figure P4.9
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure P4.9

The illustration shows an aperture along the horizontal axis ξ with two rectangular slits, one above the axis and the other below it, each X units long and Y units wide. The slits are aligned. The vertical axis η passes through the center of the two slits.

10. 4-10.

1. Sketch the aperture described by the amplitude transmittance function

$$t_A(\xi, \eta) = \text{rect}(\xi) \text{rect}(\eta) * 1 \Delta \text{comb}(\eta/\Delta) \delta(\xi) \text{rect}(\eta/N\Delta)$$

$$t_A(\xi, \eta) = \left[\text{rect}\left(\frac{\xi}{X}\right) \text{rect}\left(\frac{\eta}{Y}\right) \right] * \left[\frac{1}{\Delta} \text{comb}\left(\frac{\eta}{\Delta}\right) \delta(\xi) \right] \text{rect}\left(\frac{\eta}{N\Delta}\right)$$

where N is an odd integer and $\Delta > Y$.

2. Find an expression for the intensity distribution in the Fraunhofer diffraction pattern of that aperture, assuming illumination by a normally incident plane wave.
3. What relationship between Y and Δ can be expected to minimize the strength of the even-order diffraction components while leaving the zero-order component approximately unchanged?
11. 4-11. Find an expression for the intensity distribution in the Fraunhofer diffraction pattern of the aperture shown in [Fig. P4.11](#). Assume unit-amplitude, normally incident plane-wave illumination. The aperture is square and has a square central obscuration.

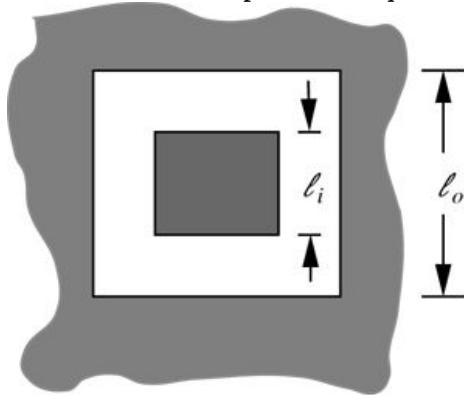


Figure P4.11

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure P4.11

12. 4-12. Find an expression for the intensity distribution in the Fraunhofer diffraction pattern of the aperture shown in [Fig. P4.12](#). Assume unit-amplitude, normally incident plane-wave

illumination. The aperture is circular and has a circular central obscuration.

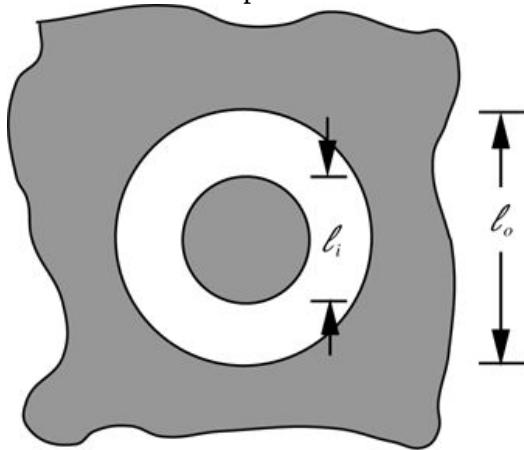


Figure P4.12

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure P4.12

13. 4-13. Two discrete spectral lines of a source are said to be “just resolved” by a diffraction grating if the peak of the q^q th-order diffraction component due to source wavelength $\lambda_1 \lambda_1$ falls exactly on the first zero of the q^q th-order diffraction component due to source wavelength $\lambda_2 \lambda_2$. The *resolving power* of the grating is defined as the ratio of the mean wavelength $\lambda \lambda$ to the minimum resolvable wavelength difference $\Delta\lambda \Delta\lambda$. Show that the resolving power of the sinusoidal phase grating discussed in this chapter is

$$\lambda\Delta\lambda = q\ell f_0 = qM$$

$$\frac{\lambda}{\Delta\lambda} = q\ell f_0 = qM$$

where q^q is the diffraction order used in the measurement, $\ell \ell$ is the width of the square grating, and $M M$ is the number of spatial periods of the grating contained in the aperture. What phenomenon limits the use of arbitrarily high diffraction orders?

14. 4-14. Calculate the diffraction efficiency into the first diffraction order for a grating with amplitude transmittance given by

$$tA(\xi) = \cos\pi\xi L$$

$$t_A(\xi) = \left| \cos\left(\frac{\pi\xi}{L}\right) \right|$$

15. 4-15. The amplitude transmittance function of a thin square-wave absorption grating is shown in [Fig. P4.15](#). Find the following properties of this grating:

1. The fraction of incident light that is absorbed by the grating.
2. The fraction of incident light that is transmitted by the grating.
3. The fraction of light that is transmitted into a single first order.

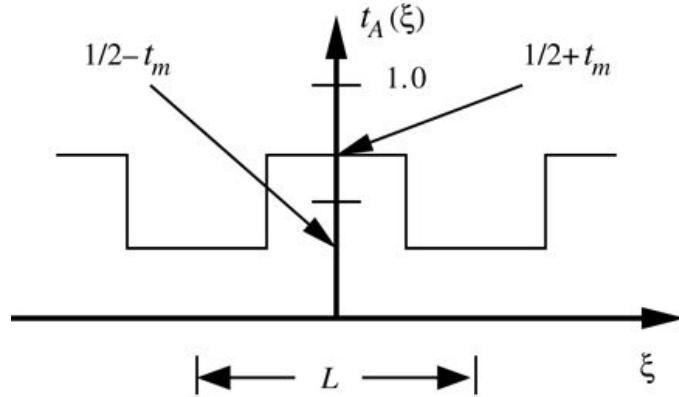


Figure P4.15

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure P4.15

The illustration shows horizontal axis ξ and a vertical axis $t_A(\xi)$. A series of continuous horizontal and vertical lines at right angles runs between the levels of points on the vertical axis marked $1/2 + t_m$ and $1/2 - t_m$. The path is as follows. To the right of the vertical axis, beginning at $1/2 + t_m$, is a horizontal line that drops vertically to the $1/2 - t_m$ level and then runs horizontally rightward and then rises vertically to the original level of $1/2 + t_m$ and then runs horizontally rightward again. This path is mirrored through the vertical axis onto the other side. The letter L marks the distance between the midpoints of the two horizontal segments on either sides of the vertical axis.

16. 4-16. A thin square-wave *phase* grating has a thickness that varies periodically (period L) such that the phase of the transmitted light jumps between 0 radians and ϕ radians.

1. Find the diffraction efficiency of this grating for the first diffraction orders.

2. What value of ϕ yields the maximum diffraction efficiency, and what is the value of that maximum efficiency?

17. 4-17. A “sawtooth” phase grating is periodic with period L and has a distribution of phase within one period from 0 to L given by

$$\phi(\xi) = 2\pi\xi L.$$

$$\phi(\xi) = \frac{2\pi\xi}{L}.$$

1. Find the diffraction efficiencies of all of the orders for this grating.
2. Suppose that the phase profile of the grating is of the more general form

$$\phi(\xi) = \phi_0 \xi L.$$

$$\phi(\xi) = \frac{\phi_0 \xi}{L}.$$

Find a general expression for the diffraction efficiency into all the orders of this new grating.

18. 4-18. An aperture Σ in an opaque screen is illuminated by a spherical wave converging towards a point P located in a parallel plane a distance z behind the screen, as shown in Fig. P4.18.

1. Find a quadratic-phase approximation to the illuminating wavefront in the plane of the aperture, assuming that the coordinates of P in the (x, y) plane are $(0, Y)$.
2. Assuming *Fresnel* diffraction from the plane of the aperture to the plane containing P , show that in the above case the observed intensity distribution is the *Fraunhofer* diffraction pattern of the aperture, centered on the point P .

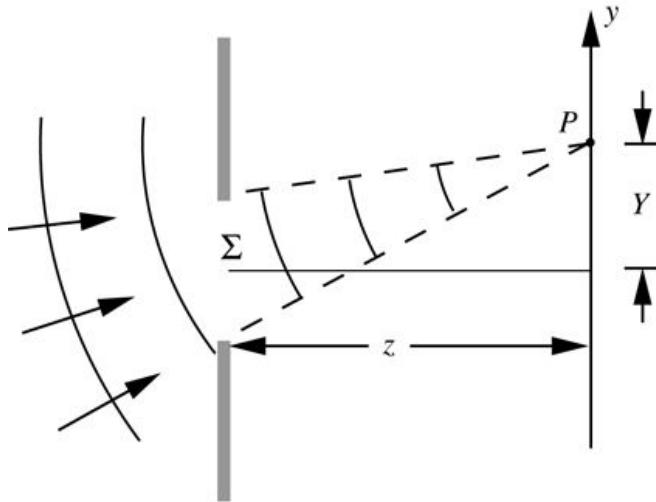


Figure P4.18
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure P4.18

The illustration shows two upward sloping dotted lines passing through aperture sigma, one from the upper extreme and the other from the lower extreme, to mark the converging path of a rightward moving spherical wave. The wave converges at point P on a vertical plane y located at a horizontal distance of z from the aperture. Point P is at a distance of Y from a point on plane y that is horizontally aligned to the center of the aperture.

19. 4-19. Find the intensity distribution *on the aperture axis* in the *Fresnel* diffraction patterns of apertures with the following transmittance functions (assume normally incident, unit-amplitude, plane-wave illumination):

$$1. t_A(\xi, \eta) = \text{circ} \sqrt{\xi^2 + \eta^2}$$

$$2. t_A(\xi, \eta) = \begin{cases} 1 & a \leq \sqrt{\xi^2 + \eta^2} < b \\ 0 & \text{otherwise} \end{cases}$$

$$\text{where } a < 1, b < 1 \text{ and } a < b$$

20. 4-20. Consider a general one-dimensional periodic object with an amplitude transmittance having an arbitrary periodic profile. Neglect the size of any bounding aperture, ignore the evanescent wave phenomenon, and assume that paraxial conditions hold. Show that at certain distances behind this object, perfect images of the amplitude transmittance are found. At what distances do these “self-images” appear?
21. 4-21. A certain two-dimensional *non-periodic* object has the property that all of the frequency components of its amplitude transmittance fall on circles in the frequency plane, the radii of the circles being given by

$$\rho_m = 2m a = 0, 1, 2, 3, \dots,$$

$$\rho_m = \sqrt{2m a} = 0, 1, 2, 3, \dots,$$

where a is a constant. Assume uniform plane-wave illumination, neglect the finite size of the object and the evanescent wave phenomenon, and assume that paraxial conditions hold. Show that perfect images of the object form at periodic distances behind the object. Find the locations of these images.

22. 4-22. A certain circularly symmetric object, infinite in extent, has amplitude transmittance

$$t_A(r) = 2\pi J_0(2\pi r) + 4\pi J_0(4\pi r)$$

$$t_A(r) = 2\pi J_0(2\pi r) + 4\pi J_0(4\pi r)$$

where J_0 is a Bessel function of the first kind, zero order, and r is radius in the two-dimensional plane. This object is illuminated by a normally incident, unit-amplitude plane wave. Paraxial conditions are assumed to hold. At what distances behind this object will we find a field distribution that is of the same form as that of the object, up to possible complex constants? (Hint: The Fourier transform of the circularly symmetric

function $J_0(2\pi r)$ is the circularly symmetric spectrum $\frac{1}{2\pi} \delta(\rho - 1)$.)

23. 4-23. An expanding cylindrical wave falls on the “input” plane of an optical system. A paraxial approximation to that wave can be written in the form

$$U(y_1) = \exp[j\lambda z_0(y_1 - y_0)/2],$$

$$U(y_1) = \exp\left\{j\frac{\pi}{\lambda z_0}[(y_1 - y_0)^2]\right\},$$

where λ is the optical wavelength, while z_0 and y_0 are given constants. The optical system can be represented by a paraxial $ABCD$ matrix (see [Appendix B, Section B.3](#)) that holds between the input and output planes of the system. Find a paraxial expression for the complex amplitude of the field across the “output” plane of the optical system, expressing the results in terms of arbitrary elements of the ray matrix. Assume that the refractive index in the input and output planes is unity. You may treat this problem as one-dimensional.

24. 4-24. In this problem, the goal is to characterize the mapping of complex field amplitudes in a one-dimensional Fourier transforming system and a one-dimensional imaging system using

ray matrices and (4-36).⁵

1. An input complex field $U_1(y)$ is input to the front focal plane of a positive lens (assume infinite extent) and the field $U_2(y)$ is observed in the rear focal plane. Analyzing this system using the cascade 3 successive ray matrices, show that a Fourier transform relationship between the fields results if $A=0$ and $D=0$. Using their definitions, show that indeed A must be zero and D must be zero.
2. Perform a similar analysis for a one-dimensional imaging system. In this case the input U_1 is located at distance z_1 in front of the lens (again assumed infinite in extent), and the output U_2 is located at distance z_2 behind the lens. The focal length of the lens is f , and the lens law $1/z_1 + 1/z_2 = 1/f$ is obeyed.
Show that

$$U_2(y_2) = e^{jkL_0} A \exp(j\pi Cy_2^2) U_1(y_1),$$

$$U_2(y_2) = \frac{e^{jkL_0}}{\sqrt{A}} \exp\left(\frac{j\pi Cy_2^2}{\lambda A}\right) U_1\left(\frac{y_1}{A}\right),$$

and A is the magnification. Hints: Complete the square in the exponent and note that since the determinant of the ray matrix is unity, $(D - A^{-1})/B = C/A$.

5 Computational Diffraction and Propagation

Often it is desirable to be able to predict the form of a diffraction pattern for apertures that do not yield closed-form analytical solutions of the diffraction formulas. Alternatively, one may wish to propagate the fields through an optical system using discrete versions of the propagation equations we have derived earlier. In such cases, one must resort to numerical calculation. Such

computations involve sampling the aperture function $U(x, y, 0)$ densely enough to minimize aliasing, padding this array with a suitable number of zeros, and performing the discrete analog of a continuous operation. In this section we discuss several techniques for such numerical calculation of diffraction patterns.

When we show calculated diffraction patterns, they will often be for a one-dimensional or two-dimensional uniform square aperture. This example, while perhaps overly simple, has the advantage that we can obtain exact results analytically and determine when the numerically calculated results are accurate. In the final sections of this chapter we discuss extensions to more complex apertures.

5.1 Approaches to Computational Diffraction

Four different approaches for computing a diffraction pattern created by a given aperture can be identified:

1. **The convolution approach.** The convolution form of the Fresnel diffraction pattern can be calculated numerically. Thus, from (4-14),

$$U(x, y, z) = e^{jkz} j \lambda z \int_{-\infty}^{\infty} \int U(\xi, \eta, 0) \exp[j\pi\lambda zx - \xi^2 + y - \eta^2] d\xi d\eta,$$

$$U(x, y, z) = \frac{e^{jkz}}{j\lambda z} \int_{-\infty}^{\infty} \int U(\xi, \eta, 0) \exp\left\{j\frac{\pi}{\lambda z}[(x - \xi)^2 + (y - \eta)^2]\right\} d\xi d\eta,$$

(5-1)

and the intensity can be found by taking the squared modulus of the result. This integral must be discretized in the computational version. There are two ways to perform the required convolution, one directly and the other by means of discrete Fourier transforms (DFTs). We will consider both methods.

2. **The Fresnel transform or the single DFT approach.** By moving the $\exp[jk2z(x^2+y^2)]$ $\exp\left[j\frac{k}{2z}(x^2+y^2)\right]$ out of the integral, we are left with a Fourier transform of the product of $U(\xi, \eta, 0)$ $U(\xi, \eta, 0)$ and the quadratic-phase factor of the form $\exp[jk2z(\xi^2+\eta^2)]$ $\exp\left[j\frac{k}{2z}(\xi^2+\eta^2)\right]$ (cf. (4-17)),

$$U(x, y, z) = e^{jkz} j \lambda z \exp[j\pi\lambda z(x^2+y^2)] \int_{-\infty}^{\infty} \int U(\xi, \eta, 0) \exp[j\frac{\pi}{\lambda z}(\xi^2+\eta^2)] e^{-j\frac{2\pi}{\lambda z}(x\xi+y\eta)} d\xi d\eta.$$

$$\begin{aligned} U(x, y, z) &= \frac{e^{jkz}}{j\lambda z} e^{j\frac{\pi}{\lambda z}(x^2+y^2)} \\ &\times \int_{-\infty}^{\infty} \int \left[U(\xi, \eta, 0) e^{j\frac{\pi}{\lambda z}(\xi^2+\eta^2)} \right] e^{-j\frac{2\pi}{\lambda z}(x\xi+y\eta)} d\xi d\eta. \end{aligned}$$

(5-2)

The discretized version of this equation is a discrete Fourier transform of the product of the array representing the aperture field and the array representing the quadratic-phase function.

3. **The Fresnel transfer function approach.** The convolution of method 1 can be Fourier transformed to yield the product of the spectrum of $U(x, y, 0)$ $U(x, y, 0)$ and the Fresnel approximation to the transfer function of free-space propagation (cf. (4-20)),

$$U(x, y, z) = \mathcal{F}^{-1} \mathcal{F} U(x, y, 0) \times e^{jkz} \exp[-j\pi\lambda z f_x^2 + f_y^2].$$

$$U(x, y, z) = \mathcal{F}^{-1}\left\{\mathcal{F}[U(x, y, 0)] \times e^{jxz} \exp[-j\pi\lambda z(f_X^2 + f_Y^2)]\right\}.$$

(5-3)

The discretized version of this approach is carried out with DFTs.

4. ***The exact transfer function approach.*** Finally, if paraxial assumptions are violated and the Fresnel approximation is in question, the more exact transfer function of (4-19) can be used,

$$U(x,y,z)=\mathcal{F}^{-1}\mathcal{F}[U(x,y,0)] \times \exp[j2\pi z\lambda -(\lambda f_X)^2-(\lambda f_Y)^2].$$

$$U(x, y, z) = \mathcal{F}^{-1}\left\{\mathcal{F}[U(x, y, 0)] \times \exp\left[j2\pi\frac{z}{\lambda}\sqrt{1 - (\lambda f_X)^2 - (\lambda f_Y)^2}\right]\right\}.$$

(5-4)

Again DFTs form the basic tool in this approach.

After a short detour discussing sampling of quadratic-phase exponentials, we consider each of these approaches in what follows. For some references to relevant literature on this problem, see [\[144\]](#), [\[241\]](#), [\[270\]](#), [\[359\]](#), [\[358\]](#), and [\[192\]](#).

5.2 Sampling a Space-Limited Quadratic-Phase Exponential

The quadratic-phase exponential

$$f(x) = e^{-j\lambda z \exp(j\pi \lambda z x^2)}$$

$$f(x) = \frac{1}{\sqrt{\lambda z}} \exp\left[j\frac{\pi}{\lambda z}x^2\right]$$

(5-5)

and its two-dimensional equivalent $f(x)f(y) f(x)f(y)$ occur frequently in diffraction problems that involve the Fresnel approximation. Because of its common occurrence, it is worthwhile to briefly discuss the sampling requirements for such a function when it is space-limited to a finite interval (see also [359]).

The integrals involving such a function in general have finite limits of integration, say $(-\ell/2, \ell/2)$, so it is the behavior within these limits that is of concern. The bandwidth of this function within the finite limits can be estimated by the following procedure. First calculate an analytic expression for the Fourier transform of this function,

$$F(f_X) = \int_{-\ell/2}^{\ell/2} e^{-j\lambda z \exp(j\pi \lambda z x^2)} e^{-j2\pi x f_X} dx,$$

$$F(f_X) = \frac{1}{\sqrt{\lambda z}} \int_{-\ell/2}^{\ell/2} \exp\left(j\frac{\pi}{\lambda z}x^2\right) \exp(-j2\pi x f_X) dx,$$

(5-6)

as was done in [Section 4.5.1](#). Second, use numerical integration to find the “equivalent bandwidth” ([37], p. 148) of the squared magnitude of this result,

$$B_X = \int_{-\infty}^{\infty} |F(f_X)|^2 df_X / |F(0)|^2.$$

$$B_X = \frac{\int_{-\infty}^{\infty} |F(f_X)|^2 df_X}{|F(0)|^2}.$$

(5-7)

The result will be found to be a function of the Fresnel number $N_F = (\ell/2)^2 / (\lambda z)$

The above operations can be carried out with the help of *Mathematica*, with the result shown in [Fig. 5.1](#). Note the different behaviors of the bandwidth, so-defined, for $N_F > 0.25$ and $N_F < 0.25$. The asymptote on the right ($N_F > 0.25$) is

$$\lambda z B_X = 4N_F \text{ or } B_X = \frac{\ell}{\lambda z},$$

$$\sqrt{\lambda z} B_X = \sqrt{4N_F} \quad \text{or} \quad B_X = \frac{\ell}{\lambda z},$$

(5-8)

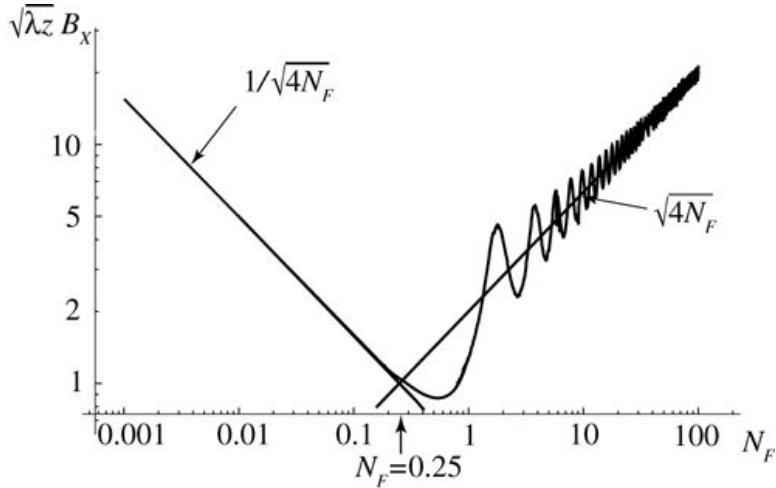


Figure 5.1

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 5.1 Equivalent width of the squared magnitude of the Fourier transform of the finite-width quadratic-phase exponential. The straight solid lines are the asymptotes that are approached in the regions $N_F < 0.25$ (on the right) and $N_F > 0.25$ (on the left). On the right the bandwidth is determined by the quadratic-phase exponential function. On the left the bandwidth is determined by the finite size of the aperture over which that function is defined.

The graph plots values ranging from 0.001 to 100 along horizontal axis N_F and values 0 to 10 along vertical axis $\sqrt{\lambda z} B_x$. It plots downward sloping line representing $1/\sqrt{4N_F}$ and upward sloping line representing $\sqrt{4N_F}$. The two lines intersect at $(0.25, 1)$. On the downward sloping line, a curve begins near the intersection and zigzags up the upward sloping line.

which is the same result one finds by considering the local spatial frequency distribution of the quadratic-phase exponential (see [Prob. 5-1](#)). That is, under the condition $N_F > 0.25$, the spectral extent is found to be well predicted by the occupancy distribution of the local spatial frequency, with greater and greater accuracy as N_F increases. For $N_F < 0.25$, this is not the case.

Given the above results, there are two issues in determining a sampling rate required for the quadratic-phase function. The first issue is that the bandwidth, as defined above, depends on the Fresnel number N_F . Clearly the sampling rate required in a given calculation depends on the Fresnel number of interest. Second, the particular measure of bandwidth we have used in obtaining [Fig. 5.1](#) is an imperfect measure when it comes to determining the required sampling rate. On the right of the figure, it is clear that there are oscillations that exceed the asymptote, although with decreasing amplitude as N_F increases.

On the left of the figure, the problems with the bandwidth estimate are more serious. In this region, which includes the conditions for Fraunhofer diffraction, the bandwidth as defined above depends almost entirely on the width of the Fourier transform of the rectangular aperture that limits the quadratic-phase exponential function, and is not well predicted by the occupancy distribution of local spatial frequency. The bandwidth predicted in this region is

$$\lambda z B_X = 14N_F \text{For} B_X = 1\ell.$$

$$\sqrt{\lambda z} B_X = \frac{1}{\sqrt{4N_F}} \quad \text{or} \quad B_X = \frac{1}{\ell}.$$

(5-9)

If this bandwidth were accepted as the required sampling rate, there would be at most two samples at the edges of the rectangular aperture of width ℓ , an entirely inadequate rate due to the severe aliasing that will occur in the frequency domain and in the diffraction pattern. To avoid such aliasing, we choose instead $B_X = M/\ell$, where $M > 1$ can be regarded as a constant chosen large enough to reduce aliasing at the edge of the diffraction pattern to a predetermined level. In accord with the sampling theorem, the spacing between samples must obey $\Delta\xi \leq \ell/M$. The parameter $\Delta\xi$ will play an important role in later sections. It will be found to determine the period of the spectrum that results from sampling of the aperture function. At the edges of the primary period of the calculated spectrum (at what we will refer to as the “fold-over” frequencies), there will be aliasing from adjacent periods of the periodic spectrum.

There is one further subtlety that should be mentioned, a constraint on the minimum number of samples K needed to properly represent the array f_k . The discrete version of the function $f(x)$ is given by

$$f_k = \lambda z \exp(j\pi\lambda z \Delta x k^2) \quad k = -K, \dots, K$$

$$f_k = \frac{1}{\sqrt{\lambda z}} \exp\left[j\frac{\pi}{\lambda z} \Delta x^2 k^2\right] \quad k = \begin{cases} -\frac{K}{2}, \dots, \frac{K}{2} - 1 & K \text{ even} \\ -\frac{K-1}{2}, \dots, \frac{K-1}{2} & K \text{ odd} \end{cases}$$

(5-10)

where Δx is the spacing between samples. The sample spacing Δx depends on whether $N_F > 0.25$ or $N_F < 0.25$.

For $N_F > 0.25$, a quadratic-phase exponential function truncated to finite length ℓ has a bandwidth $\ell/(\lambda z)$ (cf. (5-8)) and the sample spacing should be no greater than the reciprocal of the bandwidth, $\Delta x \leq \lambda z/\ell$. Substitution using the largest allowable sample spacing in the equation for f_k yields a discrete array

$$f_k = \lambda z \exp(j\pi\lambda z \lambda z \ell^2 k^2) = \lambda z \exp(j\pi 4N_F k^2).$$

$$f_k = \frac{1}{\sqrt{\lambda z}} \exp \left[j \frac{\pi}{\lambda z} \left(\frac{\ell}{\ell} \right)^2 k^2 \right] = \frac{1}{\sqrt{\lambda z}} \exp \left[j \frac{\pi}{4N_F} k^2 \right].$$

(5-11)

The number of elements in the array must be given by at least

$$K = \ell / \Delta x \geq \ell 2 \lambda z = 4N_F.$$

$$K = \ell \left| \Delta x \geq \frac{\ell^2}{\lambda z} = 4N_F. \right.$$

(5-12)

Thus when $N_F > 0.25$, to be safe we must require that $K > 4N_F$.

If on the other hand $N_F < 0.25$, the finite aperture limiting the extent of the quadratic-phase exponential determines bandwidth, which is given by $M/\ell M / \ell$. The parameter M is determined by an aliasing criterion. The sample spacing should therefore be $\Delta x \leq \ell/M$ $\Delta x \leq \ell / M$. Choosing the largest allowable spacing, the quadratic-phase exponential sequence becomes

$$f_k = 1 \lambda z \exp j \pi \lambda z \ell M^2 k^2 = 1 \lambda z \exp j \pi 4N_F M^2 k^2,$$

$$f_k = \frac{1}{\sqrt{\lambda z}} \exp \left[j \frac{\pi}{\lambda z} \left(\frac{\ell}{M} \right)^2 k^2 \right] = \frac{1}{\sqrt{\lambda z}} \exp \left[j \pi \frac{4N_F}{M^2} k^2 \right],$$

(5-13)

and the total number of elements in the array is $K > M$.

With the results of this section, we can now turn to considering the different approaches to diffraction calculations.

5.3 The Convolution Approach

The convolution approach requires that we discretize the aperture function and the quadratic-phase exponential function found in the convolution equation

$$U(x, y, z) = e^{jkz} \int_{-\infty}^{\infty} U(\xi, \eta, 0) \exp[j\pi\lambda z x - \xi^2 - \eta^2] d\xi d\eta,$$

$$U(x, y, z) = \frac{e^{jkz}}{j\lambda z} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U(\xi, \eta, 0) \exp\left\{j\frac{\pi}{\lambda z}[(x - \xi)^2 + (y - \eta)^2]\right\} d\xi d\eta,$$

(5-14)

or, in one dimension,

$$U(x, z) = 1/\lambda z \int_{-\infty}^{\infty} U(\xi, 0) \exp\left\{j\frac{\pi}{\lambda z}[(x - \xi)^2]\right\} d\xi,$$

$$U(x, z) = \frac{1}{\sqrt{\lambda z}} \int_{-\infty}^{\infty} U(\xi, 0) \exp\left\{j\frac{\pi}{\lambda z}[(x - \xi)^2]\right\} d\xi,$$

(5-15)

where we have dropped the pure phase term preceding the integral by simply redefining the phase reference. Note that the quadratic-phase exponential has infinite bandwidth, but when it is truncated, as it must be to create a finite length sequence, its bandwidth becomes approximately finite.

5.3.1 Bandwidth and Sampling Considerations

We focus on a uniform one-dimensional rectangular aperture for simplicity. More complex apertures will be considered in later sections. Let the width of the aperture function be represented by ℓ , and the width of the truncated quadratic-phase exponential function be represented by L . If we are to do a discrete convolution, both sequences must be finite in length, though not in general of the same length, and the sample spacings in both functions must be identical. The number of samples of the truncated exponential function and the aperture function will be

represented by K and M , respectively. Also let $NF = (\ell/2)/(\lambda z)$ be the Fresnel number of the finite-length aperture function. The value of ℓ is presumably known and the required value of L as a function of NF is to be found.

Assume for the moment that we are dealing with continuous (unsampled) functions and a quadratic-phase exponential function that is infinitely long. In the spectral domain, the convolution theorem implies that the spectrum of the aperture function is multiplied by the spectrum of the quadratic-phase exponential, neither being strictly bandlimited. However, the spectrum of the aperture function is falling off as frequency increases, and can be approximated as having a width

M/ℓ or M/ℓ , where M is chosen large enough to reduce the height of the spectrum to an acceptably low level, thereby determining an aliasing level in the spectral domain in the discrete case. It follows that the product of the two spectra must likewise be acceptably small at frequencies beyond this same limit. If we now allow the length of the quadratic-phase exponential function to shrink from infinity to a finite length L , we will be reducing the bandwidth of this function from infinity to approximately $L/(\lambda z)$, its bandwidth when L is finite.¹ This bandwidth can be reduced until it equals the bandwidth of the aperture function without significant loss of accuracy in the convolution. At that point we have

$$L\lambda z = M\ell \text{ or } L = \lambda z M / \ell = M N_F \ell.$$

$$\frac{L}{\lambda z} = \frac{M}{\ell} \quad \text{or} \quad L = \lambda z \frac{M}{\ell} = \frac{M}{4N_F} \ell.$$

(5-16)

If we now divide both sides by the common sampling increment $\Delta\xi$, we find that the minimum number of samples K in the quadratic phase exponential function should satisfy

$$K = M 24 N_F,$$

$$K = \frac{M^2}{4N_F},$$

(5-17)

where M is the number of samples within the aperture function (K can be larger than this limit at the cost of greater computational complexity). The total length of the result of the convolution can be shown to be

$$N = K + M.$$

$$N = K + M.$$

(5-18)

In view of the results above, we have

$$N = M 24 N_F + M.$$

$$N = \frac{M^2}{4N_F} + M.$$

(5-19)

Note that since we require $M > 4N_F$, this equation does not imply that N decreases with increasing N_F . In fact the opposite is true because the required M increases with N_F .

We know that when the Fresnel number of the aperture function is less than 0.25, the amplitude of the diffraction pattern approaches a Fourier transform of the aperture amplitude distribution. How can the convolution yield something approximating a Fourier transform? The answer is that, as the aperture array slides over the quadratic-phase exponential array, at each position it must be overlapping an approximately linear phase region of the complex exponential. The linear phase approximation to the quadratic phase must vary across the extent of that array such that all the local linear phase slopes that correspond to the spatial frequencies of interest are encountered as the aperture array moves over the exponential array. Thus the length L of the quadratic-phase exponential array must be increased to allow all necessary linear phase slopes to be “seen” by the aperture function and therefore represented in the convolution. The maximum local frequency in the quadratic-phase exponential function will be $L/(2\lambda z)$ and this frequency should equal $M/(2\ell)$ of the spectrum of the aperture function, yielding the same equations (5-16) and (5-17) derived above.

5.3.2 Discrete Convolution Equations

Turning attention back to the convolution itself, the discretized form of the convolution integral of (5-15) can be written

$$U_n(z) = \sum_{k=\max(n-K+1, 1)}^{\min(n, M)} U_k(0) h_{n-k+1}, \quad n = 1, 2, \dots, K + M - 1,$$

$$U_n(z) = \sum_{k=\max(n-K+1, 1)}^{\min(n, M)} U_k(0) h_{n-k+1}, \quad n = 1, 2, \dots, K + M - 1, \quad (5-20)$$

$$\text{where } U_n(z) = U(n\Delta x, z), \quad U_k(0) = U(k\Delta\xi, 0),$$

$$h_k = (1/\lambda z) \exp(j\pi((k-K/2)\Delta\xi)^2/2z) \quad 0 \leq k \leq K-1 \\ 0 \quad \text{otherwise},$$

$$h_k = \begin{cases} (1/\sqrt{\lambda z}) \exp\left[j\pi\frac{((k-K/2)\Delta\xi)^2}{\lambda z}\right] & 0 \leq k \leq K-1 \\ 0 & \text{otherwise.} \end{cases}$$

$$(5-21)$$

and the strange limits on the summation are there to account for cases of partial overlap of the arrays. The value of K is taken to be $K = M^2 / (4N_F)$ for all values of N_F . Using $\Delta\xi = \ell/M$, the expression for h_k becomes

$$h_k = (1/\lambda z) \exp(j4\pi N_F M^2 (k - K/2)^2 / (M^2 z)) \quad 0 \leq k \leq K-1 \\ 0 \quad \text{otherwise.}$$

$$h_k = \begin{cases} (1/\sqrt{\lambda z}) \exp\left[j\frac{4\pi N_F}{M^2} (k - K/2)^2\right] & 0 \leq k \leq K-1 \\ 0 & \text{otherwise.} \end{cases}$$

$$(5-22)$$

In two dimensions, the equivalent result is

$$U_{n,m}(z) = \sum_{k=\max(n-K+1, 1)}^{\min(n, M)} \sum_{p=\max(n-K+1, 1)}^{\min(n, M)} U_{k,p}(0) h_{n-k+1, m-p+1},$$

$$U_{n,m}(z) = \sum_{k=\max(n-K+1, 1)}^{\min(n, M)} \sum_{p=\max(n-K+1, 1)}^{\min(n, M)} U_{k,p}(0) h_{n-k+1, m-p+1},$$

(5-23)

with $n = 1, 2, \dots, K + M - 1$, $m = 1, 2, \dots, K + M - 1$, and $k = 1, 2, \dots, K + M - 1$. Here $U_{n,m}(z) = U(n\Delta x, m\Delta y, z)$, $U_{n,m}(0) = U(n\Delta\xi, m\Delta\eta, 0)$,

$$h_{n,m} = 1/(\lambda z) \exp(j4\pi N_F (n - K/2)^2 + (m - K/2)^2)$$

$$h_{n,m} = 1 \left| \left(\lambda z \right) \exp \left[j \frac{4\pi N_F}{M^2} ((n - K/2)^2 + (m - K/2)^2) \right] \right|$$

(5-24)

for $0 \leq n \leq K - 1$ and $0 \leq m \leq K - 1$, and we have assumed square arrays with equal spacings between samples, i.e. $\Delta x = \Delta y$.

The number of complex multiply-and-adds required to perform the convolution by sliding one array over the other and performing products and additions can be shown to be KM

$$= M^3 / (4N_F) \quad \text{in the one-dimensional case and } (KM)^2 = M^3 / (4N_F) 2$$

$$(KM)^2 = (M^3 / (4N_F))^2 \quad \text{in the two-dimensional case.}$$

5.3.3 Simulation Results

In Fig. 5.2 we show the dependence of M , K and N over a wide range of N_F , as determined by simulations². Nine values of N_F are used and the results are joined by first-order interpolation. In part (a), the level of intensity at the edge of the pattern, normalized by the maximum intensity, is 10^{-2} while in part (b) it is 10^{-4} . In Fig. 5.3 the diffraction pattern obtained by the direct convolution method is shown for the case $N_F = 1$ and $M = 150$ and an intensity aliasing criterion of 10^{-4} , plotted on both a linear scale and a logarithmic scale.

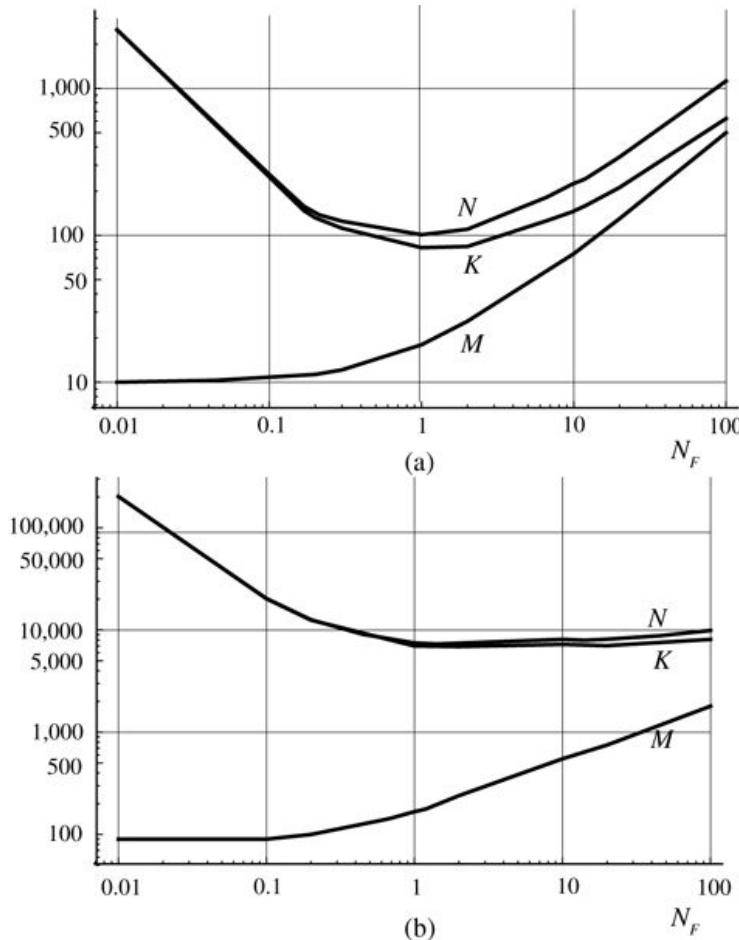


Figure 5.2

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 5.2 Dependence of M , K and N on N_F for the direct convolution method, as determined by simulations.

The two cases are (a) normalized intensity sidelobe level $\leq 10^{-2}$ at edge of the diffraction pattern and (b) normalized intensity sidelobe level $\leq 10^{-4}$ at edge of the pattern. The normalization is by the maximum value of the intensity in the diffraction pattern.

Graph A plots values ranging from 0.01 to 100 along the horizontal axis N_F and values 10 to 1,000 along the vertical axis. The M curve is an upward sloping curve that begins at (0.01, 10) and rises gently up to (100, 500). The curves for N and K begin at a point above (0.01, 1,000) and take a downward sloping path to the right, overlapping up to around (0.3, 200), beyond which both curves begin an upward sloping path, the N curve begins around (1, 100) and the K curve begins a little below it. The N curve reaches a point a little above (100, 1,000) and the K curve reaches a point a little above (100, 500). Graph B plots values ranging from 0.01 to 100 along the horizontal axis N_F and values 100 to 100,000 along the vertical axis. The M curve is an upward sloping curve that begins at (0.01, 100) and rises gently up to (100, 2,000). The curves for N and K begin on the vertical axis at a point above (0.01, 100,000) and take a downward sloping path to the right, overlapping up to (1, 7,500), beyond which both curves slightly diverge as they

take a gentle upward sloping path. The N curve reaches (100, 10,000) and the K curve reaches a point a little below it.

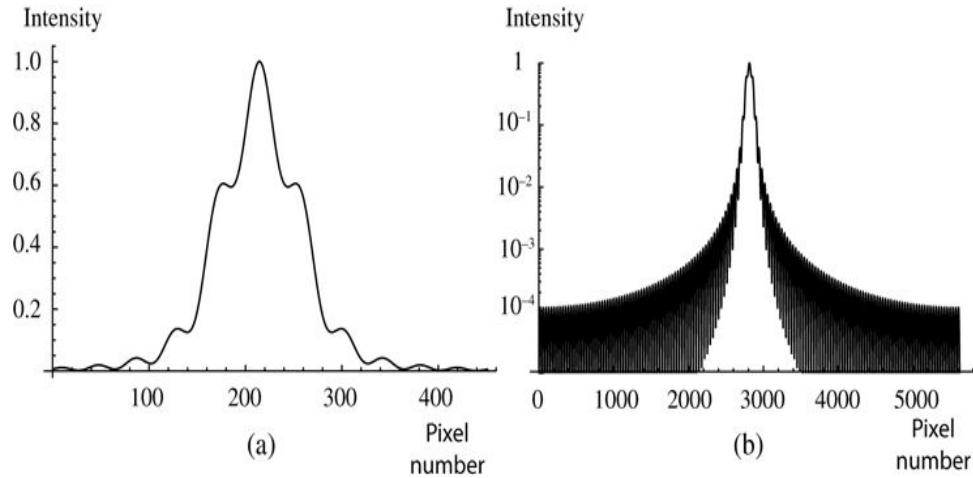


Figure 5.3

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 5.3 Diffraction pattern intensity distributions obtained by the direct convolution method with $N_F = 1$, $M = 150$ and a normalized intensity aliasing criterion of 10^{-4} . (a) Linear plot of the central part of the diffraction pattern. (b) Logarithmic plot of the entire diffraction pattern.

Graph A plots pixel number values 0 to 500 on the horizontal axis and intensity values 0 to 1 on the vertical axis. The curve begins at the origin and rises in a wavelike pattern, rising gently up to pixel number 100 and thereafter making steeper upward slopes in three waves, reaching its peak around (220, 1). The curve continues mirroring the curve thus far such that a perpendicular if dropped from the peak to the horizontal axis would be the line of symmetry. Graph B plots pixel number values 0 to 6,000 on the horizontal axis and intensity values 0 to 1 on the vertical axis. Two curves are shown to rise to the approximate location of (2750, 1), one from a point that is a little lower than the 0.0001 mark on the vertical axis and another from a point just after the 2000 mark on the horizontal axis. The area between the two curves is shaded and mirrored to the right such that the perpendicular from (2750, 1) if dropped to the horizontal axis would be the line of symmetry. The right half begins contact with the horizontal axis around the 3600 mark.

At the other extreme of $N_F > 0.25$, K is also rising. In this case the rise is due to the fact that in this region of N_F , the sampling rate is determined by the quadratic-phase exponential, and that the sampling rate is increasing with N_F due to the increasing rate of fluctuations of the phase. As a result, the increase of sampling rate with N_F leads to more and more samples within the quadratic-phase exponential array, and a consequent increase in K .

5.3.4 Convolution by Fourier Transforms

Of course the convolution can also be formed indirectly by expanding both arrays to $N = K + M$ elements using zero padding, and implementing the convolution by means of fast Fourier transforms. In this case we must apply an FFT to each of the expanded arrays, multiply their transforms, and perform an inverse FFT on that product array. Each FFT of length N

requires approximately $N \log_2 N$ complex multiply-and-adds in one dimension³, and to that number we must add N^2 products of the transforms. Since there are 2 forward FFTs and one inverse FFT required, the total operation count becomes $3N \log_2 N + N^2$ in one-dimensional and $6N^2 \log_2 N + N^2$ in two-dimensional. Here N must satisfy (5-19) with equality. In a later section we compare the computational complexities of these methods to those of other methods.

As we turn to considering the Fresnel transform approach, it is important to point out that, with the convolution approach, the sampling interval Δx in the diffraction pattern domain is identical with the sampling interval $\Delta \xi$ in the aperture domain. As we shall see, this is *not* the case for the Fresnel transform approach.

5.4 The Fresnel Transform Approach

Again assuming separability of the aperture function, the one-dimensional version of (5-2) can be written

$$U(x, z) = e^{j\pi\lambda z x} \int_{-\infty}^{\infty} U(\xi, 0) \exp(j\pi\lambda z \xi^2) \exp(-j2\pi\lambda z x \xi) d\xi,$$

$$U(x, z) = \frac{e^{j\frac{\pi}{\lambda z} x^2}}{\sqrt{\lambda z}} \int_{-\infty}^{\infty} U(\xi, 0) \exp\left(j\frac{\pi}{\lambda z} \xi^2\right) \exp\left(-j\frac{2\pi}{\lambda z} x \xi\right) d\xi,$$

(5-25)

where we have dropped a pure phase factor e^{jkz}/\sqrt{j} by properly defining the phase reference. Note that as λz shrinks, the rapidity of the oscillations of the quadratic-phase exponential in the integrand increases, suggesting that the Fresnel transform approach will be most efficient when N_F is small.

5.4.1 Sampling Increments

Usually it is desirable to embed a transmitting aperture of width ℓ in a field of zeros, so that the entire input field is of width $L > \ell$. If we choose to have N samples in the zero-padded aperture plane of width L , the spacing between samples must be

$$\Delta\xi = LN = \ell M,$$

$$\Delta\xi = \frac{L}{N} = \frac{\ell}{M},$$

(5-26)

where M is the number of samples in the aperture itself.

From (2-80) we know that $\Delta\xi\Delta f_X = 1/N$, where N is the total number of samples. In this case, the frequency variable f_X is equal to $x/(\lambda z)$, and we have

$$\Delta\xi\Delta x = \lambda z N$$

$$\Delta\xi\Delta x = \frac{\lambda z}{N}$$

(5-27)

where N^N is the total number of samples in both the ξ^ξ and the x^x domains. Note that for fixed $\Delta\xi \Delta\xi$ in the aperture plane, the spacing $\Delta x \Delta x$ of samples in the diffraction plane increases as z^z grows larger, thus compensating for the increasing spread of the diffraction pattern in the Fraunhofer diffraction region.

5.4.2 Sampling Ratio Q^Q

The ratio of full sequence length N^N to the length of the aperture sequence M^M is called the *sampling ratio*, i.e. $Q = N/M = L/\ell$, where $Q \geq 1$. The value of Q^Q that is required in any given calculation is a function of the Fresnel number NF^{NF} . The quantity $(N - M)/N = 1 - 1/Q$ represents the fraction of the total sequence length N^N that must be devoted to zero padding.

It is important to keep in mind that zero padding determines how finely the diffraction pattern is sampled, whereas for a fixed aperture length ℓ^ℓ , the number of samples M^M within the aperture determines the amount of aliasing found at the edges of the primary period of the periodic diffraction pattern. The choice of Q^Q can be influenced by the interpolation function that is used; first-order interpolation (straight lines between samples in the display) can show discrete steps in the diffraction pattern, leading to the need for finer sampling. Second-order interpolation, which uses quadratics for interpolation, gives smoother results. Most plotting routines allow the user to choose the order of the interpolation. Both Q^Q and M^M should be chosen to give results that are essentially indistinguishable from the results obtained by continuous integration.

Aliasing occurs in the amplitude domain, and then is subjected to magnitude squaring in finding intensity. The intensity errors at the edges of the calculated diffraction pattern should be examined to determine the level of aliasing.⁴ Thus for all values of NF^{NF} , we must choose M^M to ensure aliasing is acceptable and smoothness of the sidelobes is acceptable. Once the aliasing criterion is satisfied by proper choice of M^M , the amount of zero-padding required can be found starting with a large value of Q^Q and gradually reducing it until changes in the shape of the diffraction pattern are noticed. Note that the minimum value of Q^Q will depend on whether the diffraction pattern is being examined on a linear scale or a log scale, the latter case generally leading to a larger value required of Q^Q , particularly if accuracy of the sidelobes near the edge of the diffraction pattern is of concern. In many applications, significantly smaller values of Q^Q than used here will be adequate. The values found here can be used as starting points.

The Fresnel transform approach has a property that none of the other approaches discussed here share—the physical distances between samples in the aperture plane and in the diffraction plane are not the same. The relationship between sample spacings $\Delta x \Delta x$ and $\Delta\xi \Delta\xi$ in the diffraction plane and the aperture plane, respectively, can be found by the following argument.

From (5-27), we know that $\Delta x = \lambda z / (N\Delta\xi)$. In addition, by definition, $\Delta\xi = \ell/M$. It follows that

$$\Delta x \Delta\xi = \lambda z N \times M \ell \times M \ell = M^4 QNF.$$

$$\frac{\Delta x}{\Delta \xi} = \frac{\lambda z}{N} \times \frac{M}{\ell} \times \frac{M}{\ell} = \frac{M}{4QN_F}.$$

(5-28)

Thus the spacing between samples in the aperture plane and the diffraction plane are in general different. For $N_F < 0.25$, where M and Q will be found to be relatively constant for a given aliasing criterion (see [Section 5.4.5](#)), the sample spacings in the diffraction plane become larger and larger as N_F decreases below 0.25, thus compensating for the spread of the diffraction pattern in space. On the other hand, as N_F grows larger and larger than $N_F = 0.25$, Q asymptotically approaches unity and the ratio $M/(4N_F)$ approaches its lower bound of unity. Thus at large values of N_F , the spacing between samples in the diffraction plane asymptotically approaches the spacing in the aperture plane. Neither the convolution approach nor the Fresnel transfer function approach share this property.

5.4.3 Finding the Required M , Q , and N

The approach taken here to finding the parameters M , Q and N is described as follows:

1. Assume a particular value for the Fresnel number N_F .
2. Start with a large value of M (i.e. the number of samples in the aperture), remembering the requirement that $M > 4N_F$, and also start with a large value of Q . Gradually reduce M until the chosen aliasing criterion has been reached, as determined from a log plot of the diffraction pattern.
3. Now reduce Q until the diffraction pattern starts to show defects. In the cases examined here, the diffraction pattern is plotted on a log scale and the shape of the diffraction pattern is examined to determine whether there are any obvious defects. The most subtle defects are found if one examines the sidelobes of the diffraction pattern in the vicinity of the edge of the primary period of the pattern (near the fold-over frequency). If Q is too small, discontinuities may be seen in the shape of these sidelobes, or the sidelobe shape will be distorted in other ways. Ideally, Q should be chosen so that the level of these discontinuities is below the level set for aliasing. Note that there is a subjective aspect to this part of the procedure and different observers may arrive at somewhat different values of Q using such an approach. If the pattern is plotted on a linear scale, smaller values of Q will be satisfactory.
4. Since $N = QM$, determination of M and Q also determines N .

5.4.4 The Discrete Diffraction Formulas

The discrete summation required for calculating the diffraction pattern in one dimension can now be written

$$U_n(z) = \exp[j\pi\lambda z(n-N/2)2\Delta x^2] \sum_{k=0}^{N-1} U_k(0) \exp[j\pi\lambda z(k-N/2)2\Delta\xi^2] \\ \times \exp[-j2\pi\lambda z(kn\Delta\xi\Delta x)] \quad n=0,1,\dots,N-1.$$

$$U_n(z) = \exp\left[j\frac{\pi}{\lambda z}(n-N/2)^2\Delta x^2\right] \sum_{k=0}^{N-1} U_k(0) \exp\left[j\frac{\pi}{\lambda z}(k-N/2)^2\Delta\xi^2\right] \\ \times \exp\left[-j\frac{2\pi}{\lambda z}(kn\Delta\xi\Delta x)\right] \quad n=0,1,\dots,N-1.$$

(5-29)

Noting (5-27) and that $\Delta\xi = \ell/M$, this equation can be rewritten

$$U_n(z) = \exp[j\pi 4Q2NF(n-N/2)2] \sum_{k=0}^{N-1} U_k(0) \exp[j\pi 4NFM2(k-N/2)2] \\ \times \exp(-j2\pi Nnk).$$

$$U_n(z) = \exp\left[j\frac{\pi}{4Q^2N_F}(n-N/2)^2\right] \sum_{k=0}^{N-1} U_k(0) \exp\left[j\pi\frac{4N_F}{M^2}(k-N/2)^2\right] \\ \times \exp\left(-j\frac{2\pi}{N}nk\right).$$

(5-30)

The quadratic-phase exponential sequence above has been constructed to be centered at index $N/2$ and the aperture sequence should be shifted to be centered at the same index. Note that the zero-frequency component of the DFT will fall on index $n=0$, and therefore the sequence $U_n(z)$ should also be shifted to be centered in the array. Also note that the larger NF , the more rapidly the quadratic-phase exponential within the summation varies with k , consistent with our previous statements that the Fresnel transform approach is most efficient for small NF .

A generalization of this result in two dimensions takes the form

$$U_{n,m}(z) = \exp\{j\pi 4Q2NF[(n-N/2)2+(m-N/2)2]\} \\ \times \sum_{k=0}^{N-1} \sum_{p=0}^{N-1} U_{k,p}(0) \exp\{j\pi 4NFM2[(k-N/2)2+(p-N/2)2]\} \\ \times \exp[-j2\pi N(nk+mp)]$$

$$U_{n,m}(z) = \exp\left\{j\frac{\pi}{4Q^2N_F}\left[(n-N/2)^2 + (m-N/2)^2\right]\right\} \\ \times \sum_{k=0}^{N-1} \sum_{p=0}^{N-1} U_{k,p}(0) \exp\left\{j\pi\frac{4N_F}{M^2}\left[(k-N/2)^2 + (p-N/2)^2\right]\right\} \\ \times \exp\left[-j\frac{2\pi}{N}(nk+mp)\right]$$

(5-31)

(5-31)

where it has been assumed that the arrays within the summation have been centered at $(N/2, N/2)$.

If only the intensity of the diffraction pattern is needed, then the quadratic-phase exponential preceding the summations can be ignored.

5.4.5 Examples of the Dependence of $M M$ and $N N$ on NF^N_F

Using the procedure outlined above, the values of $M M$, $Q Q$ and $N N$ have been calculated for 11 values of NF^N_F and for two choices of aliasing level. [Figure 5.4](#) shows the resulting dependences of $M M$, $N N$ and $Q Q$ on NF^N_F for intensity values at the edge of the pattern no more than $10^{-2} 10^{-2}$ and $10^{-4} 10^{-4}$, relative to the maximum value of the pattern.

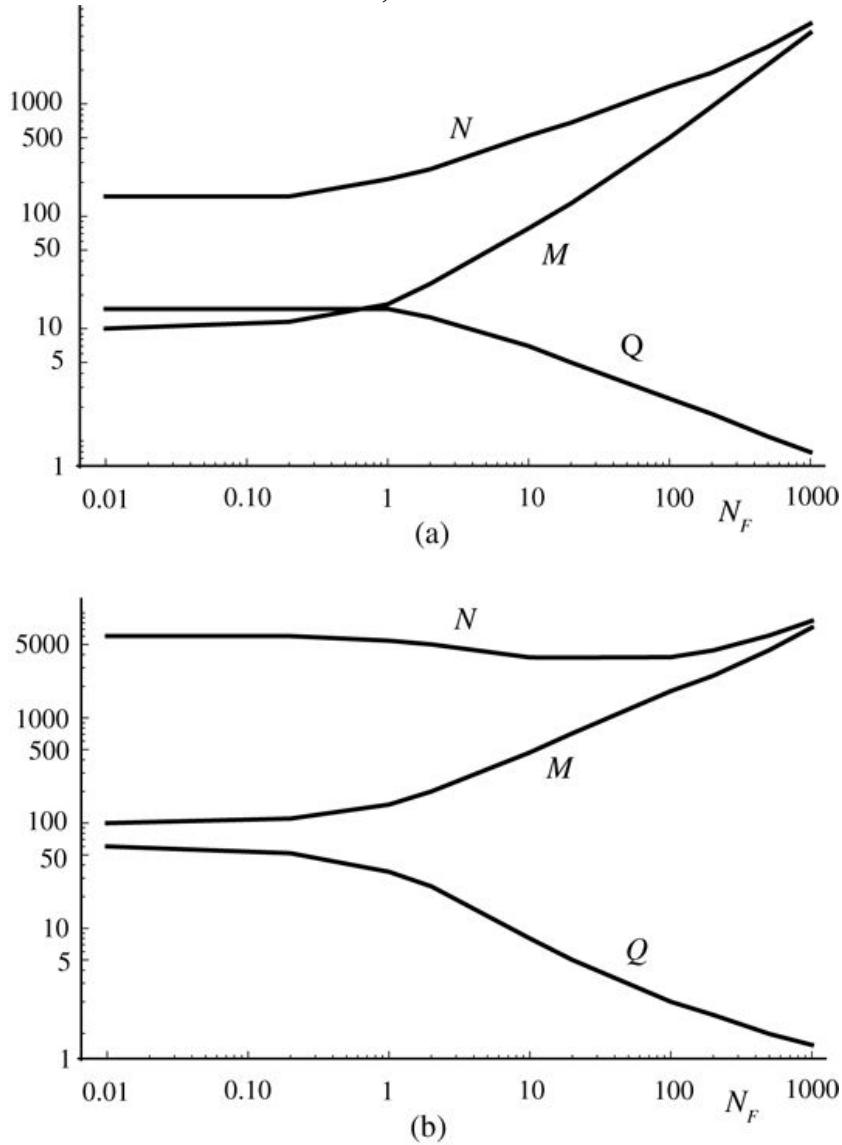


Figure 5.4
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 5.4 Plots of $N N$ and $M M$ and $Q Q$ as a function of NF^N_F for the Fresnel transform approach when the normalized intensity level at the edge of the diffraction pattern is no greater than (a) $10^{-2} 10^{-2}$ and (b) $10^{-4} 10^{-4}$.

Their approximate paths are as follows. Graph A plots N subscript F values from 0 to 1000 on the horizontal axis and N, M, and Q values from 1 to 1000 and beyond on the vertical axis. The Q curve begins around (0.01, 20), runs parallel to the horizontal axis up to (1, 20), and then slopes downward to almost reach the 1000 mark on the horizontal axis. The M curve begins around (0.01, 10) and runs mostly parallel to the horizontal axis, rising only marginally up to point (1, 20). Thereafter the curve rises steeply to a point higher than (1000, 1000). The N curve begins around (0.01, 200) and runs parallel to the horizontal axis up to the point (0.4, 200). Thereafter the curve rises to a point higher than (1000, 1000). Graph B plots N subscript F values from 0 to 1000 on the horizontal axis and N, M, and Q values from 1 to 5000 and beyond on the vertical axis. The Q curve begins around (0.01, 70), runs nearly parallel to the horizontal axis up to (0.4, 50), and then slopes downward to almost reach the 1000 mark on the horizontal axis. The M curve begins around (0.01, 100) and runs mostly parallel to the horizontal axis, rising only marginally up to point (1, 200). Thereafter the curve rises steeply to a point higher than (1000, 5000). The N curve begins around (0.01, 5000) and runs parallel to the horizontal axis up to the point (1, 5000). Thereafter the curve dips marginally and reaches a point higher than (1000, 5000).

5.4.6 Summary of Steps Using the Fresnel Transform Approach

The sequence of operations required to perform a two-dimensional diffraction calculation by this method can be summarized as follows:

1. Given λ , z , and the width ℓ of the aperture at its widest point, calculate the Fresnel number N_F of the overall aperture.
2. Choose an allowable level of intensity at the edge of the diffraction pattern, thus defining an aliasing criterion.
3. With the help of [Fig. 5.4](#), choose a value of M , N and Q that would suffice for a rectangular aperture.
4. Create the aperture array and the quadratic-phase exponential array, *both of size $M \times M$* (*not $N \times N$*), and both centered at index $(M/2, M/2)$ ($M/2, M/2$).
5. Multiply the two arrays together element by element.
6. Pad the $M \times M$ resulting array with zeros to create an $N \times N$ array with the $M \times M$ array centered inside of it.
7. Perform a DFT on the $N \times N$ array, using the FFT algorithm.
8. Shift the center of the resulting sequence from $(n,m)=(0,0)$ ($n, m = 0, 0$) to the center of the array, i.e. to $(n,m)=(N/2,N/2)$ ($n, m = (N/2, N/2)$).
9. If the field in the diffraction pattern is of interest, multiply the result of the DFT by an appropriate quadratic-phase factor, as in [\(5-30\)](#). If intensity is of interest, take the magnitude squared of each element in the result of the DFT.
10. Experiment by increasing or decreasing M and examining the aliasing level at the edge of the pattern and increase or decrease Q while monitoring the sidelobe shape at the edge of the diffraction pattern on a log plot to determine the minimum allowable value of Q .

5.4.7 Computational Complexity of the Fresnel Transform Approach

Consider now the computational complexity of this approach to diffraction pattern calculation. We assume that intensity is of interest, allowing us to ignore the quadratic-phase exponential factor preceding the summation, but we determine the complexity of calculating the field, as we did in the previous section. We will ignore the computations involved in ultimately squaring the magnitude of the field, as well as the time consumed by padding the arrays. In one-dimensional, the number of complex multiply-and-adds will be the sum of the M products involved in multiplying the two arrays, and the $N \log_2 N$ operations involved in performing an FFT of length N , the length of the padded aperture array. Thus in one dimension⁵

$$C_{1D}^{FRT} = N \log_2 N + M,$$

$$C_{1D}^{FRT} = N \log_2 N + M,$$

(5-32)

where N is illustrated in Fig. 5.4 for two different aliasing criteria. In two-dimensional, the corresponding result for an $N \times N$ array is

$$C_{2D}^{FRT} = 2N^2 \log_2 N + M^2.$$

$$C_{2D}^{FRT} = 2N^2 \log_2 N + M^2.$$

(5-33)

We defer comparisons with other methods to a later section of this chapter.

5.5 The Fresnel Transfer Function Approach

The Fresnel transfer function approach is described by [Eq.\(5-3\)](#), which we repeat here⁶:

$$U(x,y,z) = \mathcal{F}^{-1} \mathcal{F} U(x,y,0) \times e^{jkz} \exp[-j\pi\lambda z(f_X^2 + f_Y^2)].$$

$$U(x,y,z) = \mathcal{F}^{-1} \left\{ \mathcal{F} \{U(x,y,0)\} \times e^{jkz} \exp[-j\pi\lambda z(f_X^2 + f_Y^2)] \right\}.$$

(5-34)

Thus the Fresnel transfer function approach performs a Fourier transform of the aperture function, multiplies that transform by the known Fresnel approximation to the transfer function of free space,

$$H(f_X, f_Y) = e^{jkz} \exp[-j\pi\lambda z(f_X^2 + f_Y^2)],$$

$$H(f_X, f_Y) = e^{jkz} \exp[-j\pi\lambda z(f_X^2 + f_Y^2)],$$

(5-35)

and inverse Fourier transforms the product. Note that for this approach, the variations of the quadratic-phase exponent become slower and slower as z approaches the aperture, the opposite of the behavior in the Fresnel transform approach considered in the previous section. Again we first consider the one-dimensional case with a space-variable x in the aperture plane, and generalize to two dimensions later. Once again we embed an open aperture of width ℓ in a field of zeros, yielding a total width L .

5.5.1 Sampling Considerations

Let N be the total number of samples in our simulation. The spacing of samples in the aperture plane will be $\Delta x = L/N$ and the spacing of samples in the frequency domain will be $\Delta f_X = 1/L = 1/(N\Delta x)$. The number of samples within the open aperture will be $M = (\ell/L)N$.

The transfer function is multiplied by the Fourier transform of the zero-padded aperture function. That Fourier transform is not bandlimited, so there will always be some level of aliasing in the frequency domain. The separation of the spectral islands in the frequency domain is $1/\Delta x$, i.e. the inverse of the sample spacing in the aperture plane, and thus by increasing the number of samples M within the fixed size ℓ of the aperture, we diminish the effects of spectral aliasing. The effective width of the product of the transfer function and the Fourier transform of the padded aperture function is eventually decreased by the falling spectrum of the aperture function at higher spatial frequencies.

5.5.2 Finding N , M and Q for each N_F

We now use an approximate analysis to find explicit expressions for N^N , M^M that will yield good results. We again use the local frequency concept, this time in reverse, examining the local frequencies associated with the phase of the quadratic-phase transfer function to find the required extent L^L of the padded aperture.

For the quadratic-phase exponential transfer function, the relation between location in the space domain and the location in the frequency domain is found from the equation

$$x(\ell)(f_X) = 12\pi ddf_X \pi \lambda z f_X^2 = \lambda z f_X.$$

$$x^{(\ell)}(f_X) = \frac{1}{2\pi df_X} (\pi \lambda z f_X^2) = \lambda z f_X.$$

(5-36)

The bandwidth B_X in the frequency plane is limited by the fall-off of the spectrum of the aperture function, and is given by $M/\ell^M / \ell^\ell$. It follows that the length L^L of the zero-padded aperture should be

$$L = \lambda z B_X = \lambda z M / \ell.$$

$$L = \lambda z B_X = \lambda z M / \ell.$$

(5-37)

Solving for M^M we obtain

$$M = L \ell \lambda z = 4QNF.$$

$$M = \frac{L\ell}{\lambda z} = 4QN_F.$$

(5-38)

The expression for N^N becomes

$$N = QM = 4Q^2 NF.$$

$$N = QM = 4Q^2 N_F.$$

(5-39)

Because the calculations are discrete, the array in the space domain is periodic with period $N = QM$, and any calculation must check the level of aliasing in the space domain, as well as the shape of the overall pattern. Again the parameters M^M , N^N , and Q^Q are interwoven, with a choice of M^M affecting the values of Q^Q and N^N , for example. For a given NF^N_F , a choice of M^M can be made and Q^Q is chosen to satisfy (5-38), i.e. $Q = M/4NF$. If M^M is chosen too small, aliasing occurs at the edge of the diffraction pattern, and the peaks of the diffraction pattern may be misshapen.

5.5.3 The Discrete Diffraction Formulas

Since $\Delta f_X = 1/L$, the one-dimensional discrete transfer function, centered at index $N/2$, can be written as the sequence

$$H(k\Delta f_X) = \exp -j\pi\lambda z k -N/2 L^2 = \exp -j\pi\lambda z L^2(k -N/2)2 = \exp -j\pi \ell L^2(k -N/2)24NF = \exp -j\pi 4Q2NF(k -N/2)2 \quad k = 0, \dots, N - 1.$$

$$\begin{aligned} H(k\Delta f_X) &= \exp \left[-j\pi\lambda z \left(\frac{k - N/2}{L} \right)^2 \right] = \exp \left[-j\pi \frac{\lambda z}{L^2} (k - N/2)^2 \right] \\ &= \exp \left[-j\pi \left(\frac{\ell}{L} \right)^2 \frac{(k - N/2)^2}{4N_F} \right] \\ &= \exp \left[-j \frac{\pi}{4Q^2 N_F} (k - N/2)^2 \right] \quad k = 0, \dots, N - 1. \end{aligned}$$

(5-40)

Note that we have translated this array to be centered at the element with index $N/2$. The aperture itself has an amplitude distribution that should also be represented by the centered sequence. We must apply a DFT to the padded aperture sequence and shift the resulting discrete spectrum so that its zero-frequency component is lined up with the center of the transfer function. We then multiply this spectrum by the discrete version of the transfer function. Finally we perform an inverse DFT. The result of this final DFT will have its center at $n = 0$, so it should be circularly shifted to be centered at index $n = N/2$.

In terms of the DFT operator, the sequence of operations required to compute the diffracted field can be represented as follows:

$$U_n(z) = \mathcal{DFT}^{-1} \{ \mathcal{DFT}\{U_k(0)\} \exp \left(-j \frac{\pi}{4Q^2 N_F} (k - N/2)^2 \right) \}.$$

$$U_n(z) = \mathcal{DFT}^{-1} \left\{ \mathcal{DFT}\{U_k(0)\} \exp \left(-j \frac{\pi}{4Q^2 N_F} (k - N/2)^2 \right) \right\},$$

(5-41)

where the indices of the shifted aperture sequence and the quadratic-phase exponential sequence run from 0 to $N - 1$ and it is understood that the center of the discrete spectrum of the aperture array should be circularly shifted to be centered at $k = N/2$. A two-dimensional version of this equation is

$$U_{n,m}(z) = \mathcal{DFT}^{-1} \left\{ \mathcal{DFT}\{U_{k,p}(0)\} \exp \left(-j \frac{\pi}{4Q^2 N_F} ((k - N/2)^2 + (p - N/2)^2) \right) \right\},$$

$$U_{n,m}(z) = \mathcal{DFT}^{-1} \left\{ \mathcal{DFT}\{U_{k,p}(0)\} \exp \left[-j \frac{\pi}{4Q^2 N_F} ((k - N/2)^2 + (p - N/2)^2) \right] \right\},$$

(5-42)

where the DFT operation is now a two-dimensional DFT, the aperture array has been assumed to be centered in two dimensions, and the DFT of the aperture array is circularly shifted to have its $(0,0)$ -indexed component at index $(N/2, N/2)$. Note that the zero-frequency component of the diffracted field $U_{n,m}(0)$ will occur at $n = m = 0$, so this should also be circularly shifted to be centered at index $(N/2, N/2)$.

As an additional comment on this approach, the inequality $M > 4NF$ must still be obeyed. This inequality is ensured when $Q > 1$, as is always the case.

5.5.4 Examples of the Dependence of M , N and Q on NF

Through a series of simulations that calculate diffraction patterns, we find the dependence of M , N and Q on the Fresnel number NF . Two aliasing constraints are considered, one requiring that the sidelobes at the edge of the diffraction pattern intensity be no more than 10^{-2} times the peak value, and the other that they be no more than 10^{-4} times the peak value. In the results shown, large values of M were assumed, and then M was reduced until the aliasing constraint was just met, and then the value of Q was reduced to until defects were observed in the diffraction pattern displayed on a log scale. The resulting values of Q closely track the theoretical relation $Q = M / (4NF)$ for all NF . The results are shown in Fig. 5.5, obtained from 11 values of NF . A comment on the results in this figure is in order. We observed earlier that the Fresnel transfer function method is most efficient for large NF , unlike the Fresnel transform method, which is most efficient for small NF . The rise of N for small NF is clear in this figure, indicating the inefficiency of this method for that case. However, the required values of N are not falling for large NF , which might be expected if the efficiency were improving. The reason for this behavior lies in the fact that as NF increases, the structure in the diffraction pattern becomes finer and finer, requiring larger values of both M and N . We conclude that, while the Fresnel transfer function is more efficient for large NF than for small NF , we cannot conclude that it is more efficient than the Fresnel transform method for large NF . A comparison of their computational efficiencies will be examined in Section 5.7.

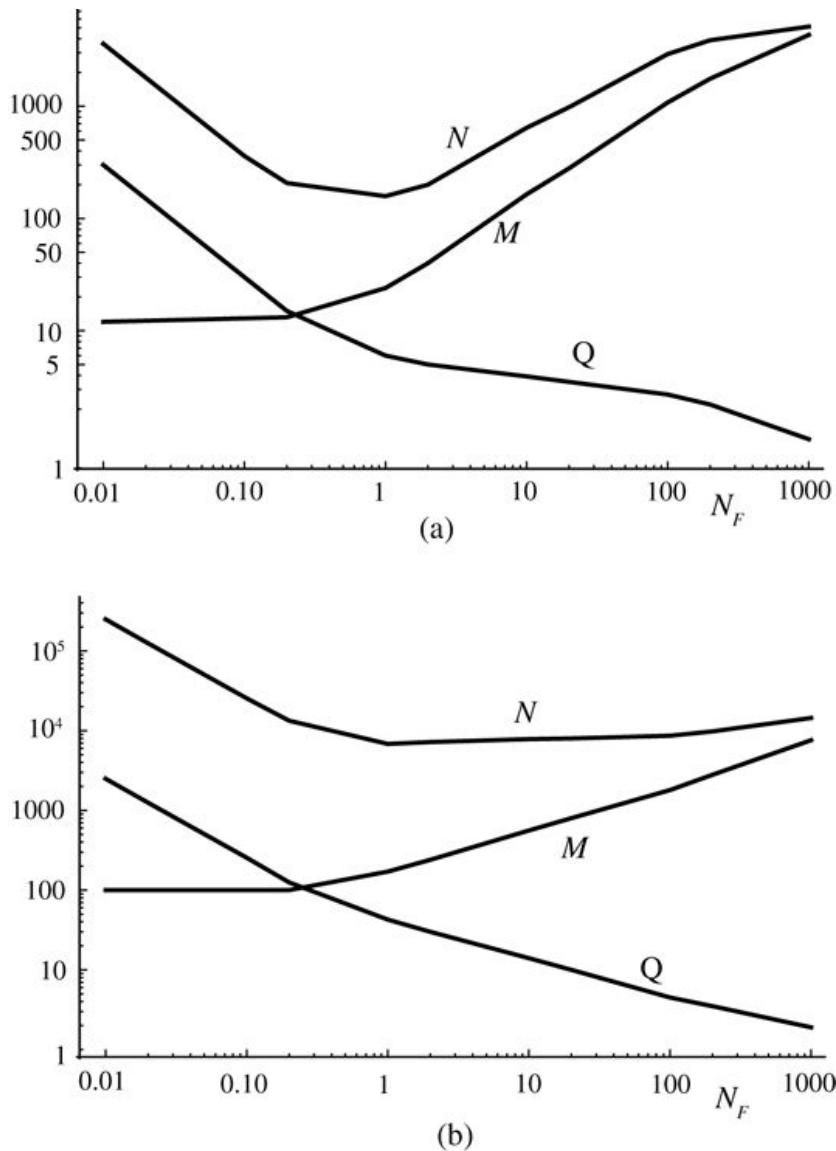


Figure 5.5
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 5.5 (a) Values of M , N , and Q required for the Fresnel transfer function approach as a function of Fresnel number N_F for aliasing intensity constraints (a) 10^{-2} and (b) 10^{-4}

Their approximate paths are as follows. Graph A plots N values from 0 to 1000 on the horizontal axis and N , M , and Q values from 1 to 1000 and beyond on the vertical axis. The Q curve begins around (0.01, 400) and slopes downward to almost reach the 1000 mark on the horizontal axis. The M curve begins around (0.01, 15) and runs mostly parallel to the horizontal axis up to around (0.3, 15), rising only marginally up to (1, 20). Thereafter the curve rises steeply to a point above (1000, 1000). The N curve begins above (0.01, 1000) and slopes downward up to (1, 200). Thereafter the curve is an upward slope to a point higher than (1000, 1000). Graph B plots N values from 0 to 1000 on the horizontal axis and N , M , and Q values from 1 to 1000 and beyond on the vertical axis. The Q curve begins around (0.01, 4000) and slopes

downward to almost reach the 1000 mark on the horizontal axis. The M curve begins around (0.01, 100) and runs mostly parallel to the horizontal axis up to around (0.3, 100). Thereafter the curve rises to (1000, 10,000). The N curve begins above (0.01, 100,000) and slopes downward up to (1, 8,000). Thereafter the curve is a slight upward slope to (1000, 20,000).

5.5.5 Summary of Steps Using the Fresnel Transfer Function Approach

A summary of the steps in this approach to diffraction calculation is as follows:

1. Given λ , z , and the largest width ℓ of the aperture, calculate the Fresnel number N_F appropriate for the entire aperture.
2. Choose an acceptable level of aliasing at the edge of the diffraction pattern.
3. Start by assuming that the aperture is rectangular, even though it may not be. Choose M , Q , and $N = QM$ based on [Fig. 5.5](#).
4. Create the padded aperture array of size $N \times N \times N$, centered at index $(N/2, N/2)$ $(N/2, N/2)$.
5. Create the quadratic-phase exponential transfer function array of length $N \times N \times N$, centered at index $(N/2, N/2)$ $(N/2, N/2)$.
6. Perform a DFT on the aperture array, using an FFT algorithm⁷. Center the result at index $(N/2, N/2)$ $(N/2, N/2)$.
7. Multiply the two spectral arrays together element by element and perform an inverse DFT on the product, again using the FFT algorithm.
8. Circularly shift the center of the resulting sequence from indices $(0,0)$ to indices $(N/2, N/2)$ $(N/2, N/2)$.
9. If intensity is of interest, take the squared magnitude of each element in the result of the inverse DFT.
10. Experiment by increasing or decreasing M to determine its minimum allowable value that will satisfy the aliasing criterion. Choose $Q = M/(4N_F)$ and adjust this value if necessary to yield good results.

5.5.6 Computational Complexity of the Fresnel Transfer Function Approach

Now we determine the computational complexity of the Fresnel transfer function approach. Since the Fresnel transfer function is known in advance, there are no calculations needed to find it. We do need to apply a DFT to the aperture sequence, which in one dimension requires $N \log_2 N$ operations. Next we multiply the two transforms, requiring N^2 operations, and finally we apply an inverse DFT to the product, requiring another $N \log_2 N$ operations. Thus the total operation count is

$$C_{\text{DFTF}} = 2N \log_2 N + N,$$

$$C_{\text{1D}}^{\text{FTF}} = 2N \log_2 N + N,$$

(5-43)

where N^N is given by (5-39). In two dimensions the two-dimensional FFTs require $2N^2 \log_2 N$ operations and the product of the spectral sequences requires $N^2 N^2$ operations. Thus the two-dimensional result can be written

$$C_{2D}^{FTF} = 4N^2 \log_2 N + N^2.$$

$$C_{2D}^{FTF} = 4N^2 \log_2 N + N^2.$$

(5-44)

Once more we defer comparing the computational complexity of the various methods to a later section.

In closing, there is one important additional point to note. In the Fresnel transfer function approach (as well as the exact transfer function approach), the spacings Δx of samples in the diffraction pattern are the same as the spacings $\Delta \xi$ of the samples in the aperture function. This is similar to the convolution approach, but different from the Fresnel transform approach.

5.6 The Exact Transfer Function Approach

Our final approach uses a transfer function that is valid without the Fresnel approximation, namely

$$U(x,y,z) = \mathcal{F}^{-1}\mathcal{F} U(x,y,0) \times \exp j2\pi z \lambda 1 - (\lambda f_X)^2 - (\lambda f_Y)^2 .$$

$$U(x, y, z) = \mathcal{F}^{-1} \left\{ \mathcal{F} \{U(x, y, 0)\} \times \exp \left[j2\pi \frac{z}{\lambda} \sqrt{1 - (\lambda f_X)^2 - (\lambda f_Y)^2} \right] \right\}.$$

(5-45)

Now the transfer function of interest is

$$\begin{aligned} H(f_X, f_Y) &= \exp j2\pi z \lambda 1 - (\lambda f_X)^2 - (\lambda f_Y)^2 , \\ H(f_X, f_Y) &= \exp \left[j2\pi \frac{z}{\lambda} \sqrt{1 - (\lambda f_X)^2 - (\lambda f_Y)^2} \right], \end{aligned}$$

(5-46)

where $(\lambda f_X)^2 + (\lambda f_Y)^2 < 1$ for propagating (i.e. nonevanescent) waves. If the radius in the frequency plane exceeds $1/\lambda$, the transfer function can be considered to be zero.

5.6.1 Sampling in the Frequency Domain

This transfer function is not separable into a function of f_X times a function of f_Y . As a consequence, we must retain the two-dimensionality of the problem from the start. For a square $L \times L$ zero-padded region, we have $\Delta x = \Delta y = L/N$, where N is the number of samples in the x and y directions, and $\Delta f_X = \Delta f_Y = 1/L$. For a square aperture of dimension $\ell \times \ell$, we have $M = (\ell/L)N$ pixels in both directions. The mathematical operation required to find the diffracted field is then

$$U_{n,m}(z) = \mathbb{DFT}^{-1} \mathbb{DFT} U_{n,m}(0) H(k\Delta f_X, p\Delta f_Y),$$

$$U_{n,m}(z) = \mathbb{DFT}^{-1} \left\{ \mathbb{DFT} \{U_{n,m}(0)\} H(k\Delta f_X, p\Delta f_Y) \right\},$$

(5-47)

where the DFTs are two-dimensional and the sequences should be shifted to be centered in the two-dimensional array. It follows that

$$H(k\Delta f_X, p\Delta f_Y) = \exp j2\pi z \lambda 1 - \lambda L^2 (k - N/2)^2 + (p - N/2)^2 ,$$

$$H(k\Delta f_X, p\Delta f_Y) = \exp \left[j \frac{2\pi z}{\lambda} \sqrt{1 - \left(\frac{\lambda}{L}\right)^2} \left((k - N/2)^2 + (p - N/2)^2 \right) \right],$$

where k and p run from 0 to $N-1$. An expression for L in terms of M and other parameters will be found in the next subsection.

5.6.2 Sampling in the Space Domain

As in the case of the Fresnel transfer function, we use the fact that a convolution in the space domain is equivalent to a multiplication in the frequency domain. If we adopt an aliasing criterion, we can consider the bandwidth of the aperture function to be M/ℓ , where M has been chosen large enough to satisfy the aliasing goal. Unlike the case of the Fresnel transfer function, which extended over all frequencies, the exact transfer function has a cutoff frequency at radius $1/\lambda$ in the frequency plane. Thus we should consider two cases, one in which the width of the exact transfer function, $2/\lambda$, exceeds M/ℓ , and one in which the extent of the exact transfer function is less than M/ℓ . If the width of the transfer function exceeds the width M/ℓ , the product spectrum can be regarded as being limited to M/ℓ . If the width of the transfer function is less than M/ℓ , then the bandwidth will be determined by the frequency cutoff of the exact transfer function. Let $B_H = 2/\lambda$ represent the bandwidth of the transfer function, and $B_A = M/\ell$ represent the bandwidth of the aperture, subject to the required aliasing condition.

Consider first the case $B_H \gg B_A$. In this case, there is no difference between the calculations using the Fresnel transfer function and the exact transfer function. Thus the flat magnitude of the exact transfer function is attenuated by the spectrum of the aperture function, and the actual bandwidth of interest is only the bandwidth of the aperture function, subject to an aliasing criterion in the frequency domain. We assume a square aperture function of dimensions $\ell \times \ell$ and select a value of M , the number of samples within one dimension of that aperture, such that the fold-over frequencies where aliasing occurs are located at $\pm M/(2\ell)$ $\pm M/(2\ell)$. The choice of M depends on our choice of a spectral aliasing criterion.

Next consider the case when the bandwidth of the transfer function B_H is still greater than the bandwidth of the aperture function B_A , but this time with B_H being only slightly larger, so that the effects of the evanescent cutoff in the frequency domain can perhaps be seen in the resulting diffraction pattern. The variations of the exact transfer function phase at various spatial frequencies define the locations of significant diffracted light arriving in the space domain. Here we will be using the local frequency concept in reverse. The phase of the exact transfer function within the frequency region lower than the evanescent cutoff is

$$\theta(f_X, f_Y) = 2\pi z \lambda [1 - (\lambda f_X)^2 - (\lambda f_Y)^2].$$

$$\theta(f_X, f_Y) = 2\pi \frac{z}{\lambda} \sqrt{1 - (\lambda f_X)^2 - (\lambda f_Y)^2}.$$

(5-48)

The space-domain equivalents to local frequencies along the f_X and f_Y axes are given by

$$x(\ell)(f_X) = 12\pi \partial \partial f_X \theta(f_X, 0) = -f_X z \lambda_1 - f_X 2 \lambda_2 y(\ell)(f_Y) = 12\pi \partial \partial f_Y \theta(0, f_Y) = -f_Y z \lambda_1 - f_Y 2 \lambda_2.$$

$$\begin{aligned} x^{(\ell)}(f_X) &= \frac{1}{2\pi \partial f_X} \theta(f_X, 0) = -\frac{f_X z \lambda}{\sqrt{1 - f_X^2 \lambda^2}} \\ y^{(\ell)}(f_Y) &= \frac{1}{2\pi \partial f_Y} \theta(0, f_Y) = -\frac{f_Y z \lambda}{\sqrt{1 - f_Y^2 \lambda^2}}. \end{aligned}$$

(5-49)

Since the aperture has been assumed to be square, we can focus on the length L of the padded array along the x axis, and results along the y axis will immediately be evident. The length L is found as the difference between $x(\ell) - M/(2\ell)$ and $x(\ell)M/(2\ell)$, since $\pm M/(2\ell) \pm M/(2\ell)$ are frequencies at the edges of the primary period of the periodic spectrum. These are the highest local frequencies associated with the transfer function phase, and they will therefore, to a good approximation, determine the extent of the diffraction pattern in the space domain. Evaluating this difference we find

$$L = M z \lambda \ell_1 - M 2 \lambda_2 / (4\ell).$$

$$L = \frac{M z \lambda}{\ell \sqrt{1 - M^2 \lambda^2 / (4\ell^2)}}.$$

(5-50)

Recalling that $M = \ell / \Delta x$, where Δx is the space domain sampling interval, we see that equivalently

$$L = M / (4N_F) \ell_1 - 14 \lambda \Delta x \ell_2.$$

$$L = \frac{M / (4N_F)}{\sqrt{1 - \frac{1}{4} \left(\frac{\lambda}{\Delta x} \right)^2}} \ell.$$

(5-51)

From this result we find $Q = L / \ell$,

$$Q = M / (4N_F) \ell_1 - 14 \lambda \Delta x \ell_2.$$

$$Q = \frac{M / (4N_F)}{\sqrt{1 - \frac{1}{4} \left(\frac{\lambda}{\Delta x} \right)^2}}.$$

(5-52)

This value of Q^Q is larger than the value of $Q = M/(4NF)$ derived in the case of the Fresnel transfer function. How much larger depends on the ratio of $\lambda/\Delta x$ to $\Delta x/\lambda$. We note that when $\Delta x = \lambda/2$, the argument of the square root becomes zero, and the value of Q^Q becomes infinite. Thus the zero-padded space-domain sequence has become infinitely long. This is precisely the condition that occurs when $B_H < B_A$, for then the bandwidth of the product of the spectra is $2/\lambda$, due to the fact that the transfer function falls to zero at $\pm 1/\lambda$. Infinitely long sequences are of course not useful in practice, but considering them does lend some insight into the effects of the exact transfer function versus the Fresnel transfer function.

Note that when $\Delta x \gg \lambda$, the second term under the square root sign can be ignored, and the value of Q^Q agrees with the value found for the Fresnel transfer function. Results based on the Fresnel transfer function are accurate in such a case. However, note that to avoid an imaginary square root in the expression for Q^Q , which would be an indication that we have gone beyond the evanescent cutoff, the inequality $\Delta x > \lambda/2$ must be satisfied. Since there is a lower limit to Δx and since for accuracy there is a lower limit to the number M of samples in the aperture, there must exist a lower limit to the length of the aperture $\ell = M\Delta x$ that can be analyzed by this approach. This lower limit depends on the Fresnel number since the lower limit on M depends on NF .

Our task now is to start with values of M that satisfy the assumed spectral aliasing criterion, find the corresponding values of Q^Q and NF , and perform simulations to find the required values of these parameters for accurate results. In this case we do not have analytical results for comparison purposes, so accuracy will have to be determined by increasing the parameter values until no further changes of the diffraction pattern intensity shape are found to result.

5.6.3 Simulation Results

Simulations in the case of the exact transfer function are inherently more time consuming due to the necessary two-dimensional aspect of the problem. In addition, there is a new parameter $\Delta x/\lambda$ that affects the results in significant ways, as well as a constraint on how small $\Delta x/\lambda$ can be once M and NF are chosen. Here we will only illustrate with one example, namely a square aperture of width ℓ on a side, and a Fresnel number $NF = 10$. [Figure 5.6](#) shows plots of the normalized intensity in central slices through the diffraction patterns on both a linear scale (left side) and a log scale (right side) for two different values of $\Delta x/\lambda$. In all plots, the value of M is 180. In part (a), the value of $\Delta x/\lambda = 111$, and therefore the Fresnel transfer function and the exact transfer function yield identical results. That the result is accurate can be confirmed by comparing it with the analytical result in [Fig. 4.15](#). In part (b), $\Delta x/\lambda = 0.509$, which is just slightly larger than the lower allowable

limit. On the left, a central subsequence of the intensity is shown on a linear scale. The results plotted on the right show the entire calculated diffraction pattern on a log scale, so that the level of aliasing at the edge of the diffraction pattern can be assessed. A number of observations can be made by comparing these results:

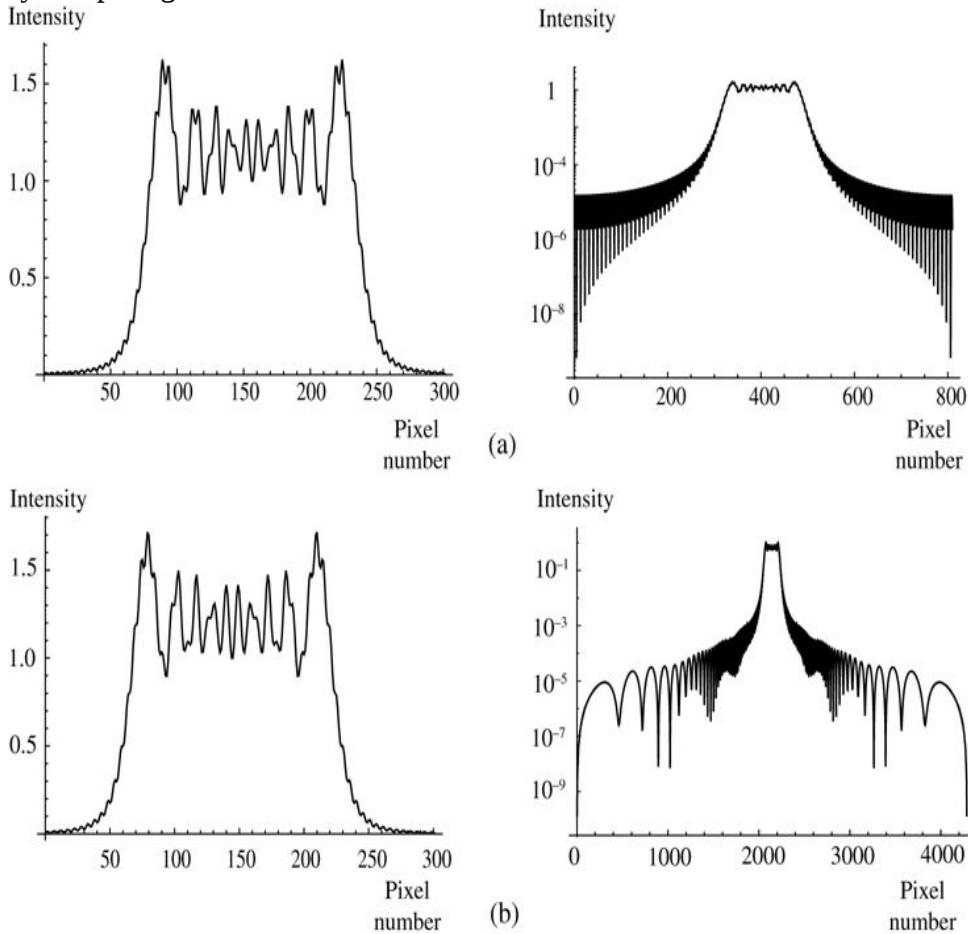


Figure 5.6

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 5.6 Plots of central slices through the two-dimensional diffraction pattern intensity on a linear scale (left) and log scale (right) using the exact transfer function method for two values of $\Delta x/\lambda$: (a) $\Delta x/\lambda = 111$ and (b) $\Delta x/\lambda = 0.509$. For both cases, $N_F = 10$ and $M = 180$. For (a), $Q = 4.5$ and $N = 810$ were used, while for (b) $Q = 24$ and $N = 4320$ were used.

Illustration a shows two graphs. The first graph plots pixel numbers from 0 to 300 along the horizontal axis and intensity from 0 to 1.7 along the vertical axis. The curve runs almost parallel to the horizontal axis and rises steeply near the 50 mark to (90, 1.6). Thereafter, the curve in a zigzag movement slopes about 10 times between the 0.8 and 1.6 levels on the vertical axis. At (225, 1.6), the curve takes a steep downward slope to a point just above the 250 mark on the horizontal axis and then slopes marginally to reach the 300 mark. The second graph in illustration a plots pixel numbers 0 to 800 on the horizontal axis and intensity 0 to 1 on the vertical axis. A slightly upward curving triangular area between the points (0, 0), (0, 10 to the power minus 5), and (350, 1) is shown shaded. It tapers almost to a line after the point (260, 10 to the power minus 4). An upward

sloping curve from (0, 10 to the power minus 6) to (350, 1) divides the triangular area into two, the one above is so dense it is black and the one below is filled with a series of close vertical lines. Between (350, 1) and (475, 1) the curve is a horizontal squiggle. The graph continues rightward such that a perpendicular from (400, 1) would serve as a line of symmetry. Illustration b shows two graphs. The first graph plots pixel numbers from 0 to 300 along the horizontal axis and intensity from 0 to 1.7 along the vertical axis. The curve runs almost parallel to the horizontal axis and rises steeply near the 50 mark to (80, 1.7). Thereafter, the curve in a zigzag movement slopes about 10 times between the 0.9 and 1.7 levels on the vertical axis. At (210, 1.7), the curve takes a steep downward slope to a point just above the 250 mark on the horizontal axis and then slopes marginally to reach the 300 mark. The second graph in illustration b plots pixel numbers 0 to 4000 on the horizontal axis and intensity 0 to 10 on the vertical axis. The graph is a series of arches of varying heights that progressively become narrow as they approach the (2000, 10) mark, thus turning denser as it approaches this point. The path does not go below the 10 to the power minus 8 level on the vertical axis. Between (2050, 10) and (2200, 10) the curve is a dense horizontal squiggle. The graph continues rightward such that a perpendicular from (2100, 10) would serve as a line of symmetry.

1. Consider first the two plots of the normalized intensity of the diffraction pattern on a linear scale. Differences in the fine structure of the linearly plotted diffraction patterns on the left are quite evident, and thus the Fresnel transfer function and the exact transfer function approaches can yield different results when $\Delta x/\lambda$ is close to its lower limit.
2. Examination of the two log-scale plots shows that for a fixed M , there is much less aliasing at the edge of the diffraction pattern when $\Delta x/\lambda$ is near its lower limit than when it is very large. Evidently the diffraction pattern is driven to near zero at its edge by the evanescent wave phenomenon. In this particular case, the value of the normalized intensity at the edges of the diffraction pattern is approximately $10 \cdot 10^{-10}$, while in the Fresnel transfer function case it is close to $10 \cdot 10^{-3}$
3. The previous observation suggests that it is possible to lower the value of M below that used in the Fresnel transfer function case when $\Delta x/\lambda$ is near its lower limit, with the result that both Q and N can be lower than their values used here. In this particular example, with $NF = 10$ and $\Delta x/\lambda = 0.509$, the value of M can be lowered from 180 to 100 without appreciable changes in the results, with accompanying lowering of values of Q from 24 to 3 and N from 4320 to 298. The level of normalized aliasing at the edge of the intensity diffraction pattern becomes 6×10^{-4} with these smaller parameters.

5.6.4 Computational Complexity of the Exact Transfer Function Approach

The expressions for the computational complexity of the exact transfer function method are the same as those for the Fresnel transfer function method when expressed as a function of N :

$$C_{DET} = 2N \log_2 N + N,$$

$$C_{1D}^{ETF} = 2N \log_2 N + N,$$

(5-53)

and in two dimensions

$$C_{2D}^{ETF} = 4N^2 \log_2 N + N^2.$$

$$C_{2D}^{ETF} = 4N^2 \log_2 N + N^2.$$

(5-54)

However, as will be discussed in the next section, for a given value of N_F^{NF} , the value of N_N^N in this expression can be different than the value of N_N^N in the case of the Fresnel transfer function.

5.7 Comparison of Computational Complexities

The problem of comparing computational complexities is a tricky one, since we must invoke certain assumptions. The curves shown here and in the following are based on samples taken at a set of 9 values of N_F in all cases. We present only the results for two-dimensional computations. We should emphasize that the computational complexities presented here assume that zero-padding the functions is done in the space domain. Different results might be obtained if other methods for gaining resolution are used.

1. **Convolution approach.** If the convolution approach is performed by sliding one array over the other in the usual “direct” discrete convolution method, the computational complexity C_{2D}^D in two-dimensional is given by

$$C_{2D}^D = (KM)^2 = M^3 / (4N_F)$$

$$C_{2D}^D = (KM)^2 = \left(\frac{M^3}{4N_F} \right)^2$$

(5-55)

for all N_F , with M determined from [Fig. 5.2](#). Alternatively, if the convolution is performed by using 3 FFTs, the computational complexity C_{2D}^{FFT} is

$$C_{2D}^{FFT} = 6N^2 \log_2 N + N^2,$$

$$C_{2D}^{FFT} = 6N^2 \log_2 N + N^2,$$

(5-56)

where $N = M^2 / (4N_F) + M$. [Figure 5.7](#) shows plots of the computational complexities for the direct convolution approach C_{2D}^D and the FFT approach C_{2D}^{FFT} to convolution for two values of normalized intensity levels at the fold-over frequency. As can be seen, the FFT approach is always more efficient than the direct convolution approach.

2. **The Fresnel transform approach.** For the Fresnel transform approach in two dimensions, the following is the relevant expression:

$$C_{2D}^{FRT} = 2N^2 \log_2 N + M^2$$

$$C_{2D}^{FRT} = 2N^2 \log_2 N + M^2$$

with $N = MQ$.

[Figure 5.8](#) shows plots of the computational complexity of the FFT convolution approach and the Fresnel transform approach, for normalized intensity aliasing criteria of 10^{-2} and 10^{-4} . Note that even though the FFT method requires three Fourier transforms while the Fresnel transform method requires only one, in the range $N_F = 0.4$ to $N_F = 40$, with an aliasing criterion of 10^{-2} , the FFT approach is computationally less complex, but for most other values of N_F , the Fresnel transform approach requires fewer operations. When the aliasing criterion is 10^{-4} the Fresnel transform method always requires fewer operations than the FFT method.

3. **Fresnel transfer function approach.** Using the results for N obtained in [Section 5.5](#), and taking account that there are two two-dimensional DFTs required, the expression for computational complexity in this case is

$$C_{\text{2DFTF}} = 4N^2 \log_2 N + N^2.$$

$$C_{\text{2D}}^{\text{FTF}} = 4N^2 \log_2 N + N^2.$$

(5-57)

In [Fig. 5.9](#) we show the computational complexities for the Fresnel transform method and the Fresnel transfer function method. As can be seen, when the intensity aliasing criterion is 10^{-2} , the Fresnel transform approach and the Fresnel transfer function approach are close to one another in computational complexity for $N_F > 0.2$, with the Fresnel transfer function approach being slightly superior between approximately $N_F = 0.4$ and $N_F = 3$ and for $N_F > 60$, but the Fresnel transform approach being superior otherwise. When the intensity aliasing condition is 10^{-4} , the Fresnel transform approach is always more efficient than the Fresnel transfer function method. We see that the Fresnel transfer function method considered by itself is more efficient for large N_F than it is for small N_F , as expected.

4. **Approach based on the exact transfer function.** While the expressions for the computational complexity for the Fresnel transfer function approach and the exact transfer function approach are the same when expressed in terms of the total number of sample points N , it is not correct to say that the two approaches have the same computational complexity in general. While when $\Delta x/\lambda \gg 1$, the computational complexities are indeed the same, this is not necessarily the case when $\Delta x/\lambda$ is close to its lower limit of $1/2$. In the latter case, for fixed M and N_F , N increases as $\Delta x/\lambda$ approaches this lower limit. However, as we have seen, it is possible to decrease M as this limit is approached, due to improved aliasing, and a decrease in M results in a smaller N . The interaction of all the variables, M , Q , N , and $\Delta x/\lambda$ makes it

difficult to compare the computational complexity with previous cases. Perhaps the only firm message in this case is that when $\Delta x/\lambda$ is sufficiently large, the computational complexities of the exact transfer function method and the Fresnel transfer function method are essentially the same.

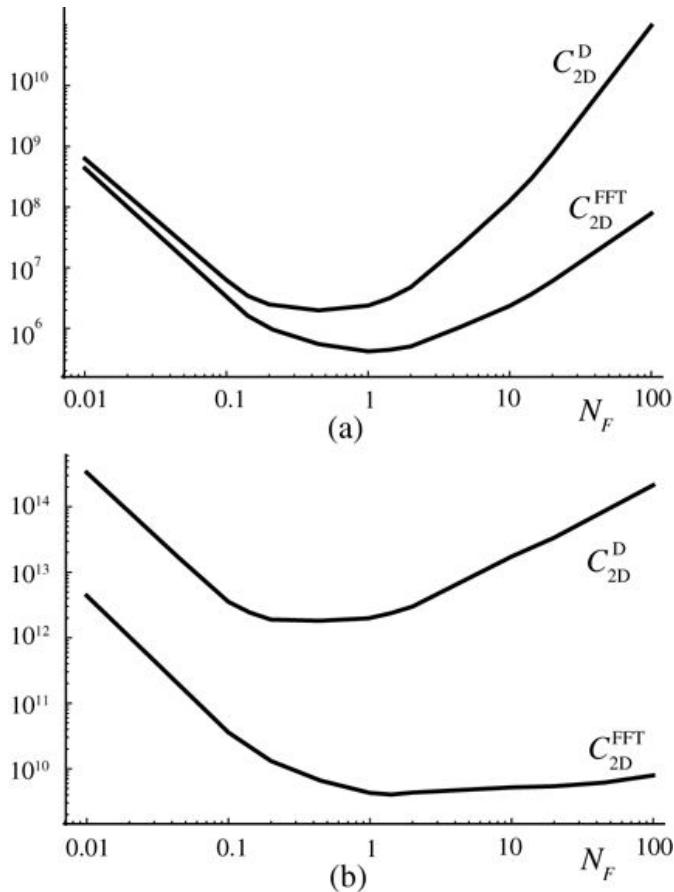


Figure 5.7

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 5.7 Computational complexities C_{2D}^D of the direct convolution approach and C_{2DFFT}^{FFT} of the FFT approach to convolution for a normalized intensity aliasing criterion of (a) 10^{-2} , and (b) 10^{-4} .

The graphs plot N_F values 0 to 100 on the horizontal axis and computational complexity on the vertical axis. The curves' approximate paths are as follows.

Graph a shows two curves. The curve for C_{2D}^{FFT} begins at $(0.01, 5.5 \times 10^8)$ and slopes downward till $(1, 500,000)$. Thereafter it rises to reach the $(100, 10^{10})$ point. The curve for C_{2D}^D begins at $(0.01, 5.5 \times 10^8)$ and slopes downward till $(0.6, 3.3 \times 10^6)$. Thereafter it rises to reach the $(100, 10^{11})$ point.

Graph b shows two curves. The curve for C_{2D}^{FFT} begins a little below the point $(0.01, 10^{13})$ and slopes downward till $(1, 5.5 \times 10^9)$. Thereafter it rises marginally to reach the $(100, 10^{10})$ point.

The curve for C_{2D}^{FFT} begins at a point above (0.01, 10^{14}) and slopes downward till a point a little above (0.3, 10^{12}). It runs parallel to the horizontal axis till it is above the 1 mark on the horizontal axis and thereafter rises to reach a point a little above (100, 10^{14}).

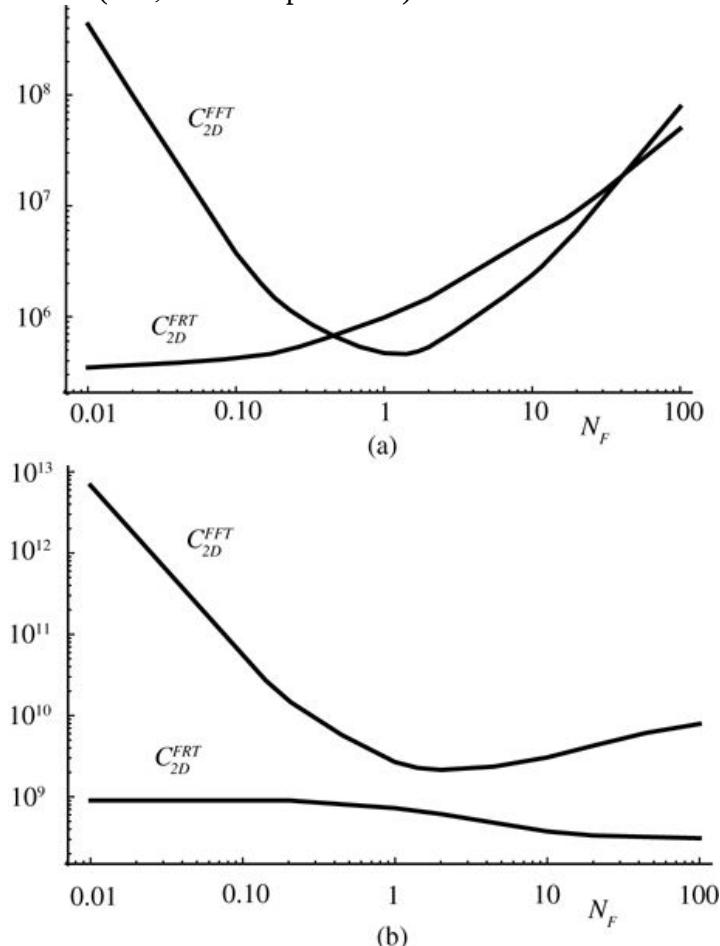


Figure 5.8

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 5.8 Computational complexities C_{2D}^{FFT} of the FFT approach to convolution and C_{2D}^{FRT} of the Fresnel transform approach to diffraction pattern calculation for normalized intensity aliasing criterions of (a) 10^{-2} and (b) 10^{-4} .

The graphs plot N_F values 0 to 100 on the horizontal axis and computational complexity on the vertical axis. The curves approximate paths are as follows.

Graph a shows two curves. The curve for C_{2D}^{FRT} begins a little below the point (0.01, 5.5×10^5) and slopes upward till a point a little above (100, 5.5×10^7). The curve for C_{2D}^{FFT} begins at a point a little below (0.01, 10^9) and slopes downward till (2, 5.5×10^5). Thereafter it rises to almost reach the (100, 10^{14}) point. Graph b shows two curves. The curve for C_{2D}^{FRT} begins a little below the point (0.01, 10^9) and runs

parallel to the horizontal axis till (0.4, 10 to the power 9). Thereafter it slopes gently downward till a point a little below (100, 5.5×10 to the power 8). The curve for C subscript 2D superscript FFT begins at a point a little below (0.01, 10 to the power 13) and slopes downward till a point that is a little below (4, 5.5×10 to the power 9). Thereafter it rises gently to almost reach the (100, 10 to the power 10) point.

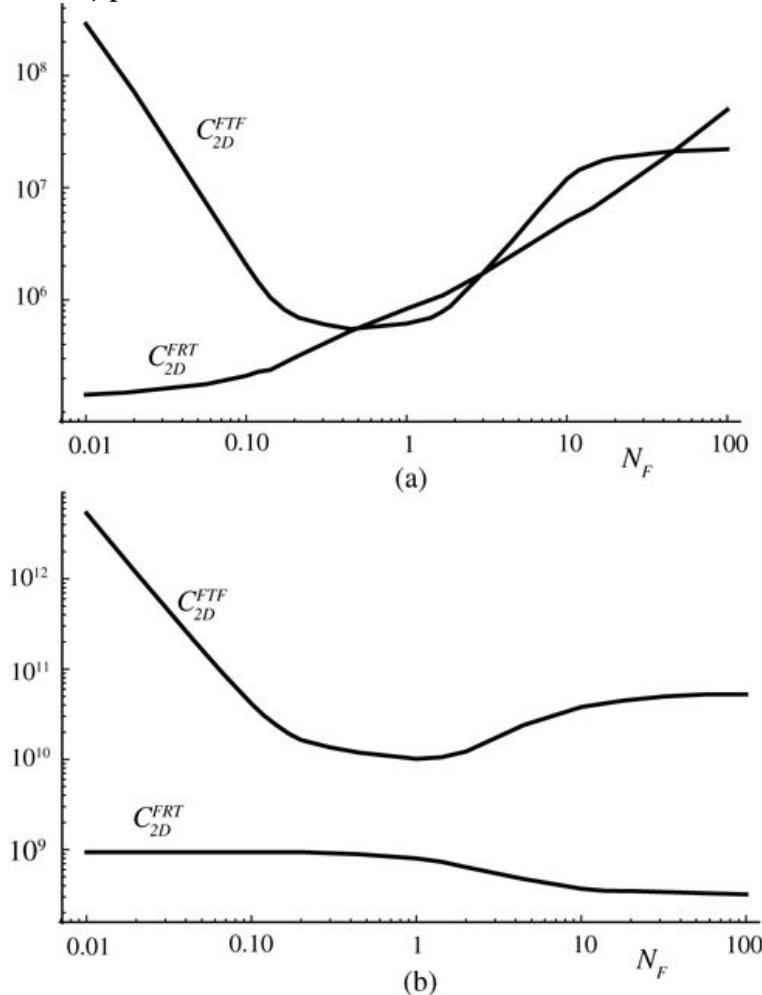


Figure 5.9

Goodman, *Introduction to Fourier Optics*, 4e,

© 2017 W. H. Freeman and Company

Figure 5.9 Computational complexities of the Fresnel transform approach C_{2D}^{FRT} and the Fresnel transfer function approach C_{2D}^{FTF} for a normalized intensity aliasing criterion of (a) 10^{-2} and (b) 10^{-4} .

The graphs plot N_F values 0 to 100 on the horizontal axis and computational complexity on the vertical axis. The curves approximate paths are as follows. Graph (a) shows two curves. The curve for C_{2D}^{FRT} begins a little below the point (0.01, 2.5×10^5) and slopes upward till a point a little below (100, 5×10^7). The curve for C_{2D}^{FTF} begins at (0.01, 3×10^8) and slopes downward till (0.5, 6.5×10^6), where it intersects the other curve. Thereafter it rises gently to once again intersect the other curve at (3, 2×10^6) to rise steeply up to (11, 2×10^9).

7). Thereafter, intersecting at (50, 2×10 to the power 7), it rises only marginally to a point a little above (100, 2×10 to the power 7). Graph b shows two curves. The curve for C subscript 2D superscript FRT begins around (0.01, 10 to the power 9) and runs parallel to the horizontal axis till (0.4, 10 to the power 9). Thereafter it slopes gently to (10, 3×10 to the power 8) and then extends to (100, 2×10 to the power 8). The curve for C subscript 2D superscript FFT begins at a point a little below (0.01, 5×10 to the power 12) and slopes downward till (1, 10 to the power 10). Thereafter it rises gently to almost reach the (100, 5×10 to the power 10) point.

5.8 Extension to More Complex Apertures

The number of aperture shapes one might consider is unlimited, and it is difficult to find a single procedure that is appropriate for all of them. In this section we illustrate with some additional examples, in hopes that the procedures used will suggest approaches that can be used in more general cases. We begin with a one-dimensional example, then generalize to aperture shapes that are separable in (x, y) coordinates, then consider the case of circular symmetry, and finally make some suggestions for the arbitrarily general case.

5.8.1 One-Dimensional Case

Our first example is that of a pair of rectangular apertures, each of width ℓ , with their centers separated by a distance 2ℓ , as illustrated in Fig. 5.10. We must first determine the bandwidth of this aperture, for that will in turn determine the sampling requirements. The smallest structure in the double aperture function is a single rectangular aperture, and from past material we know the sampling requirements for such an aperture. We define N_{F1} as being the Fresnel number for one of the sub-apertures, $(\ell/2)^2/(\lambda z)$, and the Fresnel number of the entire array as $N_{F2} = (3\ell/2)^2/(\lambda z)$. The number of samples in one aperture, represented by M_1 , and can be found from Fig. 5.4 or Fig. 5.5, depending on the method of calculation to be used. The number of samples M in the entire width 3ℓ of the aperture function must be $M = 3M_1$. For accurate results, the total number of samples in the entire array must satisfy $M > 4N_{F2}$, where N_{F2} applies to the entire array.

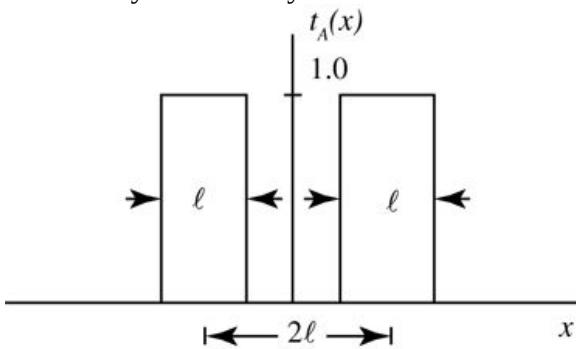


Figure 5.10

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 5.10 Aperture consisting of a pair of rectangular openings.

The identical rectangular apertures, each of width l , are placed on horizontal axis x such that the distance separating their centers is $2l$. Their height is 1.0 as marked on the vertical axis t subscript A (x), which is between the apertures and equidistant from the two.

We choose to use the Fresnel transform approach to calculate the diffraction pattern of this aperture because this approach is often the most efficient one. From Fig. 5.4 we can estimate the required sizes of M_1 and $N_1 = QM_1$ for a single rectangular aperture, and then triple them for this case to specify M and N . Figure 5.11 shows the resulting diffraction patterns calculated by this approach for $NF_1 = 100, 1$ and 0.01 . Note the fringe that exists under an envelope consisting of a sinc^2 diffraction pattern that would be produced by a single rectangular aperture when $NF_1 = 0.01$. A key step in calculating this pattern has been the determination of the bandwidth of the aperture structure, which amounts to finding the narrowest portion of the aperture function and treating it as you would a rectangular aperture.

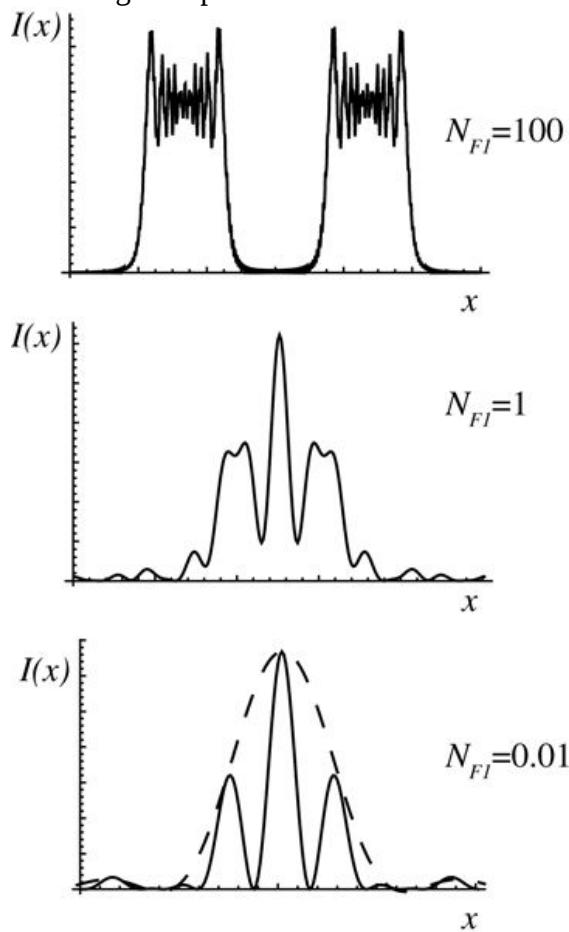


Figure 5.11

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 5.11 Diffraction patterns of the double rectangle aperture for $NF_1 = 100, 1$ and 0.01 . The horizontal scales are different for the three patterns. M_1 has been chosen to be 10 when $NF_1 = 0.01$. The dashed curve represents the sinc^2 normalized intensity distribution due to one of the sub-apertures by itself.

The three graphs are plotted on horizontal axis x and vertical axis $I(x)$. The first graph plots for $N_{F1} = 100$. The curve almost overlaps the horizontal axis for a distance of a few units and then rises almost perpendicularly to a height where it zigzags horizontally and then drops to the horizontal axis, thus forming a near symmetric pattern. The pattern is repeated to the right. The second graph plots for $N_{F1} = 1$. The curve rises and drops in a short wavelike pattern staying close to the horizontal axis and then rises steeply to form a double-peaked wave to fall and then rise higher than before to form a sharp needle-like peak. From the peak onward, the curve moves rightward to mirror the path thus far such that the perpendicular dropped from the peak would serve as a line of symmetry. The third graph plots for $N_{F1} = 0.01$. The curve rises and drops in a short wavelike pattern staying close to the horizontal axis and then rises steeply to a height and falls back to the horizontal axis and then rises again almost two times as high to form a sharp needle-like peak. From the peak onward, the curve moves rightward to mirror the path thus far such that the perpendicular dropped from the peak would serve as a line of symmetry. A dotted line in the shape of a bell runs around the central peak and the two peaks on its sides, connecting the three peaks and the horizontal axis.

5.8.2 Two-Dimensional Apertures Separable in (x, y) Coordinates

In this section we consider apertures that are separable in (x, y) coordinates, described by an amplitude transmittance function

$$tA(x,y) = tX(x)tY(y),$$

$$t_A(x, y) = t_X(x) t_Y(y),$$

(5-58)

where in general $tX(x)$ and $tY(y)$ can have different functional forms. In this case we need to follow the procedure described in the previous section, estimating the bandwidth of each of the factors of amplitude transmittance separately, and choosing the values of M and N separately for each factor, based on [Fig. 5.4](#) or [Fig. 5.5](#). Once the one-dimensional diffraction patterns as a function of x and y are calculated separately, the two-dimensional pattern can be found by multiplying those factors to form a function of (x, y) .

5.8.3 Circularly-Symmetric Apertures

To visualize the diffraction pattern in the circularly symmetric case, two approaches could be used. First, one could define a circular aperture in pixelated form, and apply any of the methods discussed previously to calculate the two-dimensional diffraction pattern. An alternative approach, which we introduce here, allows one to visualize directly the one-dimensional intensity distribution along any line through the center of the two-dimensional diffraction pattern. This approach is less computationally intensive than calculating the full two-dimensional pattern and extracting a one-dimensional slice through the origin.

The approach to be used here utilizes the projection-slice theorem presented in [Section 2.6](#). The procedure is based on the Fresnel transform approach, as detailed in the following:

1. Define the circularly symmetric aperture in pixelated form, making the diameter equal to M pixels, where M is determined from [Fig. 5.4](#) for the particular Fresnel number of interest.
2. Define the sampled quadratic-phase exponential function

$$hpq = \exp j\pi 4NFM^2 p^2 + q^2,$$

$$h_{pq} = \exp \left[j\pi \frac{4N_F}{M^2} (p^2 + q^2) \right],$$

(5-59)

where M is chosen to be an odd number in order to sample at the center of the diffraction pattern and p and q run from $-(M - 1)/2$ to $(M - 1)/2$.

3. Multiply the two arrays element by element.
4. Project the $M \times M$ array onto the horizontal axis, which we take to be the p axis. The projection operation involves simply summing all the elements in each vertical column (i.e. over the q index). The result is a one-dimensional length M array.⁸
5. Pad the length M array with zeros so that its total length is N , as taken from [Fig. 5.4](#) for the given N_F .
6. Perform a one-dimensional DFT on the length N sequence, and take the squared magnitude to obtain intensity.

[Figure 5.12](#) shows two plots of the intensity in the diffraction pattern of circular aperture with $N_F = 10$. The plot in part (a) is obtained by numerical integration of [Eq.\(4-67\)](#), while part (b) is obtained by the projection and FFT method described above. There exist small differences between the results in parts (a) and (b). For example, the intensity at the center of the diffraction pattern in part (a) goes to zero, while it does not quite reach zero in part (b). The reason for the differences is that the pixelated circle used in part (b) is not perfectly circularly symmetric, no matter how large M may be. Thus the differences arise due to the inability to represent a perfect circle with a discrete rectangular array, not due to the method used for calculation in part (b). The larger the choice for M , the closer the center pixel is to zero.⁹

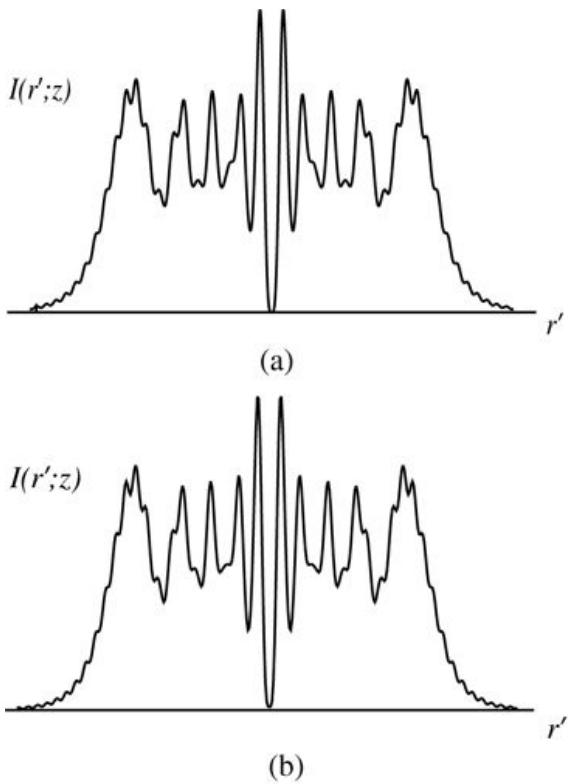


Figure 5.12
Goodman, Introduction to Fourier Optics, 4e,
 © 2017 W. H. Freeman and Company

Figure 5.12 Cross-sections of diffraction patterns calculated for a circular aperture when $NF = 10$. Part (a) shows the pattern obtained by numerical integration of (4-67), while part (b) shows the result obtained by the discrete projection and FFT approach described here.

Plotted along the horizontal axis r' , the two graphs, a and b, are identical plots of $I(r')$. The initial path of the curve is an upward slope, which is followed by horizontal zigzagging that touches the horizontal axis once right at the center. The graphs are symmetrical.

5.8.4 More General Cases

It is difficult to give a definitive recipe for calculating the diffraction pattern of an aperture of arbitrary form. In general, a reasonable approach is to first make a best guess as to the bandwidth of the aperture function, based on the minimum size ℓ of its smallest feature. Then determine the maximum bandwidth of the quadratic-phase function, based on the aperture's maximum width L and the given λ and z . Sample the aperture function with $\Delta x = \Delta y$ equal to the reciprocal of the maximum of these two bandwidths. Define $NF = (L/2)(\lambda z)$

$NF = (L/2)^2 / (\lambda z)$. Choose values of M , N and Q starting with a first guess derived from Fig. 5.4 or Fig. 5.5, depending on which calculation method will be used. Follow the procedure specified for the chosen method from this point on. Check the aliasing level at the edge of the window using a log plot. Increase M if necessary to achieve a desired aliasing criterion. Then try increasing Q to see whether the results change appreciably. When the shape of the

diffraction pattern becomes stable under these increases, you have reached the answer you have been seeking.

5.9 Concluding Comments

There are a few comments regarding the results presented in this chapter that the reader may appreciate. They are addressed in the following list.

1. There is as yet no underlying theory as to how to choose M most efficiently in the Fresnel transform and the Fresnel transfer function cases. Such a choice would have to satisfy $M > 4N_F$, and in addition M should be large enough to satisfy whatever aliasing criterion has been chosen. Beyond these two requirements, there does not yet seem to exist a theory that leads to an optimum choice of M . For this reason we resorted to simulation results to determine the best values of M for each N_F in the two cases.
2. The Fresnel transform approach and the Fresnel transfer function approach, as described in [Section 5.5](#), do not always yield exactly the same results. One can choose precisely the same values for M , Q and N for the two methods, and in some cases the width of the resulting diffraction pattern will be a smaller fraction of the length N diffraction pattern sequence when using the Fresnel transfer function approach, as compared with the fraction when using the Fresnel transform approach. This appears to be due to the different effects of Q in the two cases. In the Fresnel transform approach, the choice of Q determines the spacing of samples in the diffraction pattern. In the Fresnel transfer function approach, the choice of Q determines the spacing of samples in the frequency domain, thereby assuring that the lengths of the sequences in the space domain and the frequency domain are equal. However, in the Fresnel transfer function approach, increasing Q will not change the number of samples in the region where the calculated diffraction pattern has significant value. As a result, the width of the diffraction pattern will be a smaller fraction of the total output sequence. However, the results within the diffraction patterns themselves agree well for the two methods.
3. It has been commonly believed that the Fresnel transform approach is the most efficient approach for small N_F (long distances from the aperture) while the Fresnel transfer function approach is most efficient for large N_F (short distances from the aperture), due to their different dependencies on z . However, the results presented here suggest that, particularly for an intensity aliasing criterion smaller than 10^{-2} , the Fresnel transform approach is the most efficient of all approaches, for all values of N_F in the range 0.01 to 100.
4. The exact transfer function method has the appeal of being more accurate than the Fresnel transfer function method, but in practice the results of the two methods differ primarily under conditions for which the scalar approximation is suspect.

We conclude that there is room for the development of further understanding for this important topic.

Problems - Chapter 5

1. 5-1. Consider a one-dimensional finite length quadratic-phase exponential of the form

$$g(x) = \exp j\pi\lambda z x^2 \quad |x| \leq L/2 \text{ otherwise.}$$

$$g(x) = \begin{cases} \exp(j\frac{\pi}{\lambda z} x^2) & |x| \leq L/2 \\ 0 & \text{otherwise.} \end{cases}$$

Show that the local frequency distribution $f(\ell)(x)$ of this function is given by

$$f(\ell)(x) = x\lambda z,$$

$$f^{(\ell)}(x) = \frac{x}{\lambda z},$$

and consequently that, to a good approximation when the Fresnel number is larger than 0.25, the spectrum of $g(x)$ has no frequency components higher than $L/2\lambda z$. The spectrum of this function is therefore approximately limited to the region

$$-L/2\lambda z \leq f_x \leq L/2\lambda z.$$

$$-\frac{L}{2\lambda z} \leq f_x \leq \frac{L}{2\lambda z}.$$

2. 5-2. a. Show that for $N_F > 1$, the finest structure in the Fresnel diffraction pattern of a rectangular aperture has a spatial dimension of about $1/(4N_F)^{1/2}$ of the aperture width ℓ .

3. b. Show that under the same conditions, the number of samples within the aperture should satisfy $M \geq 4N_F$.

4. 5-3. A certain one-dimensional aperture characterized by an amplitude transmittance of the form

$$t_A(x) = \text{rect}(x/\ell) \exp(-4x^2/\ell^2)$$

$$t_A(x) = \text{rect}(x/\ell) \exp\left[-4\left(\frac{x}{\ell}\right)^2\right]$$

is illuminated by a unit amplitude monochromatic plane wave. Use any of the numerical diffraction methods we have covered to plot the intensity distribution in the diffraction plane for the following cases: (a) $N_F = 10$; (b) $N_F = 500$.

5. 5-4. A certain circularly symmetric aperture of diameter ℓ has an amplitude transmittance of unity between radius $\ell/2$ and radius $\ell/4$, and zero otherwise. Assume that this aperture is illuminated by a unit-amplitude monochromatic plane wave.

1. Use the Fourier Bessel transform to find a radial profile of the Fresnel diffraction pattern of this aperture for $N_F = 10$. Plot the result.
2. Use the numerical technique based on the projection-slice theorem and a DFT to find the same radial profile.

6 Wave-Optics Analysis of Coherent Optical Systems

The most important components of most optical systems are lenses. While a thorough discussion of geometrical optics and the properties of lenses would be helpful, such a treatment would require a rather lengthy detour. To provide the most rudimentary background, [Appendix B](#) presents a short description of the matrix theory of paraxial geometrical optics, defining certain quantities that will be important in our purely “wave-optics” approach in this chapter. The reader will be referred to appropriate material in the appendix when needed. However, the philosophy of our approach is to make minimum use of geometrical optics, and instead to develop purely wave-optic analyses of the systems of interest. The results of this approach are entirely consistent with the results of geometrical optics, with the added advantage that diffraction effects are entirely accounted for in the wave-optics approach, but not in the geometrical-optics approach. Our discussions will be limited to the case of monochromatic illumination, with generalization to nonmonochromatic light being deferred to [Chapter 7](#).

6.1 A Thin Lens as a Phase Transformation

A lens is composed of an optically dense material, usually glass with a refractive index of approximately 1.5, in which the propagation velocity of an optical disturbance is less than the velocity in air. With reference to [Appendix B](#), a lens is said to be a *thin lens* if a ray entering at coordinates (x, y) on one face exits at approximately the same coordinates on the opposite face, i.e. if there is negligible translation of a ray within the lens. Thus a thin lens simply delays an incident wavefront by an amount proportional to the thickness of the lens at each point.

Referring to [Fig. 6.1](#), let the maximum thickness of the lens (on its axis) be Δ_0 , and let the thickness at coordinates (x, y) be $\Delta(x, y)$. Then the total phase delay suffered by the wave at coordinates (x, y) in passing through the lens may be written

$$\phi(x, y) = kn\Delta(x, y) + k[\Delta_0 - \Delta(x, y)]$$

$$\phi(x, y) = kn\Delta(x, y) + k[\Delta_0 - \Delta(x, y)]$$

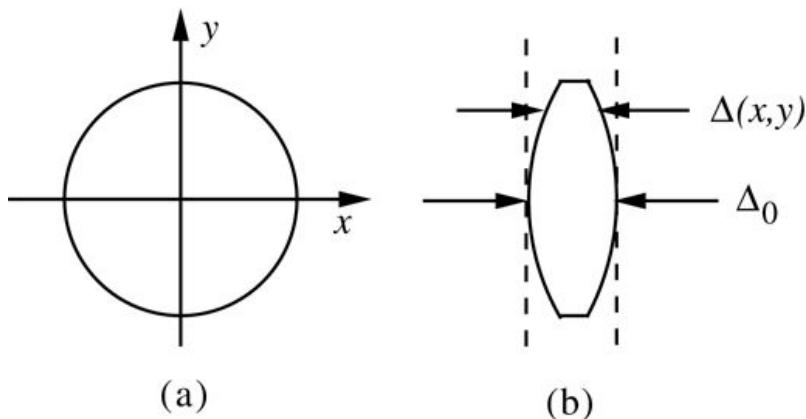


Figure 6.1

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 6.1 The thickness function. (a) Front view, (b) side view

Figure a shows a circle whose center is at the intersection of horizontal axis x and vertical axis y. Figure b shows a side view of a vertical lens that is convex on both sides of a rectangular section such that the curves are tangential to two parallel dotted vertical lines. The width of the lens at the center, where it is the thickest, is delta zero. The width near the top end is delta (x, y) .

where n is the refractive index of the lens material, $kn\Delta(x, y)$ is the phase delay introduced by the lens, and $k[\Delta_0 - \Delta(x, y)]$ is the phase delay introduced by the remaining region of free space between the two planes. Equivalently the lens may be represented by a multiplicative phase transformation of the form

$$t_l(x,y) = \exp[jk\Delta_0] \exp[jk(n-1)\Delta(x,y)].$$

$$t_l(x, y) = \exp [jk\Delta_0] \exp [jk(n - 1)\Delta(x, y)].$$

(6-1)

The complex field $U_l'(x,y)$ across a plane immediately behind the lens is then related to the complex field $U_l(x,y)$ incident on a plane immediately in front of the lens by

$$U_l'(x,y) = t_l(x,y) U_l(x,y).$$

$$U_l'(x, y) = t_l(x, y) U_l(x, y).$$

(6-2)

The problem remains to find the mathematical form of the thickness function $\Delta(x,y)$ in order that the effects of the lens may be understood.

6.1.1 The Thickness Function

In order to specify the forms of the phase transformations introduced by a variety of different types of lenses, we first adopt a sign convention: as rays travel from left to right, each *convex* surface encountered is taken to have a *positive* radius of curvature, while each *concave* surface is taken to have a *negative* radius of curvature. Thus in [Fig. 6.1\(b\)](#) the radius of curvature of the left-hand surface of the lens is a positive number R_1 , while the radius of curvature of the right-hand surface is a negative number R_2 .

To find the thickness $\Delta(x,y)$, we split the lens into three parts, as shown in [Fig. 6.2](#), and write the total thickness function as the sum of three individual thickness functions,

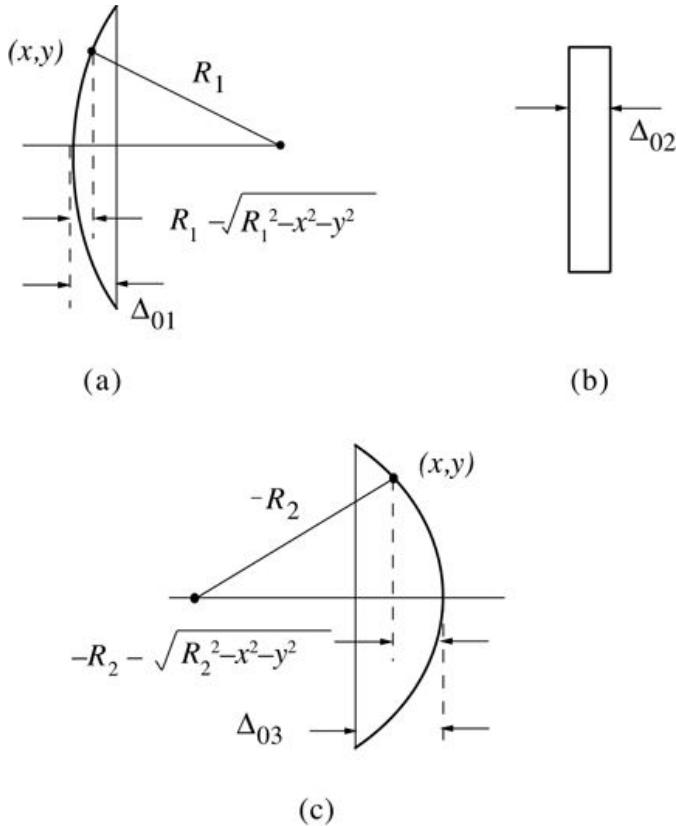


Figure 6.2
Goodman, Introduction to Fourier Optics, 4e,
 © 2017 W. H. Freeman and Company

Figure 6.2 Calculation of the thickness function. (a) Geometry for Δ_1 , (b) geometry for Δ_2 , and (c) geometry for Δ_3 .

Diagram a shows a side view of the left-hand surface of the lens that is convex on the left and plane on the right; the radius of curvature, measuring R_1 , connects the center of curvature to point (x, y) on the curved surface in the upper half of the lens. A horizontal line from the center of curvature passes through the thickest part of the lens. The distance between a vertical dotted line tangential to the curvature at the thickest point and the vertical line dropped from (x, y) is marked R_1 minus square root of (R_1 squared minus x squared minus y squared). The distance between the vertical dotted line tangential to the curvature at the thickest point and the plane on the right side of the lens is marked delta 0 1. Diagram b shows a rectangular cross-section that is vertically long, the narrow horizontal width measuring delta 02. Diagram c shows a side view of the right-hand surface of the lens that is convex on the right and plane on the left; the radius of curvature, measuring minus R_2 , connects the center of curvature to point (x, y) on the curved surface in the upper half of the lens. A horizontal line from the center of curvature passes through the thickest part of the lens. The distance between a vertical dotted line tangential to the curvature at the thickest point and the vertical line dropped from (x, y) is marked minus R_2 minus square root of (R_2 squared minus x squared minus y squared). The distance between the vertical dotted line tangential to the curvature at the thickest point and the plane on the right side of the lens is marked delta 0 3.

$$\Delta(x,y) = \Delta_1(x,y) + \Delta_2(x,y) + \Delta_3(x,y).$$

$$\Delta(x, y) = \Delta_1(x, y) + \Delta_2(x, y) + \Delta_3(x, y).$$

(6-3)

Referring to the geometries shown in that figure, the thickness function $\Delta_1(x,y)$ is given by

$$\Delta_1(x,y) = \Delta_{01} - R_1 - \sqrt{R_1^2 - x^2 - y^2}$$

$$\begin{aligned}\Delta_1(x, y) &= \Delta_{01} - \left(R_1 - \sqrt{R_1^2 - x^2 - y^2} \right) \\ &= \Delta_{01} - R_1 \left(1 - \sqrt{1 - \frac{x^2 + y^2}{R_1^2}} \right).\end{aligned}$$

(6-4)

The second component of the thickness function comes from a region of glass of constant thickness Δ_{02} . The third component is given by

$$\Delta_3(x,y) = \Delta_{03} - R_2 - \sqrt{R_2^2 - x^2 - y^2}$$

$$\begin{aligned}\Delta_3(x, y) &= \Delta_{03} - \left(R_2 - \sqrt{R_2^2 - x^2 - y^2} \right) \\ &= \Delta_{03} + R_2 \left(1 - \sqrt{1 - \frac{x^2 + y^2}{R_2^2}} \right),\end{aligned}$$

(6-5)

where we have factored the positive number $-R_2 - R_2$ out of the square root. Combining the three expressions for thickness, the total thickness is seen to be

$$\Delta(x,y) = \Delta_0 - R_1 - \sqrt{R_1^2 - x^2 - y^2} + R_2 - \sqrt{R_2^2 - x^2 - y^2}$$

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

(6-6)

$$\text{where } \Delta_0 = \Delta_{01} + \Delta_{02} + \Delta_{03}.$$

6.1.2 The Paraxial Approximation

The expression for the thickness function can be substantially simplified if attention is restricted to portions of the wavefront that lie near the lens axis, or equivalently, if only *paraxial* rays are

considered. Thus we consider only values of x^x and y^y sufficiently small to allow the following approximations to be accurate:

$$1-x^2+y^2R_1 \approx 1-x^2+y^2R_1, \quad 1-x^2+y^2R_2 \approx 1-x^2+y^2R_2.$$

$$\begin{aligned}\sqrt{1 - \frac{x^2 + y^2}{R_1^2}} &\approx 1 - \frac{x^2 + y^2}{2R_1^2} \\ \sqrt{1 - \frac{x^2 + y^2}{R_2^2}} &\approx 1 - \frac{x^2 + y^2}{2R_2^2}.\end{aligned}$$

(6-7)

The resulting phase transformation will, of course, represent the lens accurately over only a limited area, but this limitation is no more restrictive than the usual paraxial approximation of geometrical optics. Note that the relations (6-7) amount to approximations of the spherical surfaces of the lens by quadratic-phase surfaces. With the help of these approximations, the thickness function becomes

$$\Delta(x, y) = \Delta_0 - x^2 + y^2 R_1 - 1/R_2.$$

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

(6-8)

6.1.3 The Phase Transformation and Its Physical Meaning

Substitution of (6-8) into (6-1) yields the following approximation to the lens transformation:

$$t_l(x, y) = \exp[jkn\Delta_0] \exp[-jk(n-1)x^2 + y^2 R_1 - 1/R_2].$$

$$t_l(x, y) = \exp[jkn\Delta_0] \exp \left[-jk(n-1) \frac{x^2 + y^2}{2} \left(\frac{1}{R_1} - \frac{1}{R_2} \right) \right].$$

The physical properties of the lens (that is, n , R_1 , and R_2) can be combined in a single number f , called the *focal length*, which is defined by

$$f \equiv (n-1)R_1 - 1/R_2.$$

$$\frac{1}{f} \equiv (n-1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right).$$

(6-9)

Neglecting the constant phase factor, which we shall drop hereafter, the phase transformation may now be rewritten

$$t_l(x,y) = \exp -jk2f(x^2+y^2) = \exp -j\pi\lambda f(x^2+y^2).$$

$$t_l(x, y) = \exp \left[-j \frac{k}{2f} (x^2 + y^2) \right] = \exp \left[-j \frac{\pi}{\lambda f} (x^2 + y^2) \right].$$

(6-10)

This equation will serve as our basic representation of the effects of a thin lens on an incident disturbance. It neglects the finite extent of the lens, which we will account for later.

Note that while our derivation of this expression assumed the specific lens shape shown in [Fig. 6.1](#), the sign convention adopted allows the result to be applied to other types of lenses. [Figure 6.3](#) illustrates several different types of lenses with various combinations of convex and concave surfaces. In [Prob. 6-1](#), the reader is asked to verify that the sign convention adopted implies that the focal length f of a double-convex, plano-convex, or positive meniscus lens is *positive*, while that of a double-concave, plano-concave, or negative meniscus lens is *negative*. Thus [\(6-10\)](#) can be used to represent any of the above lenses, provided the correct sign of the focal length is used.

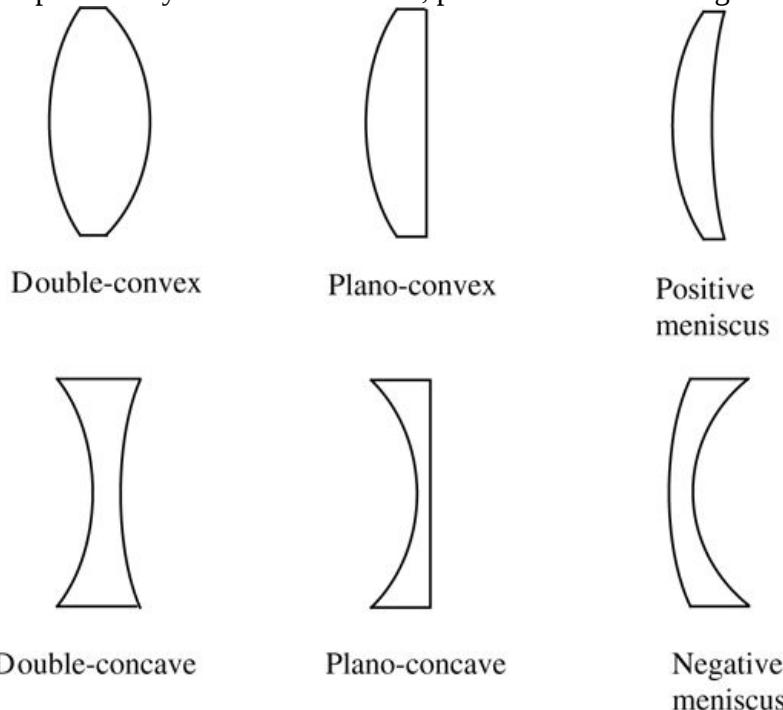


Figure 6.3
Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 6.3 Various types of lenses.

The lens types are as follows. Double convex: It is rectangular in the center with a convex curvature on each side. Plano-convex: It is rectangular with a convex curvature on one side and no curvature on the other. Positive meniscus: It has a convex curvature on the left and a concave curvature on the right such that it is thickest at the center and thinnest at the periphery. Double concave: It is rectangular in the center with a concave curvature on each side. Plano-concave: It is rectangular with a concave curvature on one side and no curvature on the other. Negative

meniscus: It has a convex curvature on the left and a concave curvature on the right such that it is thinnest at the center and thickest at the periphery.

The same result can be derived in one dimension by an entirely different method, namely by the use of ray-transfer matrices. We know from (B-14) that the ray transfer matrix for a thin lens is

$$M=10-1/f$$

$$M = \begin{bmatrix} 1 & 0 \\ -1/f & 1 \end{bmatrix}$$

(6-11)

when the refractive index of the material before and after the lens is unity. Suppose that a unit-amplitude plane wave is normally incident on this lens. The ray vector of the incident wave is

$$v \rightarrow 1 = y 0,$$

$$\vec{v}_1 = \begin{bmatrix} y \\ 0 \end{bmatrix},$$

(6-12)

where y^y is a vertical position in the plane of the thin lens. The ray vector of the wave transmitted by the thin lens is

$$v \rightarrow 2 = 10-1/f y 0 = y - y/f.$$

$$\vec{v}_2 = \begin{bmatrix} 1 & 0 \\ -1/f & 1 \end{bmatrix} \begin{bmatrix} y \\ 0 \end{bmatrix} = \begin{bmatrix} y \\ -y/f \\ f \end{bmatrix}.$$

(6-13)

From this result it is clear that the ray emerges at the same y^y -coordinate with which it entered the lens, but the ray angle is changed from $\theta_1 = 0$ to $\theta_2 = -y/f$. From (B-9) we know that the angle θ_2 is related to the local spatial frequency of the wave through

$$f(l)(y) = \theta_2 \lambda = -y/f.$$

$$f^{(l)}(y) = \frac{\theta_2}{\lambda} = -\frac{y}{\lambda f}.$$

(6-14)

But a linear dependence of local frequency on y^y corresponds to a quadratic-phase wavefront, and the ratio of the complex amplitude of the transmitted wave to the complex amplitude of the incident wave is therefore given by¹

$$t_l(y) = \exp - j \pi \lambda f y^2,$$

$$t_l(y) = \exp \left[-j \frac{\pi}{\lambda f} y^2 \right],$$

(6-15)

the one-dimensional analog of (6-10).

Returning to the two-dimensional case, the physical meaning of the lens transformation can be understood by again considering the effect of the lens on a normally incident, unit-amplitude plane wave. The field distribution U_l in front of the lens is unity, and (6-1) and (6-10) yield the following expression for U'_l behind the lens:

$$U'_l(x, y) = \exp[-jk2f(x^2 + y^2)].$$

$$U'_l(x, y) = \exp\left[-j\frac{k}{2f}(x^2 + y^2)\right].$$

We may interpret this expression as a quadratic-phase approximation to a spherical wave. If the focal length is positive, then the spherical wave is converging towards a point on the lens axis a distance f behind the lens. If f is negative, then the spherical wave is diverging from a point on the lens axis a distance $|f|$ in front of the lens. The two cases are illustrated in Fig. 6.4. Thus a lens with a positive focal length is called a *positive or converging lens*, while a lens with a negative focal length is a *negative or diverging lens*.

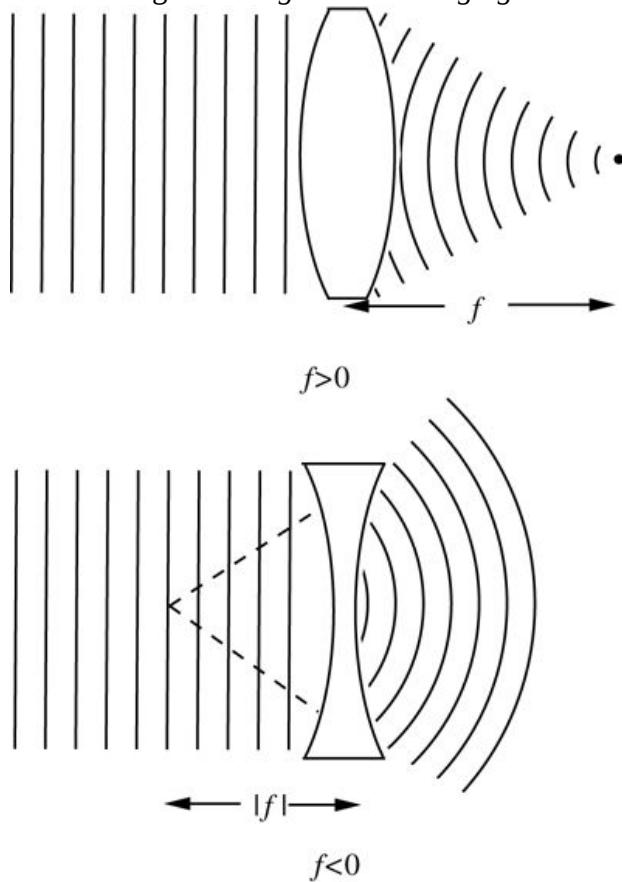


Figure 6.4
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 6.4 Effects of a converging lens and a diverging lens on a normally incident plane wave.

The illustration for $f > 0$ shows a double convex lens with successive vertical lines on the left. On the other side, successive concentric curves opening to the right diminish in length as they move up to a point that is at a distance f from the center of the lens. The illustration for $f < 0$ shows a double concave lens with successive vertical lines on the left. On the other side, successive concentric curves opening to the left increase in length as they move to the right. The shared center of the curves is at a distance of absolute value of f from the center of the lens.

Our conclusion that a lens composed of spherical surfaces maps an incident plane wave into a spherical wave is very much dependent on the paraxial approximation. Under non-paraxial conditions, the emerging wavefront will exhibit departures from perfect sphericity (called *aberrations* — see [Section 7.4](#)), even if the surfaces of the lens *are* perfectly spherical. In fact, lenses are often “corrected” for aberrations by making their surfaces aspherical in order to improve the sphericity of the emerging wavefront.

We should emphasize, however, that the results which will be derived using the multiplicative phase transformation [\(6-10\)](#) are actually more general than the analysis leading up to that equation might imply. When designing an image forming system, the lens designer strives to have nonparaxial rays come to the same image point as the paraxial rays, so that all rays come to a common point. For this reason, a paraxial treatment gives the same first-order behavior as a nonparaxial one. Thus a thorough geometrical-optics analysis of most well-corrected lens systems shows that they behave essentially in the way predicted by our more restrictive theory.

6.2 Fourier Transforming Properties of Lenses

One of the most remarkable and useful properties of a converging lens is its inherent ability to perform two-dimensional Fourier transforms when combined with proper free-space propagation before and after the lens. This complicated analog operation can be performed with extreme simplicity in a coherent optical system, taking advantage of the basic laws of propagation and diffraction of light.

In the material that follows, several different configurations for performing the transform operation are described. In all cases the illumination is assumed to be monochromatic. Under this condition the systems studied are “coherent” systems, which means that they are linear in complex amplitude, and the distribution of light amplitude across a particular plane behind the positive lens is of interest. In some cases this is the *back focal plane* of the lens, which by definition is a plane normal to the lens axis situated a distance f behind the lens (in the direction of propagation of light). The information to be Fourier-transformed is introduced into the optical system by a structure with an amplitude transmittance that is proportional to the input function of interest. In some cases this device may consist of a photographic transparency, while in others it may be a nonphotographic *spatial light modulator*, capable of controlling the amplitude transmittance in response to externally supplied electrical or optical information. Such input devices will be discussed in more detail in [Chapter 9](#). We will refer to them as input “transparencies,” even though in some cases they may operate by reflection of light rather than transmission of light. We will also often refer to the input as the “object.”

[Figure 6.5](#) shows three arrangements that will be considered here. In all cases shown, the illumination is a collimated plane wave which is incident either on the input transparency or on the lens. In case (a), the input transparency is placed directly against the lens itself. In case (b), the input is placed a distance d in front of the lens. In case (c), the input is placed behind the lens at distance d from the focal plane. An additional, more general case, will be studied in [Section 6.4](#).

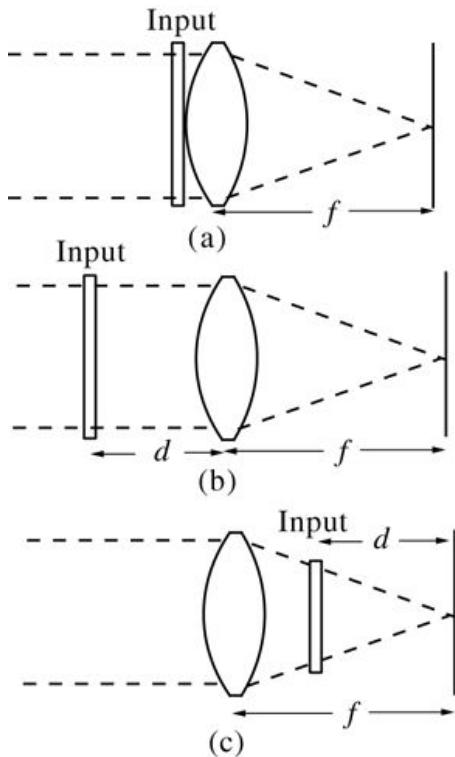


Figure 6.5

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 6.5 Geometries for performing the Fourier transform operation with a positive lens.

In all three diagrams, rays from the left fall on a double convex lens and converge on the other side at the center of a vertical line representing the focal plane. The horizontal distance between the center of the lens and the focal plane is f . In diagram a, a vertically placed thin rectangle, representing input transparency, is on the left side and in contact with the convex surface. In diagram b, the input transparency is on the left side at a distance d from the center of the lens. In diagram c, the input transparency is on the right side at distance d from the center of the focal plane; d is less than f .

For alternative discussions of the Fourier transforming properties of positive lenses, the reader may wish to consult [297], [81], or [288].

6.2.1 Input Placed against the Lens

Let a planar input transparency with amplitude transmittance $t_A(x, y)$ be placed immediately in front of a converging lens of focal length f , as shown in Fig. 6.5(a). The input is assumed to be uniformly illuminated by a normally incident, monochromatic plane wave of amplitude A , in which case the disturbance incident on the lens is

$$U_l(x, y) = A t_A(x, y).$$

$$U_l(x, y) = A t_A(x, y).$$

(6-16)

The finite extent of the lens can be accounted for by associating with the lens a *pupil function* $P(x, y)$ defined by

$$P(x, y) = \begin{cases} 1 & \text{inside the lens aperture} \\ 0 & \text{otherwise.} \end{cases}$$

$$P(x, y) = \begin{cases} 1 & \text{inside the lens aperture} \\ 0 & \text{otherwise.} \end{cases}$$

Thus the amplitude distribution behind the lens becomes, using (6-10),

$$U_l'(x, y) = U_l(x, y) P(x, y) \exp[-jk2f(x^2 + y^2)].$$

$$U_l'(x, y) = U_l(x, y) P(x, y) \exp\left[-j\frac{k}{2f}(x^2 + y^2)\right].$$

(6-17)

To find the distribution $U_f(u, v)$ in the back focal plane of the lens, the Fresnel diffraction formula, (4-17), is applied. Thus, putting $z=f$,

$$U_f(u, v) = \exp[jk2f(u^2 + v^2)] \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U_l'(x, y) \exp\left[j\frac{k}{2f}(x^2 + y^2)\right] \exp\left[-j\frac{2\pi}{\lambda f}(xu + yv)\right] dx dy,$$

$$\begin{aligned} U_f(u, v) &= \frac{\exp\left[j\frac{k}{2f}(u^2 + v^2)\right]}{j\lambda f} \\ &\times \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U_l'(x, y) \exp\left[j\frac{k}{2f}(x^2 + y^2)\right] \exp\left[-j\frac{2\pi}{\lambda f}(xu + yv)\right] dx dy, \end{aligned}$$

(6-18)

where a constant phase factor has been dropped. Substituting (6-17) in (6-18), the quadratic phase factors within the integrand are seen to exactly cancel, leaving

$$U_f(u, v) = \exp[jk2f(u^2 + v^2)] \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U_l(x, y) P(x, y) \exp[-j\frac{2\pi}{\lambda f}(xu + yv)] dx dy.$$

$$\begin{aligned} U_f(u, v) &= \frac{\exp\left[j\frac{k}{2f}(u^2 + v^2)\right]}{j\lambda f} \\ &\times \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U_l(x, y) P(x, y) \exp\left[-j\frac{2\pi}{\lambda f}(xu + yv)\right] dx dy. \end{aligned}$$

(6-19)

Thus the field distribution U_f is proportional to the two-dimensional Fourier transform of that portion of the incident field subtended by the lens aperture. When the physical extent of the input is smaller than the lens aperture, the factor $P(x, y)$ may be neglected, yielding

$$U_f(u, v) = \exp(jk2f(u^2+v^2)) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U_l(x, y) \exp(-j\frac{2\pi}{\lambda f}(xu+yu)) dx dy.$$

$$U_f(u, v) = \frac{1}{j\lambda f} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U_l(x, y) \exp\left[-j\frac{2\pi}{\lambda f}(xu+yu)\right] dx dy.$$

(6-20)

Thus we see that the complex amplitude distribution of the field in the focal plane of the lens is the *Fraunhofer diffraction pattern* of the field incident on the lens, even though the distance to the observation plane is equal to the focal length of the lens, rather than satisfying the usual distance criterion for observing Fraunhofer diffraction. Note that the amplitude and phase of the light at coordinates (u, v) in the focal plane are determined by the amplitude and phase of the input Fourier component at frequencies $(f_X = u/\lambda f, f_Y = v/\lambda f)$.

The Fourier transform relation between the input amplitude transmittance and the focal-plane amplitude distribution is not a complete one, due to the presence of the quadratic-phase factor that precedes the integral. While the phase distribution across the focal plane is not the same as the phase distribution across the spectrum of the input, the difference between the two is a simple phase curvature.

In most cases it is the *intensity* across the focal plane that is of real interest. This phase term is important if the ultimate goal is to calculate another field distribution after further propagation and possibly passage through additional lenses, in which case the complete complex field is needed. In most cases, however, the *intensity* distribution in the focal plane will be measured, and the phase distribution is of no consequence. Measurement of the intensity distribution yields knowledge of the *power spectrum* (or more accurately, the *energy spectrum*) of the input. Thus

$$I_f(u, v) = A^2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} t_A(x, y) \exp(-j2\pi\lambda f(xu+yu)) dx dy.$$

$$I_f(u, v) = \frac{A^2}{\lambda^2 f^2} \left| \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} t_A(x, y) \exp\left[-j\frac{2\pi}{\lambda f}(xu+yu)\right] dx dy \right|^2.$$

(6-21)

6.2.2 Input Placed in Front of the Lens

Consider next the more general geometry of Fig. 6.5(b). The input, located a distance d in front of the lens, is illuminated by a normally incident plane wave of amplitude A . The amplitude transmittance of the input is again represented by t_A . In addition, let $F_o(f_X, f_Y)$ represent the Fourier spectrum of the light transmitted by the input transparency, and $F_l(f_X, f_Y)$ the Fourier spectrum of the light incident on the lens; that is,

$$F_o(f_X, f_Y) = \mathcal{F}AtA F_l(f_X, f_Y) = \mathcal{F}U_l.$$

$$F_o(f_X, f_Y) = \mathcal{F}\{At_A\}F_l(f_X, f_Y) = \mathcal{F}\{U_l\}.$$

Assuming that the Fresnel or paraxial approximation is valid for propagation over distance d , then F_o and F_l are related by means of [Eq.\(4-20\)](#), giving

$$F_l(f_X, f_Y) = F_o(f_X, f_Y) \exp[-j\pi\lambda d(f_X^2 + f_Y^2)],$$

$$F_l(f_X, f_Y) = F_o(f_X, f_Y) \exp[-j\pi\lambda d(f_X^2 + f_Y^2)],$$

(6-22)

where we have dropped a constant phase delay.

For the moment, the finite extent of the lens aperture will be neglected. Thus, letting $P=1$ $P = 1$, [\(6-19\)](#) can be rewritten

$$U_f(u, v) = \exp[jk2f(u^2 + v^2)] \exp[-j\lambda f u \lambda f, v \lambda f].$$

$$U_f(u, v) = \frac{\exp[j\frac{k}{2f}(u^2 + v^2)]}{j\lambda f} F_l\left(\frac{u}{\lambda f}, \frac{v}{\lambda f}\right).$$

(6-23)

Substituting [\(6-22\)](#) into [\(6-23\)](#), we have

$$U_f(u, v) = \exp[jk2f(u^2 + v^2)] \exp[-j\lambda f u \lambda f, v \lambda f],$$

$$U_f(u, v) = \frac{\exp[j\frac{k}{2f}(1 - \frac{d}{f})(u^2 + v^2)]}{j\lambda f} F_o\left(\frac{u}{\lambda f}, \frac{v}{\lambda f}\right),$$

or

$$U_f(u, v) = A \exp[jk2f(1 - d/f)(u^2 + v^2)] \exp[-j\lambda f u \lambda f, v \lambda f] \times \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} t_A(\xi, \eta) \exp[-j\frac{2\pi}{\lambda f}(\xi u + \eta v)] d\xi d\eta.$$

$$\begin{aligned} U_f(u, v) &= \frac{A \exp[j\frac{k}{2f}(1 - \frac{d}{f})(u^2 + v^2)]}{j\lambda f} \\ &\times \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} t_A(\xi, \eta) \exp[-j\frac{2\pi}{\lambda f}(\xi u + \eta v)] d\xi d\eta. \end{aligned}$$

(6-24)

Thus the amplitude and phase of the light at coordinates (u, v) are again related to the amplitude and phase of the input spectrum at frequencies $(u/\lambda f, v/\lambda f)$. Note that a quadratic-phase factor again precedes the transform integral, but that it vanishes for the very special case $d=f$. Evidently when the input is placed in the front focal plane of the lens, the phase curvature disappears, leaving an exact Fourier transform relation!

To this point we have entirely neglected the finite extent of the lens aperture. To include the effects of this aperture, we use a geometrical optics approximation. Such an approximation is accurate if the distance (u_1, v_1) is sufficiently small to place the input deep within the region of Fresnel diffraction of the lens aperture if the light were propagating backwards from the focal plane to the plane of the input transparency. This condition is well satisfied in the vast majority of problems of interest. With reference to Fig. 6.6, the light amplitude at coordinates d is a summation of all the rays traveling with direction cosines (u_1, v_1) . However, only a finite set of these rays is passed by the lens aperture. Thus the finite extent of the aperture may be accounted for by geometrically projecting that aperture back to the input plane, the projection being centered on a line joining the coordinates $(\xi \approx u_1/f, \eta \approx v_1/f)$ with the center of the lens (see Fig. 6.6). The projected lens aperture limits the effective extent of the input, but the particular portion of t_A that contributes to the field U_f depends on the particular coordinates (u_1, v_1) being considered in the back focal plane. As implied by Fig. 6.6, the value of U_f at (u, v) can be found from the Fourier transform of that portion of the input subtended by the projected pupil function P , centered at coordinates $[\xi = -(d/f)u, \eta = -(d/f)v]$. Expressing this fact mathematically,

$$U_f(u, v) = A \exp[jk2f(1-d)f(u^2 + v^2)] \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} t_A(\xi, \eta) P(\xi, \eta) \exp[-j\frac{2\pi}{\lambda f}(\xi u + \eta v)] d\xi d\eta.$$

$$U_f(u, v) = \frac{A \exp[j\frac{k}{2f}(1 - \frac{d}{f})(u^2 + v^2)]}{j\lambda f} \times \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} t_A(\xi, \eta) P\left(\xi + \frac{d}{f}u, \eta + \frac{d}{f}v\right) \exp\left[-j\frac{2\pi}{\lambda f}(\xi u + \eta v)\right] d\xi d\eta.$$

(6-25)

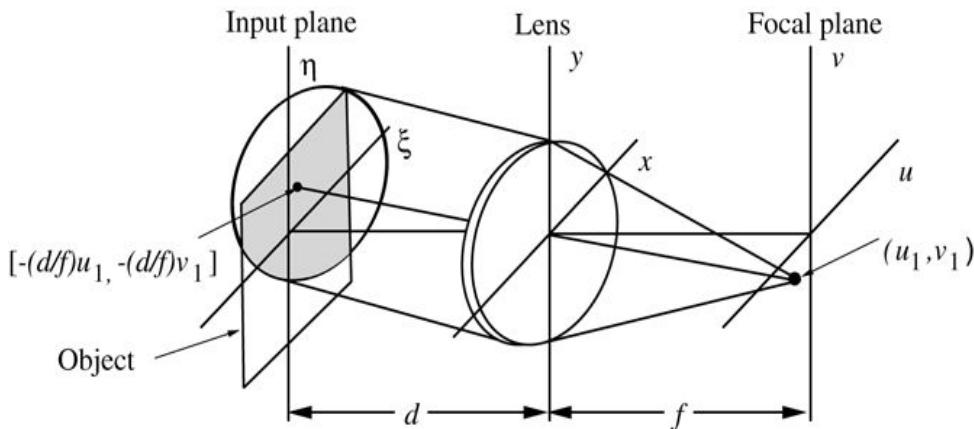


Figure 6.6

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 6.6 Vignetting of the input. The shaded area in the input plane represents the portion of the input transparency that contributes to the Fourier transform at (u_1, v_1) .

The illustration shows a lens between an input plane and a focal plane, each in a three 3D plane. The lens is set on an x y plane, where x is horizontal and y is vertical. The focal plane is set on a u v plane, where u is horizontal and v is vertical. The input plane is set on an ξ η plane, where ξ is horizontal and η is vertical. The horizontal distance between the lens and the focal plane is f and between the lens and the input plane is d . The three planes share a horizontal axis connecting (u_1, v_1) on the focal place, the center of the lens, and a point on the input plane marked [minus $(d/f) u_1$, minus $(d/f) v_1$], which lies in a shaded area that marks the overlap between the vertical rectangular object in the input plane and the circular projection of lens aperture.

The limitation of the effective input by the finite lens aperture is known as a *vignetting* effect. Note that for a simple Fourier transforming system, vignetting of the input space is minimized when the input is placed close to the lens and when the lens aperture is much larger than the input transparency. In practice, when the Fourier transform of the object is of prime interest, it is often preferred to place the input directly against the lens in order to minimize vignetting, although in analysis it is generally convenient to place the input in the front focal plane, where the transform relation is unencumbered with quadratic-phase factors.

6.2.3 Input Placed behind the Lens

Consider next the case of an input that is placed behind the lens, as illustrated in [Fig. 6.5\(c\)](#). The input again has amplitude transmittance t_A , but it is now located a distance d in front of the rear focal plane of the lens. Let the lens be illuminated by a normally incident plane wave of uniform amplitude A . Then incident on the input is a spherical wave converging towards the back focal point of the lens.

In the geometrical optics approximation, the amplitude of the spherical wave impinging on the object is Af/d , due to the fact that the linear dimension of the circular converging bundle of rays has been reduced by the factor d/f and energy has been conserved. The particular region of the input that is illuminated is determined by the intersection of the converging cone of rays with the input plane. If the lens is circular and of diameter l , then a circular region of diameter ld/f is illuminated on the input. The finite extent of the illuminating spot can be represented mathematically by projecting the pupil function of the lens down the cone of rays to the intersection with the input plane, yielding an effective illuminated region in that plane described by the pupil function $P[\xi(f/d), \eta(f/d)]$. Note that the input amplitude transmittance t_A will also have a finite aperture associated with it; the effective aperture in the input space is therefore determined by the intersection of the true input aperture with the projected pupil function of the lens. If the finite input transparency is fully illuminated by the converging light, then the projected pupil can be ignored.

Using a paraxial approximation to the spherical wave that illuminates the input, the amplitude of the wave transmitted by the input may be written

$$U_o(\xi, \eta) = Af d P[\xi(f/d), \eta(f/d)] e^{-jk2d(\xi^2 + \eta^2)} t_A(\xi, \eta).$$

$$U_o(\xi, \eta) = \left\{ \frac{Af}{d} P\left(\frac{\xi f}{d}, \frac{\eta f}{d}\right) \exp\left[-j\frac{k}{2d}(\xi^2 + \eta^2)\right] \right\} t_A(\xi, \eta).$$

(6-26)

Assuming Fresnel diffraction from the input plane to the focal plane, (4-17) can be applied to the field transmitted by the input. If this is done it is found that the quadratic-phase exponential in (ξ, η) associated with the illuminating wave *exactly cancels* the similar quadratic-phase exponential in the integrand of the Fresnel diffraction integral, with the result

$$U_f(u, v) = A \exp[jk2d(u^2 + v^2)] \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} t_A(\xi, \eta) P\left(\frac{\xi f}{d}, \frac{\eta f}{d}\right) \exp\left[-j\frac{2\pi}{\lambda d}(u\xi + v\eta)\right] d\xi d\eta.$$

$$\begin{aligned} U_f(u, v) &= \frac{A \exp\left[j\frac{k}{2d}(u^2 + v^2)\right]}{j\lambda d} \frac{f}{d} \\ &\times \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} t_A(\xi, \eta) P\left(\frac{\xi f}{d}, \frac{\eta f}{d}\right) \exp\left[-j\frac{2\pi}{\lambda d}(u\xi + v\eta)\right] d\xi d\eta. \end{aligned}$$

(6-27)

Thus, up to a quadratic-phase factor, the focal-plane amplitude distribution is the Fourier transform of that portion of the input subtended by the projected lens aperture.

The result presented in (6-27) is essentially the same result obtained when the input was placed directly against the lens itself. However, an extra flexibility has been obtained in the present configuration; namely, the scale of the Fourier transform is under the control of the experimenter. By increasing d , the distance from the focal plane, the size of the transform is made larger, at least until the transparency is directly against the lens (i.e. $d=f$). By decreasing d , the scale of the transform is made smaller. This flexibility can be of utility in spatial filtering applications (see [Chapter 10](#)), where some potential adjustment of the size of the transform can be of considerable help.

6.2.4 Example of an Optical Fourier Transform

We illustrate with a typical example the type of two-dimensional Fourier analysis that can be achieved optically with great ease. [Figure 6.7](#) shows a transparent character 3, which is placed in front of a positive lens and illuminated by a plane wave, yielding in the back focal plane the intensity distribution shown in the right-hand part of the figure. Note in particular the high-frequency components introduced by the straight edges in the input.

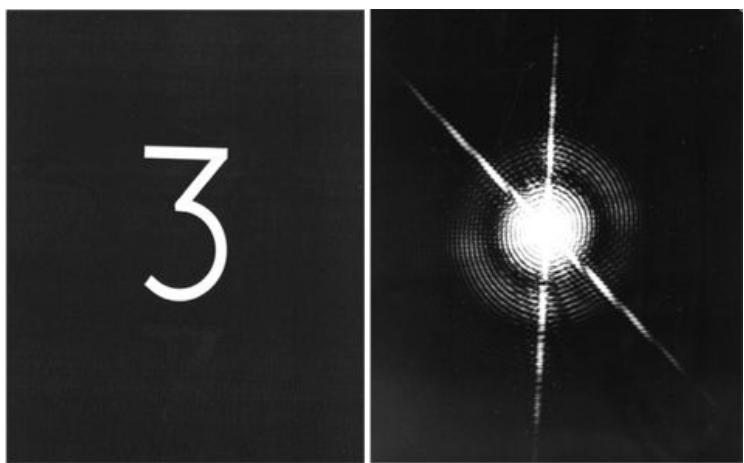


Figure 6.7

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 6.7 Optically obtained Fourier transform of the character 3.

6.3 Image Formation: Monochromatic Illumination

Certainly the most familiar property of lenses is their ability to form images. If an object is placed in front of a lens and illuminated, then under appropriate conditions there will appear across a second plane a distribution of light intensity that closely resembles the object. This distribution of intensity is called an *image* of the object. The image may be *real* in the sense that an actual distribution of intensity appears across a plane behind the lens, or it may be *virtual* in the sense that the light behind the lens appears to originate from an intensity distribution across a new plane in front of the lens.

For the present we consider image formation in only a limited context. First we restrict attention to a positive, aberration-free thin lens that forms a real image. Second, we consider only *monochromatic* illumination, a restriction implying that the imaging system is linear in complex field amplitude (see [Prob. 7-18](#)). Both of these restrictions will be removed in [Chapter 7](#), where the problem of image formation will be treated in a much more general fashion.

6.3.1 The Impulse Response of a Positive Lens

Referring to the geometry of [Fig. 6.8](#), suppose that a planar object is placed a distance z_1 in front of a positive lens and is illuminated by monochromatic light. We represent the complex field immediately behind the object by $U_o(\xi, \eta)$. At a distance z_2 behind the lens there appears a field distribution that we represent by $U_i(u, v)$. Our purpose is to find the conditions under which the field distribution U_i can reasonably be said to be an “image” of the object distribution U_o .

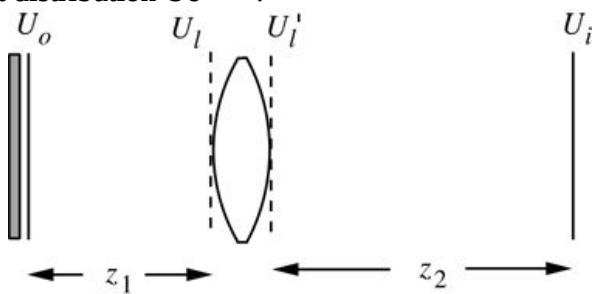


Figure 6.8

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 6.8 Geometry for image formation.

In view of the linearity of the wave propagation phenomenon, we can in all cases express the field U_i by the following superposition integral:

$$U_i(u, v) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(u, v; \xi, \eta) U_o(\xi, \eta) d\xi d\eta,$$

$$U_i(u, v) = \int_{-\infty}^{\infty} \int h(u, v; \xi, \eta) U_o(\xi, \eta) d\xi d\eta,$$

(6-28)

where $h(u, v; \xi, \eta)$ is the field amplitude produced at coordinates (u, v) by a unit-amplitude point source applied at object coordinates (ξ, η) . Thus the properties of the imaging system will be completely described if the impulse response h can be specified.

If the optical system is to produce high-quality images, then U_i must be as similar as possible to U_o . Equivalently, the impulse response should closely approximate a Dirac delta function,

$$\begin{aligned} h(u, v; \xi, \eta) &\approx K \delta(u - M\xi, v - M\eta), \\ h(u, v; \xi, \eta) &\approx K \delta(u - M\xi, v - M\eta), \end{aligned}$$

(6-29)

where K is a complex constant, and M represents the system magnification, a signed quantity that is negative in this case due to image inversion. We shall therefore specify as the “image plane” that plane where (6-29) is most closely approximated.

To find the impulse response h , let the object be a δ function (point source) at coordinates (ξ, η) . Then incident on the lens will appear a spherical wave diverging from the point (ξ, η) . The paraxial approximation to that wave is written

$$U_l(x, y) = 1j\lambda z_1 \exp[jk2z_1((x - \xi)^2 + (y - \eta)^2)].$$

$$U_l(x, y) = \frac{1}{j\lambda z_1} \exp \left\{ j \frac{k}{2z_1} [(x - \xi)^2 + (y - \eta)^2] \right\}.$$

(6-30)

After passage through the lens (focal length f), the field distribution becomes

$$U'_l(x, y) = U_l(x, y) P(x, y) \exp[-jk2f(x^2 + y^2)].$$

$$U'_l(x, y) = U_l(x, y) P(x, y) \exp \left[-j \frac{k}{2f} (x^2 + y^2) \right].$$

(6-31)

Finally, using the Fresnel diffraction equation (4-14) to account for propagation over distance z_2 , we have

$$h(u, v; \xi, \eta) = 1j\lambda z_2 \int_{-\infty}^{\infty} \int U'_l(x, y) \exp[jk2z_2((u - x)^2 + (v - y)^2)] dx dy$$

$$h(u, v; \xi, \eta) = \frac{1}{j\lambda z_2} \int_{-\infty}^{\infty} \int U_l'(x, y) \exp \left\{ j \frac{k}{2z_2} [(u - x)^2 + (v - y)^2] \right\} dx dy \quad (6-32)$$

where constant phase factors have been dropped. Combining (6-30), (6-31), and (6-32), and again neglecting a pure phase factor, yields the formidable result

$$h(u, v; \xi, \eta) = 1 \lambda 2 z_1 z_2 \exp jk 2 z_2 (u^2 + v^2) \exp jk 2 z_1 (\xi^2 + \eta^2) \times \int_{-\infty}^{\infty} \int P(x, y) \exp jk 2 z_1 + 1 z_2 - 1 f(x^2 + y^2) \times \exp -jk \xi z_1 + uz_2 x + \eta z_1 + vz_2 y dx dy.$$

$$\begin{aligned} h(u, v; \xi, \eta) &= \frac{1}{\lambda^2 z_1 z_2} \exp \left[j \frac{k}{2z_2} (u^2 + v^2) \right] \exp \left[j \frac{k}{2z_1} (\xi^2 + \eta^2) \right] \\ &\times \int_{-\infty}^{\infty} \int P(x, y) \exp \left[j \frac{k}{2} \left(\frac{1}{z_1} + \frac{1}{z_2} - \frac{1}{f} \right) (x^2 + y^2) \right] \\ &\times \exp \left\{ -jk \left[\left(\frac{\xi}{z_1} + \frac{u}{z_2} \right) x + \left(\frac{\eta}{z_1} + \frac{v}{z_2} \right) y \right] \right\} dx dy. \end{aligned} \quad (6-33)$$

Equations (6-28) and (6-33) now provide a formal solution specifying the relationship that exists between the object U_o and the image U_i . However, it is difficult to determine the conditions under which U_i can reasonably be called an image of U_o unless further simplifications are adopted.

6.3.2 Eliminating Quadratic-Phase Factors: The Lens Law

The most troublesome terms of the impulse response above are those containing quadratic-phase factors. Note that two of these terms are independent of the lens coordinates, namely

$$\exp jk 2 z_2 (u^2 + v^2) \text{ and } \exp jk 2 z_1 (\xi^2 + \eta^2),$$

$$\exp \left[j \frac{k}{2z_2} (u^2 + v^2) \right] \text{ and } \exp \left[j \frac{k}{2z_1} (\xi^2 + \eta^2) \right],$$

while one term depends on the lens coordinates (the variables of integration), namely

$$\exp jk 2 z_1 + 1 z_2 - 1 f(x^2 + y^2).$$

$$\exp \left[j \frac{k}{2} \left(\frac{1}{z_1} + \frac{1}{z_2} - \frac{1}{f} \right) (x^2 + y^2) \right].$$

We now consider a succession of approximations and restrictions that eliminate these factors. Beginning with the term involving the variables of integration (x, y) , note that the presence of a quadratic-phase factor in what otherwise would be a Fourier transform relationship will

generally have the effect of *broadening* the impulse response. For this reason we choose the distance z_2 to the image plane so that this term will identically vanish. This will be true if

$$1/z_1 + 1/z_2 - 1/f = 0.$$

$$\frac{1}{z_1} + \frac{1}{z_2} - \frac{1}{f} = 0.$$

(6-34)

Note that this relationship is precisely the classical *lens law* of geometrical optics, and must be satisfied for imaging to hold.

Consider next the quadratic-phase factor that depends only on image coordinates (u, v) . This term can be ignored under either of two conditions:

1. It is the intensity distribution in the image plane that is of interest, in which case the phase distribution associated with the image is of no consequence.
2. The image field distribution is of interest, but the image is measured on a spherical surface, centered at the point where the optical axis pierces the thin lens, and of radius z_2 (cf. [Section 4.2.5](#)).

Since it is usually the intensity of the image that is of interest, we will drop this quadratic-phase factor in the future.

Finally, consider the quadratic-phase factor in the object coordinates (ξ, η) . Note that this term depends on the variables over which the convolution operation [\(6-28\)](#) is carried out, and it has the potential to affect the result of that integration significantly. There are three different conditions under which this term can be neglected:

1. The object exists on the surface of a sphere of radius z_1 centered on the point where the optical axis pierces the thin lens.
2. The object is illuminated by a spherical wave that is converging towards the point where the optical axis pierces the lens.
3. The phase of the quadratic-phase factor changes by an amount that is only a small fraction of a radian within the region of the object that contributes significantly to the field at the particular image point (u, v) .

The first of these conditions rarely occurs in practice. The second can easily be made to occur by proper choice of the illumination, as illustrated in [Fig. 6.9](#). In this case the spherical wave illumination results in the Fourier transform of the object appearing in the pupil plane of the lens. The quadratic-phase factor of concern is exactly canceled by this converging spherical wave.

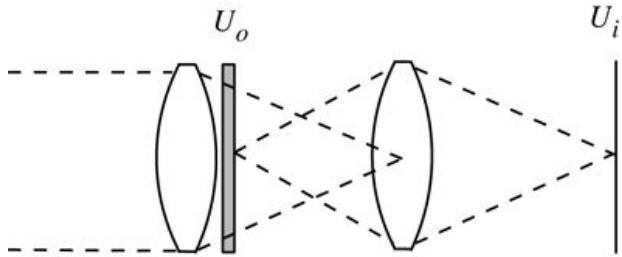


Figure 6.9

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 6.9 Converging illumination of the object.

The illustration shows two double convex lens with a vertical planar object U_o between them such that it is tangential to the curvature of lens on the left. The convergence of rays through the two lenses are as follows. Lens on the left: Parallel rays to the lens from the left pass through the lens and converge at the center of the lens on the right. Lens on the right: leftward rays from the lens converge at the center of the planar object and rightward rays from the lens converge at the center of vertical field distribution U_i , which is at the right extreme.

The third possibility for eliminating the effect of the quadratic-phase factor in object coordinates requires a more lengthy discussion. In an imaging geometry, the response of the system to an impulse at particular object coordinates should extend over only a small region of image space surrounding the exact image point corresponding to that particular object point. If this were not the case, the system would not be producing an accurate image of the object, or stated another way, it would have an unacceptably large image blur. By the same token, if we view the impulse response for a fixed image point as specifying the weighting function in object space that contributes to that image point, then only a small region on the object should contribute to any given image point². [Figure 6.10](#) illustrates this point of view. The gray patch on the left in this figure represents the area from which significant contributions arise for the particular image point

$$\frac{k}{2z_1}(\xi^2 + \eta^2)$$

on the right. If over this region the factor $\frac{k}{2z_1}(\xi^2 + \eta^2)$ changes by an amount that is only a small fraction of a radian, then the quadratic-phase factor in the object plane can be replaced by a single phase that depends on which image point (u, v) is of interest but does not depend on the object coordinates (ξ, η) (*i.e.* the phase is approximately constant over the region of interest in the object space). The replacement can be stated more precisely as

$$\exp[jk2z_1(\xi^2 + \eta^2)] \rightarrow \exp[jk2z_1u^2 + v^2M^2],$$

$$\exp\left[j\frac{k}{2z_1}(\xi^2 + \eta^2)\right] \rightarrow \exp\left[j\frac{k}{2z_1}\left(\frac{u^2 + v^2}{M^2}\right)\right],$$

(6-35)

where $M = -z_2/z_1$ is the magnification of the system. This new quadratic-phase factor in the image space can now be dropped provided that image intensity is the quantity of interest. See [Prob. 6-12](#) for consideration of the case when the complex amplitude, rather than the intensity, is desired in the image plane.

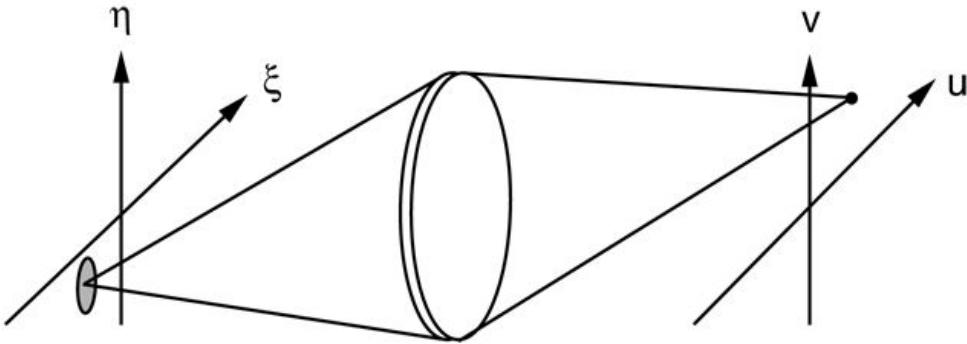


Figure 6.10

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 6.10 Region of object space contributing to the field at a particular image point.

The illustration shows two planes, u v and ξ η , where v and η are the vertical axes and u and ξ are the horizontal axes. Rays from the top right corner of the u v plane pass through the lens and converge at the bottom left corner of the ξ η plane at a point that is set in a gray circle.

[Tichenor and Goodman \[342\]](#) have examined this argument in detail and have found that the approximation stated above is valid provided the size of object is no greater than about $1/4^{1/4}$ the size of the lens aperture. For further consideration of this problem, see [Prob. 6-12](#).

The problematic quadratic-phase factors encountered in imaging with a single thin lens arise from a fundamental property of such systems. Imaging actually takes place between two spherical caps, rather than between two planes. If we construct a spherical cap in the object space, with its center on the center of the thin lens, and a spherical cap in the image space with its center on the center of the thin lens, then imaging takes place between these two spherical caps without any problematic quadratic-phase factors.

Returning to the case of plane-to-plane imaging, the end result of the arguments above is a simplified expression for the impulse response of the imaging system,

$$h(u, v; \xi, \eta) \approx 1/\lambda^2 z_1 z_2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(x, y) \times \exp[-jk(\xi z_1 + u z_2 x + \eta z_1 + v z_2 y)] dx dy.$$

$$h(u, v; \xi, \eta) \approx \frac{1}{\lambda^2 z_1 z_2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(x, y)$$

$$\times \exp\left[-jk\left(\left(\frac{\xi}{z_1} + \frac{u}{z_2}\right)x + \left(\frac{\eta}{z_1} + \frac{v}{z_2}\right)y\right)\right] dx dy.$$

(6-36)

Again recognizing that the magnification of the system is given by

$$M = -z_2 z_1,$$

$$M = -\frac{z_2}{z_1},$$

(6-37)

the minus sign being included to remove the effects of image inversion, we find a final simplified form for the impulse response,

$$h(u, v; \xi, \eta) \approx 1/\lambda z_1 z_2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(x, y) \times \exp[-j2\pi/\lambda z_2(u - M\xi)x + (v - M\eta)y] dx dy.$$

$$\begin{aligned} h(u, v; \xi, \eta) &\approx \frac{1}{\lambda^2 z_1 z_2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(x, y) \\ &\quad \times \exp \left\{ -j\frac{2\pi}{\lambda z_2} [(u - M\xi)x + (v - M\eta)y] \right\} dx dy. \end{aligned}$$

(6-38)

Thus, if the lens law is satisfied, the impulse response is seen to be given (up to an extra scaling factor $1/\lambda z_1$) by the Fraunhofer diffraction pattern of the lens aperture, centered on image coordinates $(u = M\xi, v = M\eta)$. The occurrence of a Fraunhofer diffraction formula should not be entirely surprising. By choosing z_2 to satisfy the lens law, we have chosen to examine the plane towards which the spherical wave leaving the lens is converging. From the results of [Prob. 4-18](#), we should expect the distribution of light about this point of convergence to be precisely the Fraunhofer diffraction pattern of the lens aperture that limits the extent of the spherical wave.

6.3.3 The Relation between Object and Image

Consider first the nature of the image predicted by geometrical optics. If the imaging system is perfect, then the image is simply an inverted and magnified (or demagnified) replica of the object. Thus according to geometrical optics, the ideal image field U_g and object field U_o would be related by

$$U_g(u, v) = |M| U_o(Mu, Mv).$$

$$U_g(u, v) = \frac{1}{|M|} U_o\left(\frac{u}{M}, \frac{v}{M}\right).$$

(6-39)

Indeed we can show that our wave optics solution reduces to this geometrical optics solution by using the common artifice of allowing the wavelength λ to approach zero, with the result that (see [Prob. 6-16](#))

$$h(u, v; \xi, \eta) \rightarrow 1/|M| \delta(\xi - Mu, \eta - Mv).$$

$$h(u, v; \xi, \eta) \rightarrow \frac{1}{|M|} \delta\left(\xi - \frac{u}{M}, \eta - \frac{v}{M}\right).$$

(6-40)

Substitution of this result in the general superposition equation (6-28) yields (6-39).

The predictions of geometrical optics do not include the effects of diffraction. A more complete understanding of the relation between object and image can be obtained only if such effects are included. Towards this end, we return to the expression (6-38) for the impulse response of the imaging system. As it currently stands, the impulse response is that of a linear *space-variant* system, so the object and image are related by a superposition integral but not by a convolution integral. This space-variant attribute is a direct result of the magnification and image inversion that occur in the imaging operation. To reduce the object-image relation to a convolution equation, we must normalize the object coordinates to remove inversion and magnification. Let the following normalized object-plane variables be introduced:

$$\tilde{\xi} = M\xi \quad \tilde{\eta} = M\eta$$

$$\tilde{\xi} = M\xi \tilde{\eta} = M\eta$$

in which case the impulse response of (6-38) reduces to

$$h^*(u, v; \tilde{\xi}, \tilde{\eta}) = 1/\lambda^2 z_1 z_2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(x, y) \times \exp[-j2\pi z_2(u - \tilde{\xi})x + (v - \tilde{\eta})y] dx dy,$$

$$\begin{aligned} \hat{h}(u, v; \tilde{\xi}, \tilde{\eta}) &= \frac{1}{\lambda^2 z_1 z_2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(x, y) \\ &\times \exp \left\{ -j \frac{2\pi}{\lambda z_2} [(u - \tilde{\xi})x + (v - \tilde{\eta})y] \right\} dx dy, \end{aligned} \quad (6-41)$$

which depends only on the differences of coordinates $(u - \tilde{\xi}, v - \tilde{\eta})$ and is therefore a space-invariant impulse response.

A final set of coordinate normalizations simplifies the results even further. Let

$$\tilde{x} = x/\lambda z_2 \quad \tilde{y} = y/\lambda z_2 \quad \tilde{h} = \frac{1}{|M|} \hat{h}.$$

$$\tilde{x} = \frac{x}{\lambda z_2} \quad \tilde{y} = \frac{y}{\lambda z_2} \quad \tilde{h} = \frac{1}{|M|} \hat{h}.$$

Then the object-image relationship becomes

$$\begin{aligned} U_i(u, v) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \tilde{h}(u - \tilde{\xi}, v - \tilde{\eta}) \left[\frac{1}{|M|} U_o \left(\frac{\tilde{\xi}}{|M|}, \frac{\tilde{\eta}}{|M|} \right) \right] d\tilde{\xi} d\tilde{\eta}, \\ U_i(u, v) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \tilde{h}(u - \tilde{\xi}, v - \tilde{\eta}) \left[\frac{1}{|M|} U_o \left(\frac{\tilde{\xi}}{|M|}, \frac{\tilde{\eta}}{|M|} \right) \right] d\tilde{\xi} d\tilde{\eta}, \end{aligned} \quad (6-42)$$

or, using Eq. (6-39),

$$U_i(u, v) = h^*(u, v) * U_g(u, v),$$

$$U_i(u, v) = \tilde{h}(u, v) * U_g(u, v),$$

(6-43)

where

$$\tilde{h}(u, v) = 1/\lambda^2 z_2^2 \int_{-\infty}^{\infty} \int P(x, y) \exp[-j2\pi/\lambda z_2(ux + vy)] dx dy = \int_{-\infty}^{\infty} \int P(\lambda z_2 x, \lambda z_2 y) \exp[-j2\pi/\lambda z_2(ux + vy)] dx dy$$

$$\begin{aligned} \tilde{h}(u, v) &= \frac{1}{\lambda^2 z_2^2} \int_{-\infty}^{\infty} \int P(x, y) \exp\left[-j\frac{2\pi}{\lambda z_2}(ux + vy)\right] dx dy \\ &= \int_{-\infty}^{\infty} \int P(\lambda z_2 \bar{x}, \lambda z_2 \bar{y}) \exp[-j2\pi(\bar{x}u + \bar{y}v)] d\bar{x} d\bar{y} \end{aligned}$$

(6-44)

There are two main conclusions from the analysis and discussion above:

1. The ideal image produced by a diffraction-limited optical system (i.e. a system that is free from aberrations) is a scaled and possibly inverted version of the object.
2. The effect of diffraction is to convolve that ideal image (i.e. (6-39)) with a function that is proportional to the Fraunhofer diffraction pattern of the lens pupil.

The smoothing operation associated with the convolution can strongly attenuate the fine details of the object, with a corresponding loss of image fidelity resulting. Similar effects occur in electrical systems when an input with high-frequency components passes through a filter with a limited frequency response. In the case of electrical systems, the loss of signal fidelity is most conveniently described in the frequency domain. The great utility of frequency-analysis concepts in the electrical case suggests that similar concepts might be usefully employed in the study of imaging systems. The application of filtering concepts to imaging systems is a subject of great importance and will be considered in detail in [Chapter 7](#).

6.4 Analysis of Complex Coherent Optical Systems

In the previous sections we have analyzed several different optical systems. These systems involved at most a single thin lens and at most propagation over two regions of free space. More complex optical systems can be analyzed by using the same methods applied above. However, the number of integrations grows as the number of free-space regions grows, and the complexity of the calculations increases as the number of lenses included grows. For these reasons, it is useful to find an approach to analyzing such systems that allows considerable complexity in the optical system of interest. We cover one such approach here.

6.4.1 The Ray Matrix Approach

Perhaps the simplest approach to analyzing complex optical systems is provided by use of the ABCD ray matrix approach of paraxial geometrical optics (see [Appendix B, Section B.3](#)), coupled with the result of [Section 4.2.6](#), and in particular (4-35), which in one dimension is written as

$$U_2(x) = 1/\lambda B \int_{-\infty}^{\infty} U_1(\xi) \exp[j\pi\lambda BA\xi^2 - 2\xi x + Dx^2] d\xi,$$

$$U_2(x) = \frac{1}{\sqrt{\lambda B}} \int_{-\infty}^{\infty} U_1(\xi) \exp \left[j\frac{\pi}{\lambda B} (A\xi^2 - 2\xi x + Dx^2) \right] d\xi,$$

(6-45)

and for a circularly-symmetric, non-astigmatic, two-dimensional system can be written as

$$U_2(x,y) = 1/\lambda B \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U_1(\xi, \eta) \exp[j\pi\lambda BA(\xi^2 + \eta^2) - 2(\xi x + \eta y) + D(x^2 + y^2)] d\xi d\eta.$$

$$\begin{aligned} U_2(x, y) &= \frac{1}{\lambda B} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U_1(\xi, \eta) \\ &\times \exp \left\{ j\frac{\pi}{\lambda B} [A(\xi^2 + \eta^2) - 2(\xi x + \eta y) + D(x^2 + y^2)] \right\} d\xi d\eta. \end{aligned}$$

(6-46)

In both of the above equations we have eliminated constant phase factors by appropriately redefining the phase reference.

6.4.2 Analysis of Two Optical Systems Using Ray Matrices

We illustrate the use of the ray matrix approach by analyzing two optical geometries that have not yet been treated. The first is fairly simple, consisting of two spherical lenses, each with the same focal length f , with a separation of f between them, as shown in [Fig. 6.11](#). The goal is to determine the relationship between the complex field across a plane S_1 just to the left of lens L_1 , and the complex field across a plane S_2 just to the right of the lens L_2 .

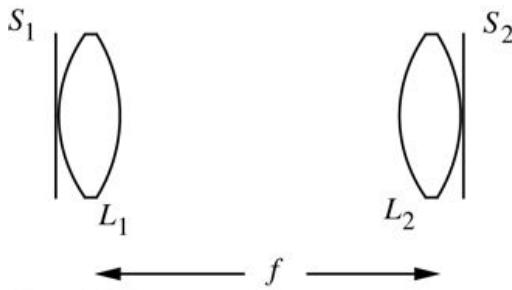


Figure 6.11

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 6.11 First problem analyzed.

The illustration shows two double convex lens, L 1 on the left and L 2 on the right, whose centers are separated by a horizontal distance of f . Plane S 1 is tangential to the left curvature of L1 and plane S2 is tangential to the right curvature of L 2.

To analyze this system, we need two ray matrices, one for passage through a lens with focal length f , and the second for free-space propagation over distance f . Assuming that the refractive index of the free-space region is unity, they are, respectively,

$M_f = 10 - 1/f$ passage through a thin lens of focal length f
 $M_z = 1/f$ free-space propagation over distance $z = f$

$$\begin{aligned} M_f &= \begin{bmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{bmatrix} && \text{passage through a thin lens of focal length } f \\ M_z &= \begin{bmatrix} 1 & f \\ 0 & 1 \end{bmatrix} && \text{free-space propagation over distance } z = f \end{aligned}$$

(6-47)

Multiplying matrices in the proper order yields

$$M_{\text{total}} = M_f M_z = f M_f = 0 f - 1/f 0.$$

$$M_{\text{total}} = M_f M_{z=f} M_f = \begin{bmatrix} 0 & f \\ -\frac{1}{f} & 0 \end{bmatrix}.$$

(6-48)

Thus $A=0$, $B=f$, $C=-1/f$ and $D=0$. It then follows that the field $U_2(x)$ is given by

$$U_2(x) = 1/\lambda f \int_{-\infty}^{\infty} U_1(\xi) \exp(-j2\pi f \xi x) d\xi.$$

$$U_2(x) = \frac{1}{\sqrt{\lambda f}} \int_{-\infty}^{\infty} U_1(\xi) \exp\left(-j\frac{2\pi}{\lambda f} \xi x\right) d\xi.$$

(6-49)

Thus the system of Fig. 6.11 performs a Fourier transform with the usual scaling factors and no premultiplying quadratic-phase factors.

The second geometry of interest is shown in Fig. 6.12. The illumination of the object in this case is a paraxial spherical wave, and the wave transmitted by the object is

$$U_1(\xi) = \exp j\pi\lambda(z_1-d)\xi^2 t_o(\xi),$$

$$U_1(\xi) = \exp \left[j\frac{\pi}{\lambda(z_1-d)}\xi^2 \right] t_o(\xi),$$

(6-50)

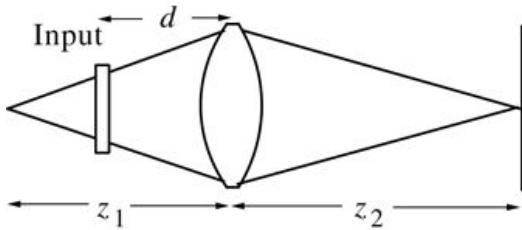


Figure 6.12

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 6.12 Second problem analyzed.

The illustration shows a biconvex lens with an input transparency to its left and a focal plane to its right. From a point on the left extreme, at a distance of z_1 from the center of the lens, rays diverge and move rightward through the input transparency to the lens; thereafter the rays converge to the center of the focal plane, located at a distance of z_2 from the center of the lens. The input transparency is at a distance of d from the center of the lens.

where $t_o(\xi)$ is the amplitude transmittance of the object, and we have assumed that the spherical wave illumination has unity intensity at the object. The lens is assumed to have focal length f , and we remove f from the answer by assuming the lens law holds for z_1 and z_2 , i.e. $f = z_1 z_2 / (z_1 + z_2)$. Propagation from the plane just behind the object to the output plane is described by the ray matrix

$$M_{\text{total}} = M_z = z_2 M_f = z_1 z_2 / (z_1 + z_2) M_z = d$$

$$M_{\text{total}} = M_{z=z_2} M_{f=z_1 z_2 / (z_1 + z_2)} M_{z=d}$$

(6-51)

$$= -z_2 / z_1 z_2 (1 - d/z_1) \quad -(z_1 + z_2) / (z_1 z_2) \quad 1 - d(z_1 + z_2) / (z_1 z_2).$$

$$= \begin{bmatrix} -z_2 / z_1 & z_2 (1 - d/z_1) & z_1 \\ -(z_1 + z_2) / (z_1 z_2) & 1 - d(z_1 + z_2) / (z_1 z_2) & z_2 \end{bmatrix}.$$

(6-52)

The ratios A/B and D/B are given by

$$AB = -z_1 z_2 d / (z_1 + z_2) z_2^2 (z_1 - d).$$

$$\begin{aligned}\frac{A}{B} &= -\frac{1}{z_1 - d} \\ \frac{D}{B} &= \frac{z_1 z_2 - d(z_1 + z_2)}{z_2^2 (z_1 - d)}.\end{aligned}$$

(6-53)

Since $\pi A/\lambda B$ is the coefficient of ξ^2 in (6-45), and since the quadratic-phase exponential of the illumination of the object in (6-50) has the negative of this coefficient, we conclude that the field $U_2(x)$ at the output is given by

$$U_2(x) = \exp j\pi\lambda z_1 z_2 - (z_1 + z_2)d z_2^2 (z_1 - d) x^2 \int_{-\infty}^{\infty} t_o(\xi) \exp \left[-j \frac{2\pi z_1}{\lambda z_2 (z_1 - d)} x \xi \right] d\xi$$

$$U_2(x) = \frac{\exp \left[j \frac{\pi z_1 z_2 - (z_1 + z_2)d}{\lambda z_2^2 (z_1 - d)} x^2 \right]}{\sqrt{\frac{\lambda z_2 (z_1 - d)}{z_1}}} \int_{-\infty}^{\infty} t_o(\xi) \exp \left[-j \frac{2\pi z_1}{\lambda z_2 (z_1 - d)} x \xi \right] d\xi$$

(6-54)

where the quadric-phase exponentials in x^2 have cancelled. Thus we obtain a scaled Fourier transform of the input object amplitude transmittance, with complicated scaling factors and a complicated quadratic-phase exponential in x .

The results of this analysis reveal some important general facts not explicitly evident in our earlier analyses. We emphasize these results because of their generality:

The Fourier transform plane need not be the focal plane of the lens performing the transform! Rather, the Fourier transform always appears in the plane where the source is imaged.

While it is not obvious without some further thought and analysis, our results show that the quadratic-phase factor preceding the Fourier transform operation is always the quadratic-phase factor that would result at the transform plane from a point source of light located on the optical axis in the plane of the input transparency.

The result presented in (6-54) can be shown to reduce to the results of the previous cases considered if z_1, z_2 , and d are properly chosen to represent those cases (see [Prob. 6-17](#)).

Problems - Chapter 6

1. 6-1. Show that the focal lengths of double-convex, plano-convex, and positive meniscus lenses are always positive, while the focal lengths of double-concave, plano-concave, and negative meniscus lenses are always negative.
2. 6-2. Consider a thin lens that is composed of a portion of a cylinder, as shown in [Fig. P6.2](#).
 1. Find a paraxial approximation to the phase transformation introduced by a lens of this form.
 2. What is the effect of such a lens on a plane wave traveling down the optical axis?

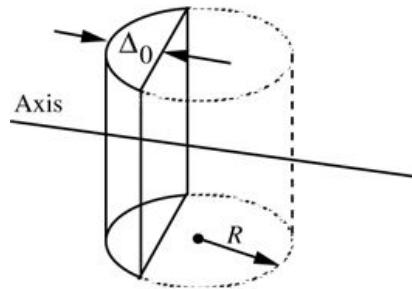


Figure P6.2
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

- Figure P6.2**
3. 6-3. A prism (illustrated in [Fig. P6.3](#)), which deflects the direction of propagation of a normally incident plane wave to angle θ with respect to the optical axis (the z axis) in the (y, z) plane, can be represented mathematically by an amplitude transmittance

$$tp(x,y)=\exp-j2\pi\lambda\sin\theta y.$$

$$t_p(x, y) = \exp \left[-j \frac{2\pi}{\lambda} (\sin\theta)y \right].$$

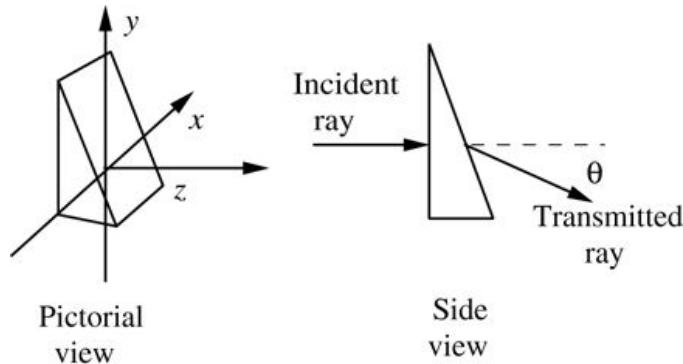


Figure P6.3

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure P6.3

The two views are of a wedge-shaped prism with a rectangular base and two rectangular sides, one perpendicular and the other slanted, such that together they form the other two vertical sides that are right triangular. In the pictorial view, the prism is set in a 3 dimensional coordinate with horizontal axis z , vertical axis y , and third axis x , such that the horizontal axis passes through the rectangular sides and the vertical axis passes through the base and the slanted side. The side view shows a right angled triangle with a rightward horizontal incident ray passing through the perpendicular side and exiting as the transmitted ray through the slanted side in a downward slope, making angle theta with a horizontal dotted line that is an extension of the incident ray.

1. Consider a thin transmitting structure with amplitude transmittance given by

$$tA(x,y) = \exp[-j\pi a^2 x^2 + (by+c)^2],$$

$$t_A(x, y) = \exp \left\{ -j\pi [a^2 x^2 + (by + c)^2] \right\},$$

with a, b, c all real and positive constants. It is claimed that this structure can be considered to consist of a sequence of one spherical lens, one cylindrical lens, and one prism, all placed in contact. Describe such a combination of thin elements that yields this transmittance, specifying the focal lengths of the lenses and the angle of deflection of the prism in terms of a, b, c , and the wavelength λ .

2. Can you think of a way to use two cylindrical lenses to achieve an amplitude transmittance

$$tA(x,y) = \exp(-j\pi dxy)$$

$$t_A(x, y) = \exp (-j\pi dxy)$$

where d is a constant? Explain your conclusion.

4. 6-4. Consider a lens that consists of the portion of a cone illustrated in [Fig. P6.4](#).

1. Show that a paraxial approximation to the phase transformation introduced by such a lens is (under the thin lens assumption)

$$tl(x,y) = \exp[jkn\Delta_o - (n-1)Ry - x^2/2f(y)]$$

$$t_l(x, y) = \exp \left\{ jk \left[n\Delta_o - \frac{(n-1)Ry}{h} - \frac{x^2}{2f(y)} \right] \right\}$$

where

$$f(y) = R(1-y/h)^{n-1}.$$

$$f(y) = \frac{R(1-y/h)^{n-1}}{n-1}.$$

2. What is the effect of such a lens on a plane wave traveling normal to the (x,y) plane?

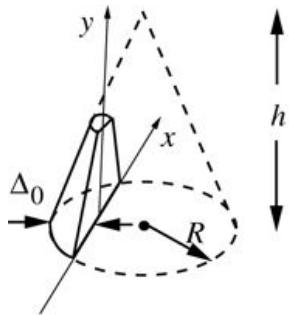


Figure P6.4
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure P6.4

The illustration shows a curved lens sectioned from a right cone of radius R and height h by a plane perpendicular to the circular base and a horizontal plane parallel to the base. At its thinnest point the lens measures Δ_0 , which is less than radius R . Axis x runs along the wider edge the plane side. Axis y is perpendicular at the center of wider edge of the plane side.

5. 6-5. An input function U_o , bounded by a circular aperture of diameter D and illuminated by a normally incident plane wave, is placed in the front focal plane of a circular positive lens of diameter L . The intensity distribution is measured across the back focal plane of the lens. Assuming $L > D$:

1. Find an expression for the maximum spatial frequency of the input for which the measured intensity accurately represents the squared modulus of the input's Fourier spectrum (free from the effects of vignetting).
2. What is the numerical value of that spatial frequency (in cycles/mm) when $L = 4$ cm, $D = 2$ cm, $f = 50$ cm, and $\lambda = 6 \times 10^{-7}$ meters?
3. Above what frequency does the measured spectrum vanish, in spite of the fact that the input may have nonzero Fourier components at such frequencies?

6. 6-6. An array of one-dimensional input functions can be represented by $U_o(\xi, \eta_k)$, where $\eta_1, \eta_2, \dots, \eta_N$ are N fixed η coordinates in the input plane. It is desired to perform a one-dimensional Fourier transform of all N functions in the ξ direction, yielding an array of transforms

$$G_o(f_X, \eta_k) = \int_{-\infty}^{\infty} U_o(\xi, \eta_k) \exp(-j2\pi f_X \xi) d\xi, k=1, 2, \dots, N.$$

$$G_o(f_X, \eta_k) = \int_{-\infty}^{\infty} U_o(\xi, \eta_k) \exp(-j2\pi f_X \xi) d\xi, \quad k = 1, 2, \dots, N.$$

Neglecting the finite extent of the lens and object apertures, use the Fourier transforming and imaging properties of lenses derived in this chapter to show how this can be done with

1. two cylindrical lenses of different focal lengths.
2. a cylindrical and a spherical lens of the same focal length.

Simplification: You need only display $|G_o|^2$, so phase factors may be dropped.

7. 6-7. A normally incident, unit-amplitude, monochromatic plane wave illuminates a converging lens of 5 cm diameter and 2 meters focal length (see Fig. P6.7). One meter behind the lens and centered on the lens axis is placed an object with amplitude transmittance

$$t_A(\xi, \eta) = 121 + \cos(2\pi f_o \xi) \operatorname{rect}\left(\frac{\xi}{L}\right) \operatorname{rect}\left(\frac{\eta}{L}\right).$$

$$t_A(\xi, \eta) = \frac{1}{2}[1 + \cos(2\pi f_o \xi)] \operatorname{rect}\left(\frac{\xi}{L}\right) \operatorname{rect}\left(\frac{\eta}{L}\right).$$

Assuming $L=1$ cm, $\lambda=0.633 \mu$ m, and $f_o=10$ cycles/mm, sketch the intensity distribution across the u axis of the focal plane, labeling the numerical values of the distance between diffracted components and the width (between first zeros) of

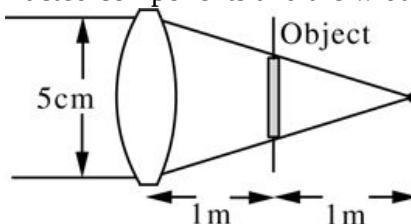


Figure P6.7
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

the individual components.

Figure P6.7

8. 6-8. In Fig. P6.8, a monochromatic point source is placed a fixed distance z_1 to the left of a positive lens (focal length f), and a transparent object is placed a variable distance d to the left of the lens. The distance z_1 is greater than f . The Fourier transform and the image of the object appear to the right of the lens.

- How large should the distance d be (in terms of z_1 and f) to ensure that the *Fourier plane* and the *object* are equidistant from the lens?
- When the object has the distance found in part (a) above, how far to the right of the lens is its image and what is the magnification of that image?

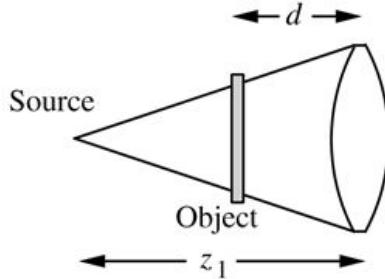


Figure P6.8
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure P6.8

9. 6-9. A unit-amplitude, normally incident, monochromatic plane wave illuminates an object of maximum linear dimension D , situated immediately in front of a larger positive lens of focal length f (see Fig. P6.9). Due to a positioning error, the intensity distribution is measured across a plane at a distance $f - \Delta$ behind the lens. How small must Δ be if the measured intensity distribution is to accurately represent the Fraunhofer diffraction

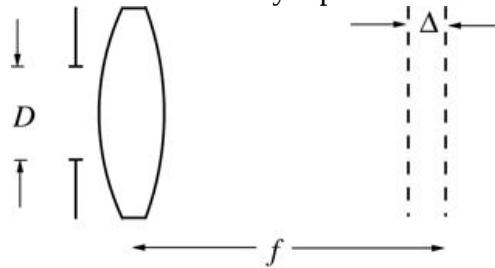


Figure P6.9
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

pattern of the object?

Figure P6.9

10. 6-10. Consider the optical system shown in Fig. P6.10. The object on the left is illuminated by a normally incident plane wave. Lens L₁ is a *negative* lens with focal length $-f$, and lens L₂ is a *positive* lens with focal length f . The two lenses are spaced by distance f . Lens L₁ is a distance $2f$ to the right of the object. Use the simplest possible reasoning to predict the distances d and z_2 , respectively, to the Fourier plane and the

image plane to the right or left of lens L_2 L_2 (specify right or left in the answers).

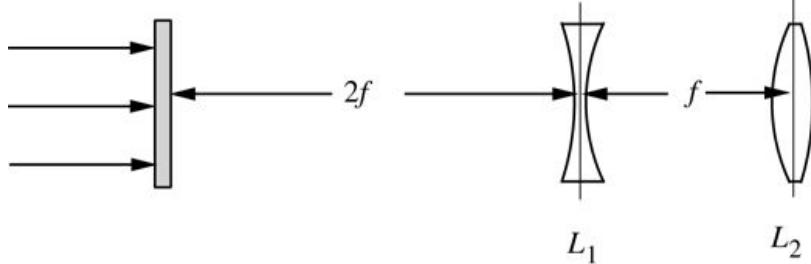


Figure P6.10

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure P6.10

11. 6-11. In the optical system shown in [Fig. P6.11](#), specify the locations of all Fourier and image planes to the left and right of the lens. The lens shown is positive and has focal length f . The illumination of the object is a converging spherical wave, as indicated.

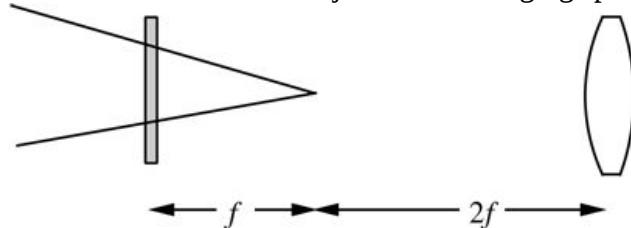


Figure P6.11

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure P6.11

12. 6-12. With reference to [Eq. \(6-35\)](#):

- At what radius r_0 in the object plane has the phase of $\exp[jk2z_1(\xi^2 + \eta^2)]$

changed by 1 radian from its value at the origin?

- Assuming a circular pupil function of radius R , what is the radius (in the object plane) to the first zero of the impulse response h , assuming that the observation point in the image space is the origin?

- From the results obtained so far, what relation between R , λ , and z_1 will allow

the quadratic-phase exponential $\exp[jk2z_1(\xi^2 + \eta^2)]$ to be replaced by a single complex number, assuming observation near the lens axis?

- With reference to [Section 6.3.2](#), in some cases the complex amplitude of the image field may be desired, rather than the intensity. This will be true if, for example, a transmitting structure is placed in the image plane and light then propagates into a second portion of the optical system. Show that, subject to the assumption that the weighting function in the object

space for any image point is confined to a very small region, the complex amplitude of the image field is given by

$$U_i(u, v) = \exp[j\pi|M|\lambda f(u^2 + v^2)(h(u, v)^* U_g(u, v))],$$

$$U_i(u, v) = \exp\left[j\frac{\pi}{|M|\lambda f}(u^2 + v^2)\right] (h(u, v)^* U_g(u, v)),$$

where

$$h(u, v) = 1/\lambda z_2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(x, y) \exp\left[-j\frac{2\pi}{\lambda z_2}(ux + vy)\right] dx dy,$$

$$h(u, v) = \frac{1}{\lambda^2 z_2^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(x, y) \exp\left[-j\frac{2\pi}{\lambda z_2}(ux + vy)\right] dx dy,$$

and $U_g(u, v)$ is the geometrical optics prediction of the image.

14. 6-14. A diffracting structure has a circularly symmetric amplitude transmittance function given by

$$t_A(r) = 12 + 12 \cos(\gamma r^2) \operatorname{circ}\left(\frac{r}{R}\right).$$

1. In what way does this screen act like a lens?
2. Give an expression for the focal length of the screen.
3. What characteristics might seriously limit the use of this screen as an imaging device for polychromatic objects?

15. 6-15. A certain diffracting screen with an amplitude transmittance

$$t_A(r) = 12 + 12 \operatorname{sgn}(\cos(\gamma r^2)) \operatorname{circ}\left(\frac{r}{R}\right)$$

$$t_A(r) = \left[\frac{1}{2} + \frac{1}{2} \operatorname{sgn}(\cos(\gamma r^2)) \right] \operatorname{circ}\left(\frac{r}{R}\right)$$

is normally illuminated by a unit-amplitude, monochromatic plane wave. Show that the screen acts as a lens with multiple focal lengths. Specify the values of these focal lengths and the relative amounts of optical power brought to focus in the corresponding focal planes. (A diffracting structure such as this is known as a *Fresnel zone plate*. Hint: The square wave shown in [Fig P6.15](#) can be represented by the Fourier series

$$f(x) = \sum_{n=-\infty}^{\infty} \sin(\pi n/2) \pi n \exp(j2\pi nx/X).$$

$$f(x) = \sum_{n=-\infty}^{\infty} \left[\frac{\sin(\pi n/2)}{\pi n} \right] \exp\left(j\frac{2\pi nx}{X}\right).$$

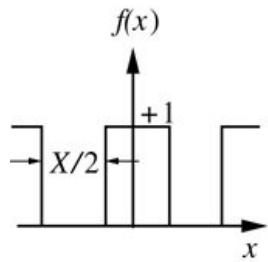


Figure P6.15
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure P6.15

The graph shows horizontal axis x with vertical axis $f(x)$ at its center. The wave is a rectangle of height +1 on the vertical axis, which it overlaps. On either side of the rectangle is a perpendicular of height +1 at a distance of $X/2$ from the rectangle. A horizontal line beginning at the top end of each perpendicular extends away from the vertical axis.

16. 6-16. Show that in the limit $\lambda \rightarrow 0$ $\lambda \rightarrow 0$, (6-38) approaches the impulse response shown in Eq. (6-40).
17. 6-17. Find the form of the general result of (6-54) under the following limiting conditions:
 1. $z_1 \rightarrow \infty$ $z_1 \rightarrow \infty$ and $d \rightarrow 0$ $d \rightarrow 0$.
 2. $z_1 \rightarrow \infty$ $z_1 \rightarrow \infty$ and $d \rightarrow f$ $d \rightarrow f$.
 3. $z_1 \rightarrow \infty$ $z_1 \rightarrow \infty$, general distance d d .
18. 6-18. Consider the simple optical system shown in Fig. P6.18.

1. Write the ray-matrix sequence that describes the successive propagation between planes and through lenses for this system.
2. Reduce this ray-matrix sequence to a simple scaling operator.

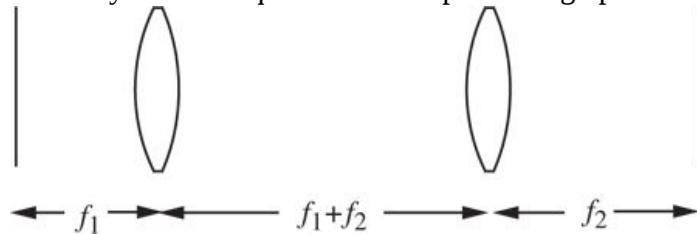


Figure P6.18

19. 6-19. Show that any perfect imaging system using a thin lens in free space, with object at distance z_1 z_1 from the lens and image at distance z_2 z_2 from the lens, must have a paraxial ray-transfer matrix of the form

$$M = M_0 - 1/f_1/M$$

$$M = \begin{bmatrix} M & 0 \\ -1/f & 1/M \end{bmatrix}$$

where $M = z_2 / z_1$ is the (signed) magnification of the system and f is the (signed) focal length of the lens.

7 Frequency Analysis of Optical Imaging Systems

Considering the long and rich history of optics, the tools of frequency analysis and linear systems theory have played important roles for only a relatively short period of time. Nevertheless, in this short time these tools have been so widely and successfully used that they now occupy a fundamental place in the theory of imaging systems.

A realization of the utility of Fourier methods in the analysis of optical systems arose rather spontaneously in the late 1930's when a number of workers began to advocate the use of sinusoidal test patterns for system evaluation. Much of the initial stimulus was provided by a French scientist, P.M. Duffieux, whose work culminated in the publication of a book, in 1946, on the use of Fourier methods in optics [97]. This book has recently been translated into English [98]. In the United States, much of the interest in these topics was stimulated by an electrical engineer, Otto Schade, who very successfully employed methods of linear systems theory in the analysis and improvement of television camera lenses [308]. In the United Kingdom, H.H. Hopkins led the way in the use of transfer function methods for the assessment of the quality of optical imaging systems, and was responsible for many of the first calculations of transfer functions in the presence of common aberrations [173]. However, it must be said that the foundations of Fourier optics were laid considerably earlier, particularly in the works of Ernst Abbe (1840–1905) and Lord Rayleigh (1842–1919).

In this chapter we shall consider the role of Fourier analysis in the theory of coherent and incoherent imaging. While historically the case of incoherent imaging has been the more important one, nonetheless the case of coherent imaging has always been important in microscopy, and it gained much additional importance with the advent of the laser. For example, the field of holography is predominantly concerned with coherent imaging.

For additional discussions of various aspects of the subject matter to follow, the reader may wish to consult any of the following references: [269], [117], [229], [84], [373].

7.1 Generalized Treatment of Imaging Systems

In the preceding chapter, the imaging properties of a single thin positive lens were studied for the case of monochromatic illumination. In the material to follow, we shall first broaden our discussion beyond a single thin positive lens, finding results applicable to more general systems of lenses, and then remove the restriction to monochromatic light, obtaining results for “quasi-monochromatic” light, both spatially coherent and spatially incoherent. To broaden the perspective, it will be necessary to draw upon some results from the theory of geometrical optics. The necessary concepts are all introduced in [Appendix B](#).

7.1.1 A Generalized Model

Suppose that an imaging system of interest is composed, not of a single thin lens, but perhaps of several lenses, some positive, some negative, with various distances between them. The lenses need not be thin in the sense defined earlier. We shall assume, however, that the system ultimately produces a *real* image in space; this is not a serious restriction, for if the system produces a virtual image, to view that image it must be converted to a real image, perhaps by the lens of the eye.

To specify the properties of the lens system, we adopt the point of view that all imaging elements may be lumped into a single “black box,” and that the significant properties of the system can be completely described by specifying only the *terminal properties* of the aggregate. Referring to [Fig. 7.1](#), the “terminals” of this black box consist of the planes containing the entrance and exit pupils (see [Appendix B](#) for a discussion of these planes).¹ It is assumed that the passage of light between the entrance pupil and the exit pupil is adequately described by geometrical optics.

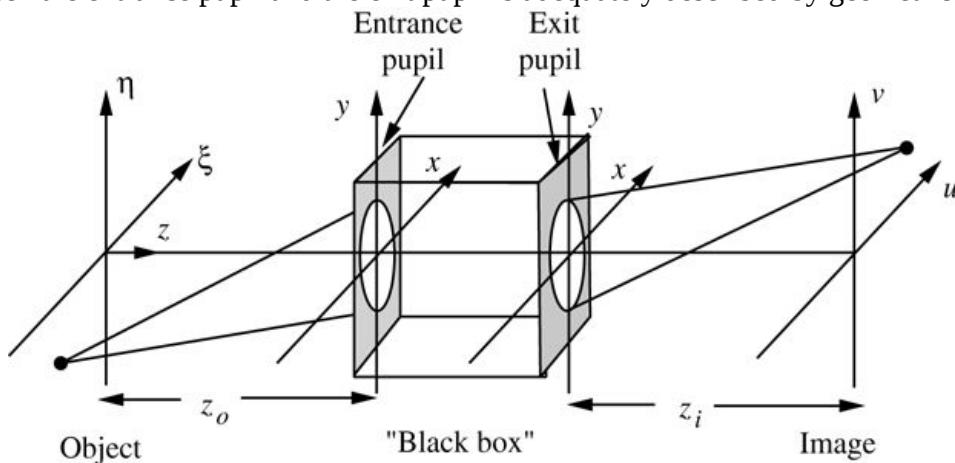


Figure 7.1

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 7.1 Generalized model of an imaging system.

The illustration shows a cube-shaped black box at the center whose left and right vertical sides are in the x y coordinate planes where y is the vertical axis and x is the third axis. Each of these two sides has a circular pupil at the center, the left one being the entrance pupil and the right one being the exit pupil. On the right side, at a distance of z subscript i from the exit pupil, is the image plane

$u v$, where v is the vertical axis and u the third axis. On the left side, at a distance of z subscript o from the entrance pupil, is the object plane $x_i \eta$, where η is the vertical axis and x_i the third axis. The four coordinate planes share a single horizontal axis z , which passes through the centers of the two pupils. Rays from a point in the bottom left corner of the object plane enter the entrance pupil and exit the exit pupil on the other side to converge at a point in the top right corner of the image plane.

The entrance and exit pupils are in fact images of the same limiting aperture within the system. As a consequence there are several different ways to visualize the origin of the spatial limitation of the wavefront that ultimately gives rise to diffraction. It can be viewed as being caused by the physical limiting aperture internal to the system (which is the true physical source of the limitation). Equivalently it can be viewed as arising from the entrance pupil or from the exit pupil of the system.

Throughout this chapter, we shall use the symbol z_o to represent the distance of the plane of the entrance pupil from the object plane, and the symbol z_i to represent the distance of the plane of the exit pupil from the image plane.² The distance z_i is then the distance that will appear in the diffraction equations that represent the effect of diffraction by the exit pupil on the point-spread function of the optical system. We shall refer either to the exit pupil or simply to the “pupil” of the system when discussing these effects.

An imaging system is said to be *diffraction-limited* if a diverging spherical wave, emanating from a point-source object, is converted by the system into a new wave, again perfectly spherical, that converges towards an ideal point in the image plane, where the transverse location of that ideal image point is related to the transverse location of the original object point through a simple scaling factor (the magnification), a factor that must be the same for all points in the image field of interest if the system is to be ideal. Thus the terminal property of a diffraction-limited imaging system is that a diverging spherical wave incident on the entrance pupil is converted by the system into a converging spherical wave at the exit pupil. For any real imaging system, this property will be satisfied, at best, over only finite regions of the object and image planes. If the object of interest is confined to the region for which this property holds, then the system may be regarded as being diffraction-limited.

If in the presence of a point-source object, the wavefront leaving the exit pupil departs significantly from ideal spherical shape, then the imaging system is said to have *aberrations*. Aberrations will be considered in [Section 7-4](#), where it is shown that they lead to defects in the spatial-frequency response of the imaging system.

7.1.2 Effects of Diffraction on the Image

Since geometrical optics adequately describes the passage of light between the entrance and exit pupils of a system, diffraction effects play a role only during passage of light from the object to the entrance pupil, or alternatively and equivalently, from the exit pupil to the image. It is, in fact, possible to associate *all* diffraction limitations with *either* of these two pupils. The two points of view that regard image resolution as being limited by (1) the finite entrance pupil seen from the object space or (2) the finite exit pupil seen from the image space are entirely equivalent, due to the fact that these two pupils are images of each other.

The view that diffraction effects result from the entrance pupil was first espoused by Ernst Abbe in 1873 [1] in studies of coherent imagery with a microscope. According to the Abbe theory, only a certain portion of the diffracted components generated by a complicated object are intercepted by this finite pupil. The components not intercepted are precisely those generated by

the high-frequency components of the object amplitude transmittance. This viewpoint is illustrated in Fig. 7.2 for the case of an object that is a grating with several orders and an imaging system composed of a single positive lens.

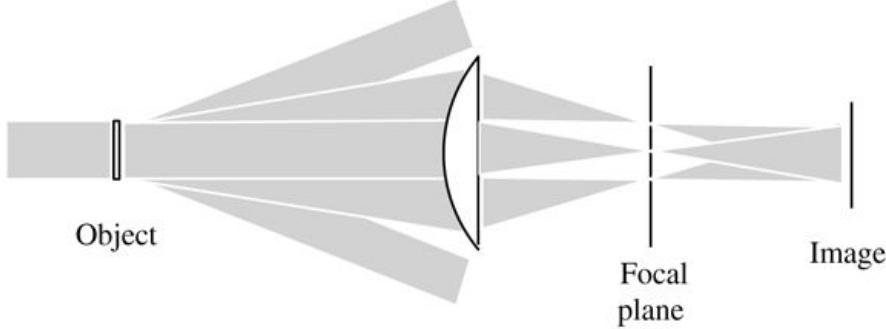


Figure 7.2

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 7.2 The Abbe theory of image formation.

The illustration shows a vertical object plane on the left extreme, rays travel from it to the curvature side of a plano-convex lens to the right. Some rays are horizontal while others are upward sloping reaching the top section of the lens and yet others are downward sloping reaching the bottom section of the lens. The rays emerge on the other side, the plane side, as three beams of rays converging on the focal plane and exiting it to diverge and fuse into an image on the image plane located still further to the right.

A view equivalent to regarding diffraction effects as resulting from the exit pupil was presented by Lord Rayleigh in 1896 [295]. This is the viewpoint that was used in Section 6.3, and we shall adopt it again here.

Again the image amplitude³ is represented by a superposition integral

$$U_i(u, v) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(u, v; \xi, \eta) U_o(\xi, \eta) d\xi d\eta,$$

$$U_i(u, v) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(u, v; \xi, \eta) U_o(\xi, \eta) d\xi d\eta,$$

(7-1)

where h is the amplitude at image coordinates (u, v) in response to a delta-function object at (ξ, η) , and U_o is the amplitude distribution transmitted by the object. In the absence of aberrations, the response h arises from a spherical wave (of limited extent) converging from the exit pupil towards the ideal image point $(u = M\xi, v = M\eta)$. We allow the magnification to be either negative or positive, according to whether the image is inverted or not.

The light amplitude about the ideal image point is simply the Fraunhofer diffraction pattern of the exit pupil,⁴ centered on image coordinates $(u = M\xi, v = M\eta)$. Thus

$$h(u, v; \xi, \eta) = 1/\lambda^2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(x, y) \exp[-j2\pi/\lambda z_i(u - M\xi)x + (v - M\eta)y] dx dy,$$

$$h(u, v; \xi, \eta) = \frac{1}{\lambda^2 z_i^2} \int_{-\infty}^{\infty} \int P(x, y) \exp \left\{ -j \frac{2\pi}{\lambda z_i} [(u - M\xi)x + (v - M\eta)y] \right\} dx dy,$$

(7-2)

where the pupil function P is unity inside and zero outside the exit pupil aperture, z_i is the distance from the exit pupil to the image plane, and (x, y) are coordinates in the plane of the exit pupil. In what follows we neglect the pure phase factors preceding the integral, as justified in [Section 6.3.2](#).

In order to achieve space invariance in the imaging operation, it is necessary to remove the effects of magnification and image inversion from the equations. This can be done by defining *reduced coordinates* in the object space⁵ according to

$$\tilde{\xi} = M\xi, \tilde{\eta} = M\eta,$$

$$\tilde{\xi} = M\xi\tilde{\eta} = M\eta,$$

in which case the amplitude point-spread function becomes

$$h(u - \tilde{\xi}, v - \tilde{\eta}) = 1/\lambda^2 z_i^2 \int_{-\infty}^{\infty} \int P(x, y) \exp \left\{ -j \frac{2\pi}{\lambda z_i} [(u - \tilde{\xi})x + (v - \tilde{\eta})y] \right\} dx dy.$$

$$h(u - \tilde{\xi}, v - \tilde{\eta}) = \frac{1}{\lambda^2 z_i^2} \int_{-\infty}^{\infty} \int P(x, y) \exp \left\{ -j \frac{2\pi}{\lambda z_i} [(u - \tilde{\xi})x + (v - \tilde{\eta})y] \right\} dx dy.$$

At this point it is convenient to define the *ideal image*, or the geometrical-optics prediction of the image for a perfect imaging system as

$$U_g(\tilde{\xi}, \tilde{\eta}) = |M| U_o(\tilde{\xi}, \tilde{\eta}),$$

$$U_g(\tilde{\xi}, \tilde{\eta}) = \frac{1}{|M|} U_o\left(\frac{\tilde{\xi}}{M}, \frac{\tilde{\eta}}{M}\right),$$

(7-3)

yielding a convolution equation for the image,

$$U_i(u, v) = \int_{-\infty}^{\infty} \int h(u - \tilde{\xi}, v - \tilde{\eta}) U_g(\tilde{\xi}, \tilde{\eta}) d\tilde{\xi} d\tilde{\eta},$$

$$U_i(u, v) = \int_{-\infty}^{\infty} \int h(u - \tilde{\xi}, v - \tilde{\eta}) U_g(\tilde{\xi}, \tilde{\eta}) d\tilde{\xi} d\tilde{\eta},$$

(7-4)

where

$$h(u, v) = 1/\lambda^2 z_i^2 \int_{-\infty}^{\infty} \int P(x, y) \exp \left\{ -j \frac{2\pi}{\lambda z_i} [ux + vy] \right\} dx dy.$$

$$h(u, v) = \frac{1}{\lambda^2 z_i^2} \int_{-\infty}^{\infty} \int P(x, y) \exp \left\{ -j \frac{2\pi}{\lambda z_i} (ux + vy) \right\} dx dy.$$

(7-5)

Thus in this general case, for a diffraction-limited system we can regard the image as being a convolution of the image predicted by geometrical optics with an impulse response that is proportional to the Fraunhofer diffraction pattern of the exit pupil.

7.1.3 Polychromatic Illumination: The Coherent and Incoherent Cases

The assumption of strictly monochromatic illumination has been present in all our discussions of imaging systems up to this point. This assumption is overly restrictive, for the illumination generated by real optical sources, including lasers, is never perfectly monochromatic. The statistical nature of the time variations of illumination amplitude and phase can, in fact, influence the behavior of an imaging system in profound ways. We therefore digress temporarily to consider the very important effects of polychromicity.

To treat this subject in a completely satisfactory way, it would be necessary to take a rather long detour through the *theory of partial coherence*. However, for our purposes such a detailed detour would not be practical. We therefore treat the subject from two points of view, one entirely heuristic, and the second more rigorous but not entirely complete. The reader interested in a more complete treatment may wish to consult [376], [34], or [135].

In the case of monochromatic illumination it was convenient to represent the complex amplitude of the field by a complex phasor U that was a function of space coordinates. When the illumination is polychromatic but narrowband, i.e. occupying a bandwidth that is small compared with its center frequency, this approach can be generalized by representing the field by a *time-varying* phasor that depends on both time and space coordinates. For the narrowband case, the amplitude and phase of the time-varying phasor are readily identified with the envelope and phase of the real optical wave.

Consider the nature of the light that is transmitted by or reflected from an object illuminated by a polychromatic wave. Since the time variations of the phasor amplitude are statistical in nature, only statistical concepts can provide a satisfactory description of the field. As we have seen previously, each object point generates an amplitude impulse response in the image plane. If the amplitude and phase of the light at a particular object point vary randomly with time, then the overall amplitude and phase of the amplitude impulse response will vary in a corresponding fashion. Thus the statistical relationships between the phasor amplitudes at the various points on the object will influence the statistical relationships between the corresponding impulse responses in the image plane. These statistical relationships will greatly affect the result of the time-averaging operation that yields the final image intensity distribution.

We shall consider only two types of illumination here. First, we consider object illumination with the particular property that the phasor amplitudes of the field at all object points vary *in unison*. Thus while any two object points may have different *relative* phases, their absolute phases are varying with time in a perfectly correlated way. Such illumination is called *spatially coherent*. Second, we consider object illumination with the opposite property that the phasor amplitudes at all points on the object are varying in totally uncorrelated fashions. Such illumination is called *spatially incoherent*. (In the future we shall refer to these types of illumination as simply *coherent*

or *incoherent*.) Coherent illumination is obtained whenever light appears to originate from a single point.⁶ The most common example of a source of such light is a laser, although more conventional sources (e.g. zirconium arc lamps) can yield coherent light, albeit of weaker brightness than a laser, if their output is first passed through a pinhole. Incoherent light is obtained from diffuse or extended sources, for example gas discharges and the sun.

When the object illumination is coherent, the various impulse responses in the image plane vary in unison, and therefore must be added on a complex amplitude basis. *Thus a coherent imaging system is linear in complex amplitude.* The results of the monochromatic analysis can therefore be applied directly to such systems, with the understanding that the complex amplitude U is now a time-invariant phasor that depends on the *relative* phases of the light.

When the object illumination is incoherent, the various impulse responses in the image plane vary in uncorrelated fashions. They must therefore be added on a power or intensity basis. Since the intensity of any given impulse response is proportional to the intensity of the point source that gave rise to it, it follows that *an incoherent imaging system is linear in intensity, and the impulse response of such a system is the squared magnitude of the amplitude impulse response.*

The preceding arguments have been entirely heuristic, and in fact have certain assumptions and approximations hidden in them. We therefore turn to a more rigorous examination of the problem. To begin, note that in the monochromatic case we obtain the phasor representation of the field by suppressing the positive-frequency component of the sinusoidal field, and doubling the remaining negative frequency component. To generalize this concept to a polychromatic wave $u(P,t)$, we suppress all positive-frequency components of its Fourier spectrum, and double its negative-frequency components, yielding a new (complex) function $u_-(P,t)$. If we further write

$$u_-(P,t) = U(P,t) \exp(-j2\pi v^- t),$$

$$u_-(P,t) = U(P,t) \exp(-j2\pi \bar{\nu} t),$$

where $v^- \bar{\nu}$ represents the mean or center frequency of the optical wave, then the complex function $U(P,t)$ may be regarded as the time-varying phasor representation of $u(P,t)$.

Under the narrowband condition assumed above, the amplitude impulse response does not change appreciably for the various frequencies contained within the optical spectrum. Therefore it is possible to express the time-varying phasor representation of the image in terms of the convolution of a wavelength-independent impulse response with the time varying phasor representation of the object (in reduced object coordinates),

$$\begin{aligned} U_i(u,v;t) &= \iint_{-\infty}^{\infty} h(u-\xi, v-\eta) U_g(\xi, \eta; t-\tau) d\xi d\eta \\ U_i(u,v;t) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(u-\xi, v-\eta) U_g(\xi, \eta; t-\tau) d\xi d\eta \end{aligned} \quad (7-6)$$

where τ is a time delay associated with propagation from $(\tilde{\xi}, \tilde{\eta})$ to (u, v) (u, v) (note that in general, τ is a function of the coordinates involved).

To calculate the image intensity, we must *time average* the instantaneous intensity represented by $|U_i(u, v; t)|^2$, due to the fact that the detector integration time is usually extremely long compared with the reciprocal of the optical bandwidth, even for narrowband optical sources. In what follows, the angle brackets represent the infinite time averaging operation. Thus the image intensity is given by $I_i(u, v) = |U_i(u, v; t)|^2$

$I_i(u, v) = \langle |U_i(u, v; t)|^2 \rangle$, or, after substitution of (7-6) and interchanging orders of averaging and integration,

$$I_i(u, v) = \iint_{-\infty}^{\infty} d\xi_1 d\eta_1 \iint_{-\infty}^{\infty} d\xi_2 d\eta_2 h(u - \xi_1, v - \eta_1) h^*(u - \xi_2, v - \eta_2),$$

$$I_i(u, v) = \int_{-\infty}^{\infty} d\tilde{\xi}_1 d\tilde{\eta}_1 \int_{-\infty}^{\infty} d\tilde{\xi}_2 d\tilde{\eta}_2 h(u - \tilde{\xi}_1, v - \tilde{\eta}_1) h^*(u - \tilde{\xi}_2, v - \tilde{\eta}_2) \\ \times \langle U_g(\tilde{\xi}_1, \tilde{\eta}_1; t - \tau_1) U_g^*(\tilde{\xi}_2, \tilde{\eta}_2; t - \tau_2) \rangle.$$

(7-7)

Now for a fixed image point, the impulse response h is nonzero over only a small region about the ideal image point. Therefore the integrand is nonzero only for points $(\tilde{\xi}_1, \tilde{\eta}_1)$ and $(\tilde{\xi}_2, \tilde{\eta}_2)$ that are very close together⁷. Hence we assume that the difference between the time delays τ_1 and τ_2 is negligible under the narrowband assumption, allowing the two delays to be dropped.

The expression for image intensity can now be written

$$I_i(u, v) = \iint_{-\infty}^{\infty} d\xi_1 d\eta_1 \iint_{-\infty}^{\infty} d\xi_2 d\eta_2 h(u - \xi_1, v - \eta_1) h^*(u - \xi_2, v - \eta_2),$$

$$I_i(u, v) = \int_{-\infty}^{\infty} d\tilde{\xi}_1 d\tilde{\eta}_1 \int_{-\infty}^{\infty} d\tilde{\xi}_2 d\tilde{\eta}_2 h(u - \tilde{\xi}_1, v - \tilde{\eta}_1) h^*(u - \tilde{\xi}_2, v - \tilde{\eta}_2) \\ \times J_g(\tilde{\xi}_1, \tilde{\eta}_1; \tilde{\xi}_2, \tilde{\eta}_2),$$

(7-8)

where

$$J_g(\tilde{\xi}_1, \tilde{\eta}_1; \tilde{\xi}_2, \tilde{\eta}_2) = U_g(\tilde{\xi}_1, \tilde{\eta}_1; t) U_g^*(\tilde{\xi}_2, \tilde{\eta}_2; t)$$

$$J_g(\tilde{\xi}_1, \tilde{\eta}_1; \tilde{\xi}_2, \tilde{\eta}_2) = \langle U_g(\tilde{\xi}_1, \tilde{\eta}_1; t) U_g^*(\tilde{\xi}_2, \tilde{\eta}_2; t) \rangle$$

(7-9)

is known as the *mutual intensity*, and is a measure of the *spatial coherence* of the light at the two object points.

When the illumination is perfectly *coherent*, the time-varying phasor amplitudes across the object plane differ only by complex constants. Equivalently we may write

$$\begin{aligned} U_g(\xi^1, \eta^1; t) &= U_g(\xi^1, \eta^1) U_g(0, 0; t) |U_g(0, 0; t)|^2 \\ U_g(\xi_1, \eta_1; t) &= U_g(\xi_1, \eta_1) \frac{U_g(0, 0; t)}{\langle |U_g(0, 0; t)|^2 \rangle^{1/2}} \\ U_g(\xi_2, \eta_2; t) &= U_g(\xi_2, \eta_2) \frac{U_g(0, 0; t)}{\langle |U_g(0, 0; t)|^2 \rangle^{1/2}} \end{aligned} \quad (7-10)$$

where the phase of the time-varying phasor at the origin has arbitrarily been chosen as a phase reference, the time-independent U_g are phasor amplitudes *relative* to the time varying phasor amplitude at the origin, and the normalizations have been performed to allow the time-independent phasors to retain correct information about the average power or intensity. Substituting these relations in the definition of mutual intensity, [Eq.\(7-9\)](#), for the coherent case we obtain

$$J_g(\xi^1, \eta^1; \xi^2, \eta^2) = U_g(\xi^1, \eta^1) U_g^*(\xi^2, \eta^2).$$

$$J_g(\xi_1, \eta_1; \xi_2, \eta_2) = U_g(\xi_1, \eta_1) U_g^*(\xi_2, \eta_2). \quad (7-11)$$

When this result is in turn substituted into [\(7-8\)](#) for the intensity, the result is

$$I_i(u, v) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(u - \xi, v - \eta) U_g(\xi, \eta) d\xi d\eta.$$

$$I_i(u, v) = \left| \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(u - \xi, v - \eta) U_g(\xi, \eta) d\xi d\eta \right|^2. \quad (7-12)$$

Finally, defining a time-invariant phasor amplitude U_i in the image space relative to the corresponding phasor amplitude at the origin, the coherent imaging system is found to be described by an amplitude convolution equation,

$$U_i(u, v) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(u - \xi, v - \eta) U_g(\xi, \eta) d\xi d\eta,$$

$$U_i(u, v) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(u - \xi, v - \eta) U_g(\xi, \eta) d\xi d\eta, \quad (7-13)$$

the same result obtained in the monochromatic case. We thus confirm that coherent object illumination yields an imaging system that is linear in *complex amplitude*.

When the object illumination is perfectly *incoherent*, the phasor amplitudes across the object vary in statistically independent fashions. This idealized property may be represented by the equation

$$\begin{aligned} \text{Ug}(\xi^1, \eta^1; t) \text{Ug}^*(\xi^2, \eta^2; t) &= \kappa I_g(\xi^1, \eta^1) \delta(\xi^1 - \xi^2, \eta^1 - \eta^2) \\ \langle U_g(\xi_1, \eta_1; t) U_g^*(\xi_2, \eta_2; t) \rangle &= \kappa I_g(\xi_1, \eta_1) \delta(\xi_1 - \xi_2, \eta_1 - \eta_2) \end{aligned} \quad (7-14)$$

where κ is a real constant. Such a representation is not exact; in actuality, the minimum distance over which coherence can exist is of the order of one wavelength (see [25], [Section 4.4](#), for more details). Nonetheless, provided the coherence area on the object is small compared with a resolution cell size in object space, (7-14) is accurate. When used in (7-9), the result

$$I_i(u, v) = \kappa \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |h(u - \xi, v - \eta)|^2 I_g(\xi, \eta) d\xi d\eta$$

$$I_i(u, v) = \kappa \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |h(u - \xi, v - \eta)|^2 I_g(\xi, \eta) d\xi d\eta \quad (7-15)$$

is obtained. Thus for incoherent illumination, the image intensity is found as a convolution of the *intensity impulse response* $|h|^2$ with the ideal image intensity I_g . Hence we have confirmed that an incoherent imaging system is linear in *intensity*, rather than amplitude. Furthermore, the impulse response of the incoherent mapping is just the squared modulus of the amplitude impulse response.

When the source of illumination is an extended incoherent source, it is possible to specify the conditions under which the imaging system will behave substantially as an incoherent system and substantially as a coherent system (see [135], p. 283). Let θ_s represent the effective angular diameter of the incoherent source that illuminates the object, θ_p the angular diameter of the entrance pupil of the imaging system, and θ_o the angular diameter of the angular spectrum of the object, all angles being measured from the object plane. Then for a planar, non-diffuse object such as a microscope slide or a lithographic mask, the system can be shown to behave as an incoherent system provided

$$\theta_s \geq \theta_o + \theta_p$$

$$\theta_s \geq \theta_o + \theta_p$$

and will behave as a coherent system when

$$\theta_s \ll \theta_p.$$

$$\theta_s \ll \theta_p.$$

For conditions between these extremes, the system will behave as a *partially coherent* system, the treatment of which is beyond the scope of this discussion. For information on partially coherent imaging systems, see, for example, [Chapter 7](#) of [135].

If the object of interest has a rough surface and transmits or reflects light diffusely, the fine structure of the surface is generally not of interest, and incoherence will be achieved if

$$\theta_s \gg \theta_p,$$

$$\theta_s \gg \theta_p,$$

or, in words, if the coherence width of the illumination is much smaller than the lateral resolution of the imaging system.

The above results are strictly valid for near-monochromatic illumination. Note that in the case of a diffuse object illuminated coherently, the image would contain significant amounts of speckle. However, if the illumination is non-monochromatic, the speckle can be smoothed out due to different speckle patterns being produced by different wavelength regions of the source (see [Section 7.5.3](#) of this book and Section 7.7.4 in [135]). The result of this superposition of independent speckle patterns is an image that can appear more incoherent than coherent, depending on the bandwidth of the source.

7.2 Frequency Response for Diffraction-Limited Coherent Imaging

We turn now to the central topic of this chapter, the frequency analysis of imaging systems. Attention in this section is devoted to imaging systems with coherent illumination. Systems with incoherent illumination will be treated in [Section 7.3](#).

As emphasized previously, a coherent imaging system is linear in complex amplitude. This implies, of course, that such a system provides a highly nonlinear intensity mapping. If frequency analysis is to be applied in its usual form, it must be applied to the linear *amplitude* mapping.

7.2.1 The Amplitude Transfer Function

Our analysis of coherent systems has yielded a space-invariant form of the amplitude mapping, as evidenced by the convolution equation ([7-13](#)). We would anticipate, then, that transfer-function concepts can be applied directly to this system, provided it is done on an amplitude basis. To do so, define the following frequency spectra⁸ of the input and output, respectively:

$$G_g(f_X, f_Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U_g(u, v) \exp[-j2\pi(f_X u + f_Y v)] du dv$$

$$\begin{aligned} G_g(f_X, f_Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U_g(u, v) \exp[-j2\pi(f_X u + f_Y v)] du dv \\ G_i(f_X, f_Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U_i(u, v) \exp[-j2\pi(f_X u + f_Y v)] du dv. \end{aligned}$$

In addition, define the *amplitude transfer function* H as the Fourier transform of the space-invariant amplitude impulse response,

$$H(f_X, f_Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(u, v) \exp[-j2\pi(f_X u + f_Y v)] du dv.$$

$$H(f_X, f_Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(u, v) \exp[-j2\pi(f_X u + f_Y v)] du dv.$$

(7-16)

Now applying the convolution theorem to ([7-13](#)), it follows directly that

$$G_i(f_X, f_Y) = H(f_X, f_Y) G_g(f_X, f_Y).$$

$$G_i(f_X, f_Y) = H(f_X, f_Y) G_g(f_X, f_Y).$$

(7-17)

Thus the effects of the diffraction-limited imaging system have been expressed, at least formally, in the frequency domain. It now remains to relate H more directly to the physical characteristics of the imaging system itself.

To this end, note that while (7-16) defines H as the Fourier transform of the amplitude point-spread function h , this latter function is itself a Fraunhofer diffraction pattern and can be expressed as a scaled Fourier transform of the pupil function (cf. 7-5). Thus

$$H(f_X, f_Y) = \mathcal{F} \left\{ 1/\lambda z_i^2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(x, y) \exp(-j2\pi/\lambda z_i(ux + vy)) dx dy \right\} = P(-\lambda z_i f_X, -\lambda z_i f_Y).$$

$$\begin{aligned} H(f_X, f_Y) &= \mathcal{F} \left\{ \frac{1}{\lambda^2 z_i^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(x, y) \exp \left\{ -j\frac{2\pi}{\lambda z_i}(ux + vy) \right\} dx dy \right\} \\ &= P(-\lambda z_i f_X, -\lambda z_i f_Y). \end{aligned}$$

(7-18)

For notational convenience we ignore the negative signs in the arguments of P (almost all applications of interest to us here have pupil functions that are symmetrical in x and y). Thus

$$H(f_X, f_Y) = P(\lambda z_i f_X, \lambda z_i f_Y).$$

$$H(f_X, f_Y) = P(\lambda z_i f_X, \lambda z_i f_Y).$$

(7-19)

This relation is of the utmost importance; it supplies very revealing information about the behavior of diffraction-limited coherent imaging systems in the frequency domain. If the pupil function P is indeed unity within some region and zero otherwise, then there exists a finite passband in the frequency domain within which the diffraction-limited imaging system passes all frequency components without amplitude or phase distortion.⁹ At the boundary of this passband the frequency response suddenly drops to zero, implying that frequency components outside the passband are completely eliminated.

Finally we give some intuitive explanation as to why the scaled pupil function plays the role of the amplitude transfer function. Remember that in order to completely remove the quadratic-phase factor across the object, the object should be illuminated with a spherical wave, in this case converging towards the point where the entrance pupil is pierced by the optical axis (cf. discussion leading up to Fig. 6.9). The converging spherical illumination causes the Fourier components of the object amplitude transmittance to appear in the entrance pupil, as well as in the exit pupil, since the latter is the image of the former (see Appendix B). Thus the pupil sharply limits the range of Fourier components passed by the system. If the converging illumination is not present, the same conclusion is approximately true, especially for an object of sufficiently small extent in the object plane, as was discussed in connection with Fig. 6.10.

7.2.2 Examples of Amplitude Transfer Functions

To illustrate the frequency response of diffraction-limited coherent imaging systems, consider the amplitude transfer functions of systems with square (width w) and circular (diameter w) pupils. For these two cases, we have, respectively,

$$P(x, y) = \text{rect}(x/w) \text{rect}(y/w) = \text{circ}(x^2 + y^2/w^2).$$

$$\begin{aligned} P(x, y) &= \text{rect}\left(\frac{x}{w}\right) \text{rect}\left(\frac{y}{w}\right) \\ P(x, y) &= \text{circ}\left(\frac{\sqrt{x^2 + y^2}}{w/2}\right). \end{aligned}$$

Thus, from (7-19), the corresponding amplitude transfer functions are

$$\begin{aligned} H(f_X, f_Y) &= \text{rect}\left(\frac{\lambda z_i f_X}{w}\right) \text{rect}\left(\frac{\lambda z_i f_Y}{w}\right) \\ (7-20) \end{aligned}$$

$$H(f_X, f_Y) = \text{circ}\left(\lambda z_i \frac{\sqrt{f_X^2 + f_Y^2}}{w/2}\right)$$

$$\begin{aligned} H(f_X, f_Y) &= \text{circ}\left(\lambda z_i \frac{\sqrt{f_X^2 + f_Y^2}}{w/2}\right) \\ (7-21) \end{aligned}$$

These functions are illustrated in Fig. 7.3. Note that a cutoff frequency f_o can be defined in both cases by

$$f_o = w/2\lambda z_i$$

$$f_o = \frac{w/2}{\lambda z_i}$$

$$(7-22)$$

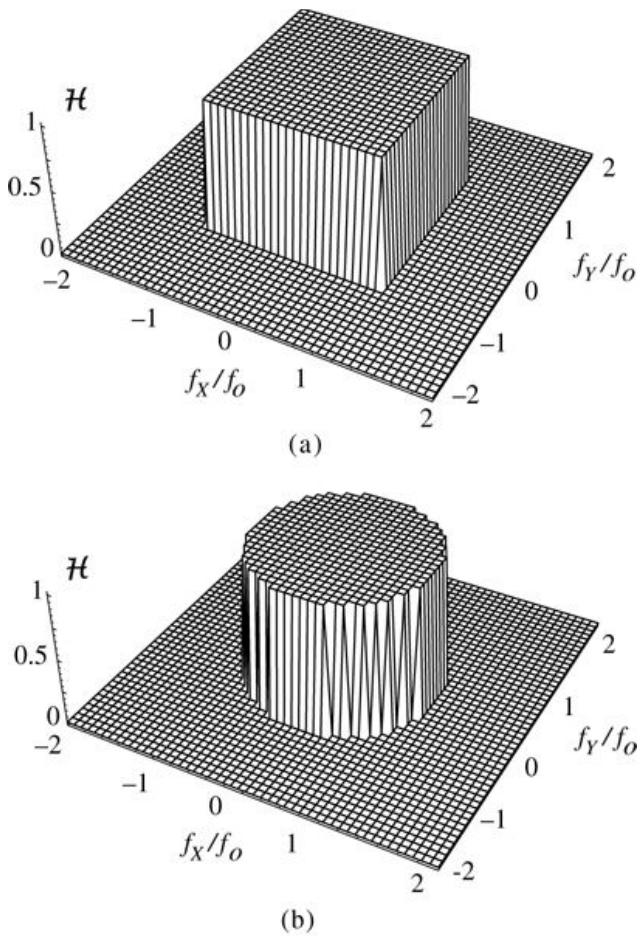


Figure 7.3

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 7.3 Amplitude transfer functions for diffraction-limited systems with (a) square and (b) circular exit pupils.

Each of the two illustrations shows a 3 dimensional projection at the center of a square horizontal plane whose adjoining sides are axes marked from minus 2 to +2; the axes represent f_x/f_o and f_y/f_o . The vertical axis H is marked from 0 to 1. In illustration a, the projection is a cube while in the illustration b, it is a cylinder.

where in the circular case this cutoff is uniform in all directions in the frequency plane, while in the square case this cutoff applies only along the f_x/f_o and f_y/f_o axes. To illustrate a particular order of magnitude of f_o , suppose that $w=2.5$ cm, $z_i=10$ cm, and $\lambda=\lambda=5\times 10^{-4}$ cm. Then the cutoff frequency is 250 cycles/mm.

7.3 Frequency Response for Diffraction-Limited Incoherent Imaging

In the coherent case, the relation between the pupil and the amplitude transfer function has been seen to be a very direct and simple one. When the object illumination is incoherent, the transfer function of the imaging system will be seen to be determined by the pupil again, but in a less direct and somewhat more interesting way. The theory of imaging with incoherent light has, therefore, a certain extra richness not present in the coherent case. We turn now to considering this theory; again attention will be centered on *diffraction-limited* systems, although the discussion that immediately follows applies to all incoherent systems, regardless of their aberrations.

7.3.1 The Optical Transfer Function

Imaging systems that use incoherent illumination have been seen to obey the *intensity* convolution integral

$$\begin{aligned} I_i(u,v) &= k \iint_{-\infty}^{\infty} |h(u - \xi, v - \eta)|^2 I_g(\xi, \eta) d\xi d\eta. \\ I_i(u, v) &= k \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |h(u - \tilde{\xi}, v - \tilde{\eta})|^2 I_g(\tilde{\xi}, \tilde{\eta}) d\tilde{\xi} d\tilde{\eta}. \end{aligned} \quad (7-23)$$

Such systems should therefore be frequency-analyzed as linear mappings of intensity distributions. To this end, let the *normalized* frequency spectra of I_g and I_i be defined by

$$\begin{aligned} \mathcal{G}_g(f_X, f_Y) &= \iint_{-\infty}^{\infty} I_g(u, v) \exp[-j2\pi(f_X u + f_Y v)] du dv \\ &= \frac{\iint_{-\infty}^{\infty} I_g(u, v) \exp[-j2\pi(f_X u + f_Y v)] du dv}{\iint_{-\infty}^{\infty} I_g(u, v) du dv} \end{aligned} \quad (7-24)$$

$$\begin{aligned} \mathcal{G}_i(f_X, f_Y) &= \iint_{-\infty}^{\infty} I_i(u, v) \exp[-j2\pi(f_X u + f_Y v)] du dv \\ &= \frac{\iint_{-\infty}^{\infty} I_i(u, v) \exp[-j2\pi(f_X u + f_Y v)] du dv}{\iint_{-\infty}^{\infty} I_i(u, v) du dv}. \end{aligned} \quad (7-25)$$

The normalization of the spectra by their “zero-frequency” values is partly for mathematical convenience, and partly for a more fundamental reason. It can be shown that any real and

nonnegative function, such as I_g or I_i , has a Fourier transform which achieves its maximum value at the origin. We choose that maximum value as a normalization constant in defining \mathcal{G}_g and \mathcal{G}_i . Since intensities are nonnegative quantities, they always have a spectrum that is nonzero at the origin. The visual quality of an image depends strongly on the “contrast” of the

image, or the relative strengths of the information-bearing portions of the image and the often-present background. Hence the spectra are normalized by that background.

In a similar fashion, the normalized transfer function of the system can be defined by

$$\mathcal{H}(f_X, f_Y) = \iint_{-\infty}^{\infty} |h(u, v)|^2 \exp[-j2\pi(f_X u + f_Y v)] du dv / \iint_{-\infty}^{\infty} |h(u, v)|^2 du dv.$$

$$\mathcal{H}(f_X, f_Y) = \frac{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |h(u, v)|^2 \exp[-j2\pi(f_X u + f_Y v)] du dv}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |h(u, v)|^2 du dv}.$$

(7-26)

Application of the convolution theorem to (7-23) then yields the frequency-domain relation

$$i(f_X, f_Y) = \mathcal{H}(f_X, f_Y) g(f_X, f_Y).$$

$$i(f_X, f_Y) = \mathcal{H}(f_X, f_Y) g(f_X, f_Y).$$

(7-27)

By international agreement, the function \mathcal{H} is known as the *optical transfer function* (OTF) of the system. Its modulus $|\mathcal{H}|$ is known as the *modulation transfer function* (MTF). Note that $\mathcal{H}(f_X, f_Y)$ simply specifies the complex weighting factor applied by the system to the frequency component at (f_X, f_Y) , relative to the weighting factor applied to the zero-frequency component.

Since the definitions of both the amplitude transfer function and the optical transfer function involve the function h , we might expect some specific relationship between the two. In fact, such a relationship exists and can be readily found with the help of the autocorrelation theorem of [Chapter 2](#). Since

$$H(f_X, f_Y) = \mathcal{F}h$$

$$H(f_X, f_Y) = \mathcal{F}\{h\}$$

and

$$\mathcal{H}(f_X, f_Y) = \mathcal{F}|h|^2 \iint_{-\infty}^{\infty} |h(u, v)|^2 du dv,$$

$$\mathcal{H}(f_X, f_Y) = \frac{\mathcal{F}\{|h|^2\}}{\iint_{-\infty}^{\infty} \int_{-\infty}^{\infty} |h(u, v)|^2 du dv},$$

it follows (with the help of Rayleigh's theorem) that

$$\mathcal{H}(f_X, f_Y) = \iint_{-\infty}^{\infty} \int_{-\infty}^{\infty} H(p', q') H^*(p' - f_X, q' - f_Y) dp' dq' \iint_{-\infty}^{\infty} \int_{-\infty}^{\infty} |H(p', q')|^2 dp' dq'.$$

$$\mathcal{H}(f_X, f_Y) = \frac{\int_{-\infty}^{\infty} \int H(p', q') H^*(p' - f_X, q' - f_Y) dp' dq'}{\int_{-\infty}^{\infty} \int |H(p', q')|^2 dp' dq'}. \quad (7-28)$$

The simple change of variables

$$p = p' - f_X 2q = q' - f_Y 2$$

$$p = p' - \frac{f_X}{2}q = q' - \frac{f_Y}{2}$$

results in the symmetrical expression

$$\mathcal{H}(f_X, f_Y) = \int_{-\infty}^{\infty} \int H(p + f_X 2q, q + f_Y 2) H^*(p - f_X 2q, q - f_Y 2) dp dq.$$

$$\mathcal{H}(f_X, f_Y) = \frac{\int_{-\infty}^{\infty} \int |H(p, q)|^2 dp dq}{\int_{-\infty}^{\infty} \int |H(p, q)|^2 dp dq}.$$

(7-29)

Thus the OTF is the normalized autocorrelation function of the amplitude transfer function!

[Equation \(7-29\)](#) will serve as our primary link between the properties of coherent and incoherent systems. Note that it is entirely valid for systems both with and without aberrations.

7.3.2 General Properties of the OTF

A number of very simple and elegant properties of the OTF can be stated based only on knowledge that it is a normalized autocorrelation function. The most important of these properties are as follows:

1. $\mathcal{H}(0,0)=1$
2. $\mathcal{H}(-f_X, -f_Y) = \mathcal{H}^*(-f_X, f_Y) = \mathcal{H}^*(f_X, f_Y)$
3. $|\mathcal{H}(f_X, f_Y)| \leq |\mathcal{H}(0,0)|$

Property 1 follows directly by substitution of ($f_X=0, f_Y=0$) ($f_X = 0, f_Y = 0$) in (7-29). The proof of Property 2 is left as an exercise for the reader, it being no more than a statement that the Fourier transform of a real function has Hermitian symmetry.

The proof that the MTF at any frequency is always less than its zero-frequency value of unity requires more effort. To prove Property 3 we use Schwarz's inequality ([\[276\]](#), p. 177), which can be stated as follows: If $X(p, q)$ and $Y(p, q)$ are any two complex-valued functions of (p, q) , then

$$\iint XY dp dq \leq \iint |X|^2 dp dq \iint |Y|^2 dp dq$$

$$\left| \iint XY dp dq \right|^2 \leq \left(\iint |X|^2 dp dq \right)^2 \left(\iint |Y|^2 dp dq \right)$$

(7-30)

with equality if and only if $Y = KX^*$ where K is a complex constant. Letting

$$X(p, q) = H + f_X p - f_Y q \quad \text{and} \quad Y(p, q) = H^* - f_X p + f_Y q$$

$$X(p, q) = H \left(p + \frac{f_X}{2}, q + \frac{f_Y}{2} \right) \quad \text{and} \quad Y(p, q) = H^* \left(p - \frac{f_X}{2}, q - \frac{f_Y}{2} \right)$$

we find

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |H(p + \frac{f_X}{2}, q + \frac{f_Y}{2})|^2 dp dq \leq \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |H(p, q)|^2 dp dq \leq \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |H(p, q)|^2 dp dq.$$

$$\begin{aligned} & \left| \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H \left(p + \frac{f_X}{2}, q + \frac{f_Y}{2} \right) H^* \left(p - \frac{f_X}{2}, q - \frac{f_Y}{2} \right) dp dq \right|^2 \\ & \leq \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left| H \left(p + \frac{f_X}{2}, q + \frac{f_Y}{2} \right) \right|^2 dp dq \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left| H \left(p - \frac{f_X}{2}, q - \frac{f_Y}{2} \right) \right|^2 dp dq \\ & = \left[\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |H(p, q)|^2 dp dq \right]^2. \end{aligned}$$

Normalizing by the right-hand side of the inequality, it follows that $|\mathcal{H}(f_X, f_Y)|$ is never greater than unity.

Finally, it should be pointed out that while the OTF is always unity at the zero frequency, this does not imply that the absolute intensity level of the image background is the same as the absolute intensity level of the object background. The normalization used in the definition of the OTF has removed all information about absolute intensity levels.

7.3.3 The OTF of an Aberration-Free System

To this point, our discussions have been equally applicable to systems with and without aberrations. We now consider the special case of a diffraction-limited incoherent system. Recall that for coherent systems we have

$$H(f_X, f_Y) = P(\lambda z_i f_X, \lambda z_i f_Y).$$

$$H(f_X, f_Y) = P(\lambda z_i f_X, \lambda z_i f_Y).$$

For an incoherent system, it follows from (7-29) with a change of variables $x = \lambda z_i f_X$ and $y = \lambda z_i f_Y$, that

$$\mathcal{H}(f_X, f_Y) = \iint_{-\infty}^{\infty} P_x + \lambda z_i f_X^2, y + \lambda z_i f_Y^2 P_x - \lambda z_i f_X^2, y - \lambda z_i f_Y^2 dx dy \iint_{-\infty}^{\infty} P(x, y) dx dy,$$

$$\mathcal{H}(f_X, f_Y) = \frac{\iint_{-\infty}^{\infty} P\left(x + \frac{\lambda z_i f_X}{2}, y + \frac{\lambda z_i f_Y}{2}\right) P\left(x - \frac{\lambda z_i f_X}{2}, y - \frac{\lambda z_i f_Y}{2}\right) dx dy}{\iint_{-\infty}^{\infty} P^2(x, y) dx dy},$$

(7-31)

where, in the denominator, when P equals only unity or zero, P^2 can be replaced by P . The expression (7-31) for \mathcal{H} lends itself to an extremely important geometrical interpretation when P equals only one or zero. The numerator represents the area of overlap of two displaced pupil functions, one centered at $(\lambda z_i f_X / 2, \lambda z_i f_Y / 2)$ and the second centered on the diametrically opposite point $(-\lambda z_i f_X / 2, -\lambda z_i f_Y / 2)$. The denominator simply normalizes the area of overlap by the total area of the pupil. Thus

$$\mathcal{H}(f_X, f_Y) = \text{area of overlap} / \text{total area}.$$

$$\mathcal{H}(f_X, f_Y) = \frac{\text{area of overlap}}{\text{total area}}.$$

To calculate the OTF of a diffraction-limited system, the steps indicated by this interpretation can be directly performed, as illustrated in Fig. 7.4. For simple geometrical shapes, closed-form expressions for the normalized overlap area can be found (see examples to follow). Note that this geometrical interpretation of the OTF implies that the OTF of a diffraction-limited system is always *real* and *nonnegative*. It is not necessarily a monotonically decreasing function of frequency, however (see, for example, Prob. 7-3).

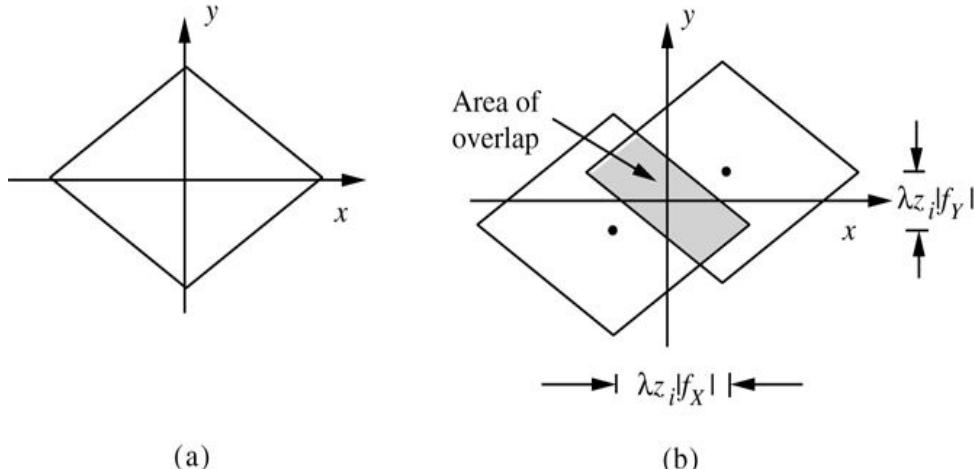


Figure 7.4

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 7.4 Geometrical interpretations of the OTF of a diffraction-limited system. (a) The pupil function-total area is the denominator of the OTF; (b) two displaced pupil functions—the shaded area is the numerator of the OTF.

Both illustrations show horizontal axis x and vertical axis y. Illustration a shows a rhombus centered at the origin with its diagonals overlapping the two axes. Illustration b shows two identical overlapping rhomboids such that the center of one is in the third quadrant while that of the other is in the first quadrant, the downward slanting sides of one are parallel to those of the other. The horizontal distance between the centers is marked $\lambda z_i |f_X|$. The vertical distance between the centers is marked $\lambda z_i |f_Y|$. The overlapping part is highlighted.

For complicated pupils, the OTF can be calculated with the help of a digital computer. A straightforward way to perform such a calculation is to inverse Fourier transform the reflected pupil function $P(-x, -y)$ (or equivalently to Fourier transform the pupil function $P(x, y)$ $P(x, y)$), thereby finding the amplitude point-spread function, take the squared magnitude of this quantity (thus finding the intensity point-spread function), and take the Fourier transform of this result to find the unnormalized OTF. A final normalization to unity at the origin should then be performed.

To lend further physical insight into the OTF, consider the ways in which a sinusoidal component of intensity at a particular frequency pair (f_X, f_Y) can be generated in the image. We claim that such a fringe can be generated only by interference of light in the image plane from two separate patches on the exit pupil of the system, with a separation between patches that is $(\lambda z_i |f_X|, \lambda z_i |f_Y|)$. Only when light contributions from two patches having this particular separation interfere can a fringe with this frequency be generated (cf. [Prob. 7-1](#)). However, there are many different pairs of patches of this separation that can be embraced by the pupil of the system. In fact, the relative weight given by the system to this particular frequency pair is determined by how many different ways such a separation can be fit into the pupil. The number of ways a particular separation can be fit into the exit pupil is proportional to the area of overlap of two pupils separated by this particular spacing. See [Fig. 7.5](#).

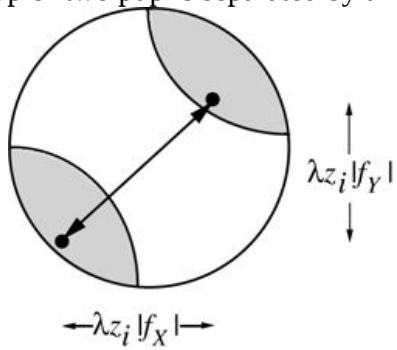


Figure 7.5

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 7.5 Light from patches separated by $(\lambda z_i |f_X|, \lambda z_i |f_Y|)$ interferes to produce a sinusoidal fringe at frequency (f_X, f_Y) . The shaded areas on the pupil are the areas within which the light patches can reside while retaining this special separation.

The illustration shows a circle with two gray patches, one near the upper end and the other near the lower end, each with a dot. A double headed arrow connects the two dots. The horizontal distance

between the dots is marked $\lambda z_i |f_X|$. The vertical distance between the dots is marked $\lambda z_i |f_Y|$.

7.3.4 Examples of Diffraction-Limited OTFs

We consider now as examples the OTFs that correspond to diffraction-limited systems with square (width w) and circular (diameter w) pupils. [Figure 7.6](#) illustrates the calculation for the square case. The area of overlap is evidently

$$A(f_X, f_Y) = (w - \lambda z_i |f_X|)(w - \lambda z_i |f_Y|) |f_X| \leq w/\lambda z_i, |f_Y| \leq w/\lambda z_i \\ 0 \text{ otherwise.}$$

$$\mathcal{A}(f_X, f_Y) = \begin{cases} (w - \lambda z_i |f_X|)(w - \lambda z_i |f_Y|) & |f_X| \leq w/\lambda z_i, \\ & |f_Y| \leq w/\lambda z_i \\ 0 & \text{otherwise.} \end{cases}$$

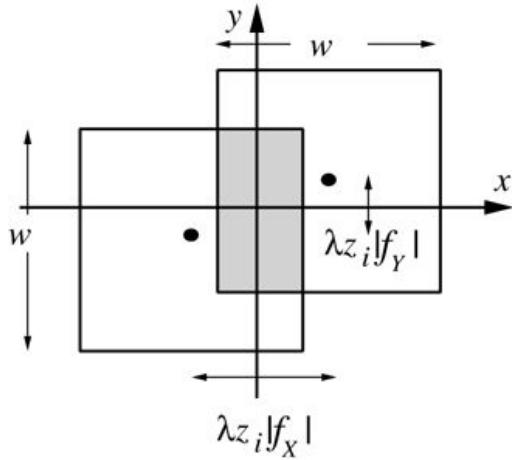


Figure 7.6

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 7.6 Calculation of the OTF for a square aperture.

The illustration shows horizontal axis x and vertical axis y with two identical overlapping squares of width w such that the center of one is in the third quadrant while that of the other is in the first quadrant, the vertical sides of one are parallel to those of the other. The horizontal distance between the centers is marked $\lambda z_i |f_X|$. The vertical distance between the centers is marked $\lambda z_i |f_Y|$. The overlapping part is highlighted.

When this area is normalized by the total area w^2 , the result becomes

$$\mathcal{H}(f_X, f_Y) = \Lambda(f_X)^2 \Lambda(f_Y)^2$$

$$\mathcal{H}(f_X, f_Y) = \Lambda\left(\frac{f_X}{2f_o}\right) \Lambda\left(\frac{f_Y}{2f_o}\right)$$

(7-32)

where Λ is the triangle function of [Chapter 2](#), and f_o is the cutoff frequency of the same system when used with *coherent* illumination,

$$f_o = w/2\lambda z_i.$$

$$f_o = \frac{w/2}{\lambda z_i}.$$

Note that the cutoff frequency of the incoherent system occurs at frequency $2f_o$ along the f_X and f_Y axes.¹⁰ The OTF represented by [Eq.\(7-32\)](#) is illustrated in [Fig. 7.7](#).

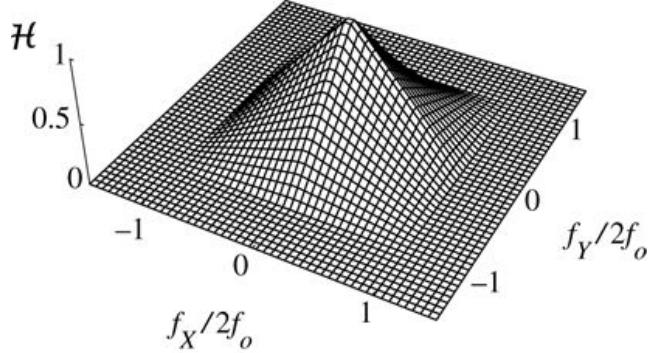


Figure 7.7

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 7.7 The optical transfer function of a diffraction-limited system with a square pupil.

The illustration shows a 3 dimensional projection at the center of a square horizontal plane whose adjoining sides are axes marked from minus 1 to +1; the axes represent $f_X / 2f_o$ and $f_Y / 2f_o$. The vertical axis H is marked from 0 to 1. The projection is conical.

When the pupil is circular, the calculation is not quite so straightforward. Since the OTF will clearly be circularly symmetric, it suffices to calculate H along the positive f_X axis. As illustrated in [Fig. 7.8](#), the area of overlap may be regarded as being equal to four times the shaded area B of the circular sector $A+B$. But the area of the circular sector is

$$\text{Area}(A+B) = \theta 2\pi (w/2)^2 = \arccos(\lambda z_i f_X/w) 2\pi (w/2)^2$$

$$\text{Area } (A+B) = \left[\frac{\theta}{2\pi} \right] (\pi(w/2)^2) = \left[\frac{\arccos(\lambda z_i f_X/w)}{2\pi} \right] (\pi(w/2)^2)$$

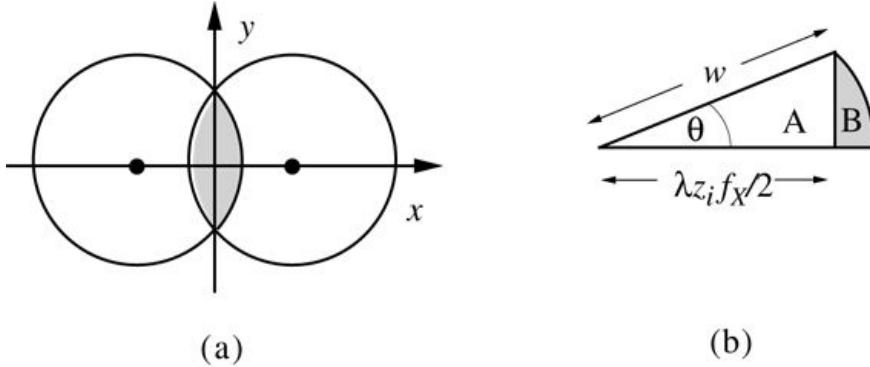


Figure 7.8

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 7.8 Calculation of the area of overlap of two displaced circles. (a) Overlapping circles, (b) geometry of the calculation.

Illustration a shows two overlapping circles with their centers on the horizontal axis *x*, one to the left of the origin and the other to the right. The circumferences of the circles intersect at points that lie on the vertical axis *y*, one above the origin and the other below. Illustration b shows a sector of angle theta and radius *w*. A perpendicular from where the radius intersects the curve is dropped to the other radius, dividing the sector into two parts: a right-angled triangle labeled *A* and the other part labeled *B* and shaded gray. The distance between the center and the vertex of the right angle is marked lambda *z* subscript *i* *f* subscript *X* / 2.

while the area of the triangle *A* is

$$\text{Area}(A) = \frac{1}{2} \lambda z_i f_X^2 w^2 - \lambda z_i f_X^2 w^2.$$

$$\text{Area } (A) = \frac{1}{2} \left(\frac{\lambda z_i f_X}{2} \right) \sqrt{w^2 - \left(\frac{\lambda z_i f_X}{2} \right)^2}.$$

Finally, we have

$$\mathcal{H}(f_X, 0) = 4[\text{area}(A+B) - \text{area}(A)]\pi w^2$$

$$\mathcal{H}(f_X, 0) = \frac{4[\text{area } (A + B) - \text{area } (A)]}{\pi w^2}$$

or, for a general radial distance ρ in the frequency plane,

$$\mathcal{H}(\rho) = 2\pi \arccos \rho_o - \rho_o^2 \rho_o^2 - \rho_o^2 \rho_o^2 \rho \leq 2\rho_o^2 \text{ otherwise.}$$

$$\mathcal{H}(\rho) = \begin{cases} \frac{2}{\pi} \left[\arccos \left(\frac{\rho}{2\rho_o} \right) - \frac{\rho}{2\rho_o} \sqrt{1 - \left(\frac{\rho}{2\rho_o} \right)^2} \right] & \rho \leq 2\rho_o \\ 0 & \text{otherwise.} \end{cases}$$

(7-33)

The quantity ρ_o is the cutoff frequency of the coherent system,

$$\rho_o = w / 2\lambda z_i.$$

$$\rho_o = \frac{w/2}{\lambda z_i}.$$

Referring to Fig. 7.9, the OTF is again seen to extend to a frequency that is twice the coherent cutoff frequency.

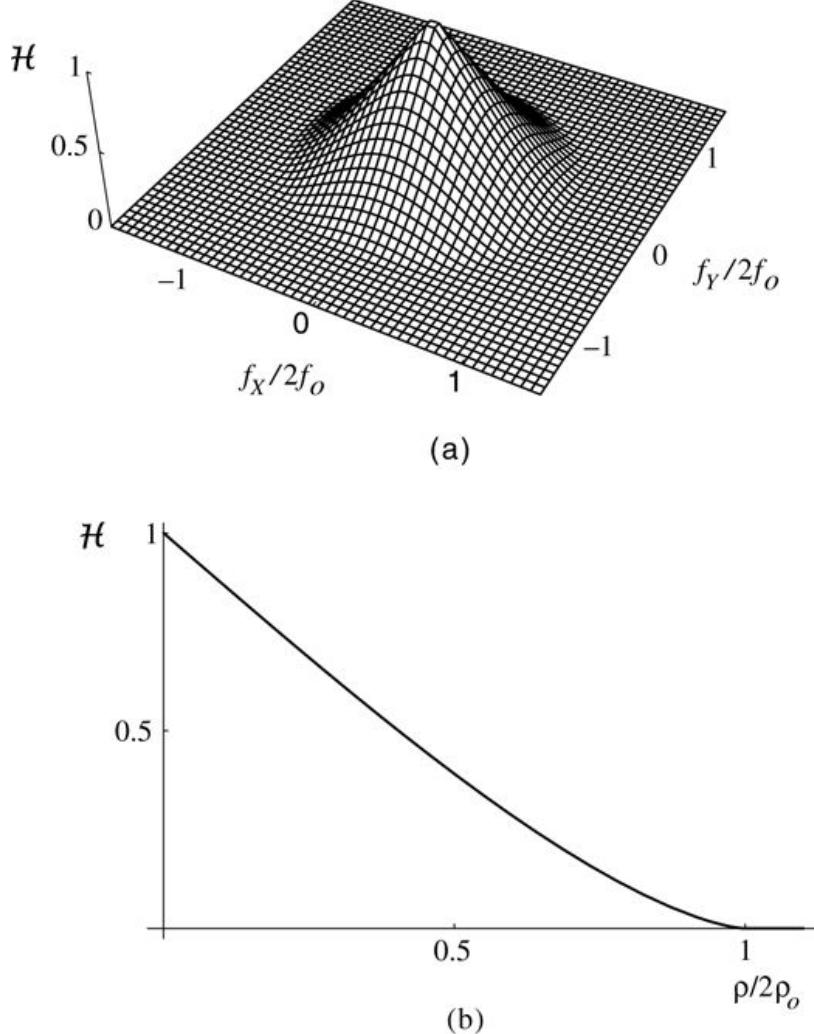


Figure 7.9
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 7.9 The optical transfer function of a diffraction-limited system with a circular pupil. (a) Three-dimensional perspective, (b) cross section.

The illustration shows a 3 dimensional projection at the center of a square horizontal plane whose adjoining sides are axes marked from minus 1 to +1; the axes represent $f_x/2f_o$ and $f_y/2f_o$. The vertical axis H is marked from 0 to 1. The projection is conical. The graph shows horizontal axis $\rho/2\rho_o$ and vertical axis H , both plotting values from 0 to 1. A smooth downward sloping line connects the 1 mark on the vertical axis to the 1 mark on the horizontal axis.

7.4 Aberrations and Their Effects on Frequency Response

In the development of a generalized model of an imaging system, it was specifically assumed that the presence of a point-source object yielded at the exit pupil a perfect spherical wave, converging toward the ideal geometrical image point. Such a system was called *diffraction-limited*. We consider now the effects of *aberrations*, or departures of the exit-pupil wavefront from ideal spherical form. Aberrations can arise in a variety of ways, ranging from a defect as simple as a focusing error to inherent properties of perfectly spherical lenses, such as spherical aberration. A complete treatment of aberrations and their detailed effects on frequency response is beyond the scope of this development. Rather we concentrate on very general effects and illustrate with one relatively simple example. For a more complete treatment of various types of aberrations and their effects on frequency response, see, for example, [373], [173], [366], or [239].

7.4.1 The Generalized Pupil Function

When an imaging system is diffraction limited, the (amplitude) point-spread function has been seen to consist of the Fraunhofer diffraction pattern of the exit pupil, centered on the ideal image point. This fact suggests a convenient artifice which will allow aberrations to be directly included in our previous results. Specifically, when wavefront errors exist, we can imagine that the exit pupil is illuminated by a perfect spherical wave, but that a phase-shifting plate exists in the aperture, thus deforming the wavefront that leaves the pupil. If the phase error at the point (x, y) is represented by $kW(x, y)$, where $k=2\pi/\lambda$ and W is an effective path-length error, then the complex amplitude transmittance $P(x, y)$ of the imaginary phase-shifting plate is given by

$$P(x, y) = P(x, y) \exp[jkW(x, y)].$$

$$\mathbf{P}(x, y) = P(x, y) \exp[jkW(x, y)].$$

(7-34)

The complex function \mathbf{P} may be referred to as the *generalized* pupil function. The amplitude point-spread function of an aberrated coherent system is simply the Fraunhofer diffraction pattern of an aperture with amplitude transmittance \mathbf{P} . The intensity impulse response of an aberrated incoherent system is, of course, the squared magnitude of the amplitude impulse response.

[Figure 7.10](#) shows the geometry that defines the aberration function W . If the system were free from aberrations, the exit pupil would be filled by a perfect spherical wave converging towards the ideal image point. We regard an ideal spherical surface, centered on the ideal image point and passing through the point where the optical axis pierces the exit pupil, as defining a *Gaussian reference sphere* with respect to which the aberration function can be defined. If we trace a ray backward from the ideal image point to the coordinates (x, y) in the exit pupil, the aberration function $W(x, y)$ is the path-length error accumulated by that ray as it passes from the Gaussian reference sphere to the actual wavefront, the latter wavefront also being

defined to intercept the optical axis in the exit pupil. The error can be positive or negative, depending on whether the actual wavefront lies to the left or to the right (respectively) of the Gaussian reference sphere.

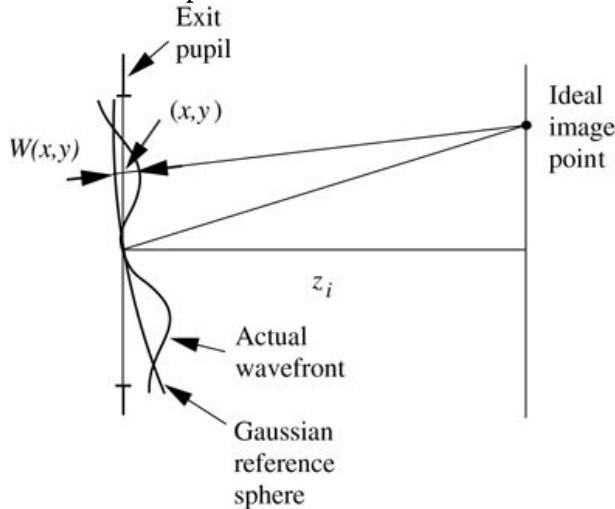


Figure 7.10

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 7.10 Geometry for defining the aberration function.

The illustration shows a vertical exit pupil to the left and the image plane to the right, their centers connected by a horizontal line of length z_i . The Gaussian reference sphere is a steep downward slanting curve passing through the center of the exit pupil. The actual wavefront is a downward slanting wavelike curve of two peaks and a valley, such that the bottom of the valley coincides with the center of the exit pupil. Point (x, y) lies in the upper half of the exit pupil. A straight line connects the ideal image point in the upper half of the image plane to (x, y) . The line when extended to the left beyond (x, y) touches the Gaussian reference sphere at $W(x, y)$.

7.4.2 Effects of Aberrations on the Amplitude Transfer Function

When considering a diffraction-limited coherent system, the transfer function was found by noting that (1) the impulse response is the Fourier transform of the pupil function, and (2) the amplitude transfer function is the Fourier transform of the amplitude impulse response. As a consequence of the two Fourier transform relations, the amplitude transfer function was found to be proportional to a scaled pupil function P^P . Identical reasoning can be used when aberrations are present, provided the generalized pupil function P^P replaces P^P . Thus the amplitude transfer function is written

$$H(f_X, f_Y) = P(\lambda z_i f_X, \lambda z_i f_Y) = P(\lambda z_i f_X, \lambda z_i f_Y) \exp[jkW(\lambda z_i f_X, \lambda z_i f_Y)].$$

$$H(f_X, f_Y) = P(\lambda z_i f_X, \lambda z_i f_Y) = P(\lambda z_i f_X, \lambda z_i f_Y) \exp[jkW(\lambda z_i f_X, \lambda z_i f_Y)].$$

(7-35)

Evidently the band limitation of the amplitude transfer function, as imposed by the finite exit pupil, is unaffected by the presence of aberrations. The sole effect of aberrations is seen to be the

introduction of *phase distortions* within the passband. Phase distortions can, of course, have a severe effect on the fidelity of the imaging system.

There is little more of a general nature that can be said about the effects of aberrations on a coherent imaging system. Again the result is a very simple one: as we shall now see, the result for an incoherent system is again more complex and, in many respects, more interesting.

7.4.3 Effects of Aberrations on the OTF

Having found the effects of aberrations on the amplitude transfer function, it is now possible, with the help of (7-29), to find the effects on the optical transfer function. To simplify the notation, the function $A(f_X, f_Y)$ is defined as the *area of overlap* of

$$P_{x-\lambda z_i f_X/2, y-\lambda z_i f_Y/2} \text{ and } P_{x+\lambda z_i f_X/2, y+\lambda z_i f_Y/2}.$$

$$P\left(x - \frac{\lambda z_i f_X}{2}, y - \frac{\lambda z_i f_Y}{2}\right) \text{ and } P\left(x + \frac{\lambda z_i f_X}{2}, y + \frac{\lambda z_i f_Y}{2}\right).$$

Thus the OTF of a diffraction-limited system is given, in this new notation, by

$$\mathcal{H}(f_X, f_Y) = \int A(f_X, f_Y) \int dx dy / \int A(0, 0) \int dx dy.$$

$$\mathcal{H}(f_X, f_Y) = \frac{\int \int dx dy}{\int \int dx dy} \frac{\int A(f_X, f_Y)}{\int A(0, 0)}.$$

(7-36)

When aberrations are present, substitution of (7-35) into (7-29) yields

$$\mathcal{H}(f_X, f_Y) = \int A(f_X, f_Y) \int e^{jkWx + \lambda z_i f_X/2, y + \lambda z_i f_Y/2 - Wx - \lambda z_i f_X/2, y - \lambda z_i f_Y/2} \int dx dy / \int A(0, 0) \int dx dy.$$

$$\mathcal{H}(f_X, f_Y) = \frac{\int \int e^{jk \left[W\left(x + \frac{\lambda z_i f_X}{2}, y + \frac{\lambda z_i f_Y}{2}\right) - W\left(x - \frac{\lambda z_i f_X}{2}, y - \frac{\lambda z_i f_Y}{2}\right) \right]} dx dy}{\int \int dx dy / \int A(0, 0) \int dx dy}.$$

(7-37)

This expression allows us, then, to directly relate the wavefront errors and the OTF.

As an important general property, it can be shown that aberrations will *never increase* the MTF (the modulus of the OTF). To prove this property, Schwarz's inequality (7-30) will be used. Let the functions X and Y of that equation be defined by

$$X(x, y) = \exp[jkWx + \lambda z_i f_X/2, y + \lambda z_i f_Y/2] Y(x, y) = \exp[-jkWx - \lambda z_i f_X/2, y - \lambda z_i f_Y/2]$$

$$\begin{aligned} X(x, y) &= \exp \left[jkW \left(x + \frac{\lambda z_i f_X}{2}, y + \frac{\lambda z_i f_Y}{2} \right) \right] \\ Y(x, y) &= \exp \left[-jkW \left(x - \frac{\lambda z_i f_X}{2}, y - \frac{\lambda z_i f_Y}{2} \right) \right]. \end{aligned}$$

Noting that $|X|^2=|Y|^2=1$, it follows that

$$|\mathcal{H}(f_X, f_Y)| \text{with aberrations}^2 = \int A(f_X, f_Y) \int e^{jkWx + \lambda z_i f_X x + \lambda z_i f_Y y - Wx - \lambda z_i f_X x - \lambda z_i f_Y y} dxdy \leq \int A(f_X, f_Y) \int dxdy = |\mathcal{H}(f_X, f_Y)| \text{without aberrations}^2.$$

$$\begin{aligned} & |\mathcal{H}(f_X, f_Y)|_{\text{with aberrations}}^2 \\ &= \left| \frac{\int \int e^{jk \left[W\left(x + \frac{\lambda z_i f_X}{2}, y + \frac{\lambda z_i f_Y}{2}\right) - W\left(x - \frac{\lambda z_i f_X}{2}, y - \frac{\lambda z_i f_Y}{2}\right) \right]} dx dy}{\int \int dx dy}_{\mathcal{A}(0,0)} \right|^2 \\ &\leq \left[\frac{\int \int dx dy}{\mathcal{A}(0,0)} \right]^2 = \left| \mathcal{H}(f_X, f_Y) \right|_{\text{without aberrations}}^2. \end{aligned}$$

Thus aberrations cannot increase the contrast of any spatial-frequency component of the image, and in general will lower the contrast. The absolute cutoff frequency remains unchanged, but severe aberrations can reduce the high-frequency portions of the OTF to such an extent that the effective cutoff is much lower than the diffraction-limited cutoff. In addition, aberrations can cause the OTF to have *negative* (or even *complex*) values in certain bands of frequencies, a result that never occurs for an aberration-free system. When the OTF is negative, image components at that frequency undergo a contrast reversal; i.e., intensity peaks become intensity nulls, and vice versa. An example of this effect will be seen in the section that follows.

7.4.4 Example of a Simple Aberration: A Focusing Error

One of the easiest aberrations to deal with mathematically is a simple error of focus. But even in this simple case, the assumption of a *square* aperture (rather than a circular aperture) is needed to keep the mathematics simple.

When a focusing error is present, the center of curvature of the spherical wavefront converging towards the image of an object point-source lies either to the left or to the right of the image plane. Considering an on-axis point for simplicity, this means that the phase distribution across the exit pupil is of the form

$$\phi(x, y) = -\pi\lambda z_a x^2 + y^2,$$

$$\phi(x, y) = -\frac{\pi}{\lambda z_a} (x^2 + y^2),$$

where $z_a \neq z_i$. The path-length error $W(x, y)$ can then be determined by subtracting the ideal phase distribution from the actual phase distribution,

$$kW(x, y) = -\pi\lambda z_a x^2 + y^2 + \pi\lambda z_i x^2 + y^2.$$

$$kW(x, y) = -\frac{\pi}{\lambda z_a} (x^2 + y^2) + \frac{\pi}{\lambda z_i} (x^2 + y^2).$$

(7-38)

The path-length error is thus given by

$$W(x, y) = -121za - 1zix^2 + y^2,$$

$$W(x, y) = -\frac{1}{2} \left(\frac{1}{z_a} - \frac{1}{z_i} \right) (x^2 + y^2),$$

(7-39)

which is seen to depend quadratically on the space variables in the exit pupil.

For a square aperture of width w , the maximum path-length error at the edge of the aperture along the x or y axes, which we represent by W_m , is given by

$$W_m = -121z_a - 1z_i(w/2)^2.$$

$$W_m = -\frac{1}{2} \left(\frac{1}{z_a} - \frac{1}{z_i} \right) (w/2)^2.$$

(7-40)

The number W_m is a convenient indication of the severity of the focusing error. Using the definition of W_m , we can express the path-length error as

$$W(x, y) = W_m x^2 + y^2 (w/2)^2.$$

$$W(x, y) = W_m \frac{x^2 + y^2}{(w/2)^2}.$$

(7-41)

For the case of a focusing error, the phase θ of the exponential in the numerator of [Eq.\(7-37\)](#) can be simplified substantially as follows:

$$\theta = kWx + \lambda z_i f_X x + \lambda z_i f_Y y - Wx - \lambda z_i f_X x - \lambda z_i f_Y y = kWm(w/2)x + \lambda z_i f_X x + \lambda z_i f_Y y - \lambda z_i f_X x - \lambda z_i f_Y y = 16\pi W_m z_i (f_X x + f_Y y),$$

$$\begin{aligned} \theta &= k \left[W \left(x + \frac{\lambda z_i f_X}{2}, y + \frac{\lambda z_i f_Y}{2} \right) - W \left(x - \frac{\lambda z_i f_X}{2}, y - \frac{\lambda z_i f_Y}{2} \right) \right] \\ &= k \frac{W_m}{(w/2)^2} \left[\left(x + \frac{\lambda z_i f_X}{2} \right)^2 + \left(y + \frac{\lambda z_i f_Y}{2} \right)^2 - \left(x - \frac{\lambda z_i f_X}{2} \right)^2 + \left(y - \frac{\lambda z_i f_Y}{2} \right)^2 \right] \\ &= \frac{16\pi W_m z_i}{w^2} (f_X x + f_Y y), \end{aligned}$$

(7-42)

where W_m is the path-length error at the edge of the pupil. The expression for the numerator of the OTF factors into an integral I_X with respect to x and an integral I_Y with respect to y . Considering just the x integral, we have

$$I_X = \int_{-\infty}^{\infty} \text{rect}(x) + \lambda z_i f_X^2 w \text{rect}(x) - \lambda z_i f_X^2 w \exp(j16\pi W_m z_i w^2 x) dx = \int_{-w/2}^{w/2} |f_X|^2 w \exp(j16\pi W_m z_i w^2 x) dx,$$

$$\begin{aligned}
I_X &= \int_{-\infty}^{\infty} \left[\text{rect}\left(\frac{x + \frac{\lambda z_i f_X}{2}}{w}\right) \text{rect}\left(\frac{x - \frac{\lambda z_i f_X}{2}}{w}\right) \right] \exp\left[j \frac{16\pi W_m z_i}{w^2} f_X x\right] dx \\
&= \int_{-\frac{w}{2}\left(1 - \frac{|f_X|}{2f_o}\right)}^{\frac{w}{2}\left(1 - \frac{|f_X|}{2f_o}\right)} \exp\left[j \frac{16\pi W_m z_i}{w^2} f_X x\right] dx,
\end{aligned}$$

(7-43)

where again $f_o = w/2\lambda z_i f_o = \frac{w/2}{\lambda z_i}$. The integral with respect to y has the same form. The result, after normalization and simplification, is

$$\mathcal{H}(f_X, f_Y) = \Lambda\left(\frac{f_X}{2f_o}\right)\Lambda\left(\frac{f_Y}{2f_o}\right)$$

$$\times \text{sinc}\left[8 \frac{W_m}{\lambda} \left(\frac{f_X}{2f_o}\right) \left(1 - \frac{|f_X|}{2f_o}\right)\right] \text{sinc}\left[8 \frac{W_m}{\lambda} \left(\frac{f_Y}{2f_o}\right) \left(1 - \frac{|f_Y|}{2f_o}\right)\right].$$

(7-44)

Plots of this OTF are shown in [Fig. 7.11](#) for various values of W_m . Note that the diffraction-limited OTF is indeed obtained when $W_m = 0$. Note also that, for values of W_m greater than $\lambda/2$, sign reversals of the OTF occur. These reversals of contrast can readily be observed if the “spoke” target of [Fig. 7.12\(a\)](#) is used as the object. The “local spatial frequency” of this target changes slowly, increasing as the radius from the center is decreased. The local contrast of fringes is thus an indication of the value of the MTF at various frequencies. The position of the fringes is determined by the phase associated with the OTF at each frequency.

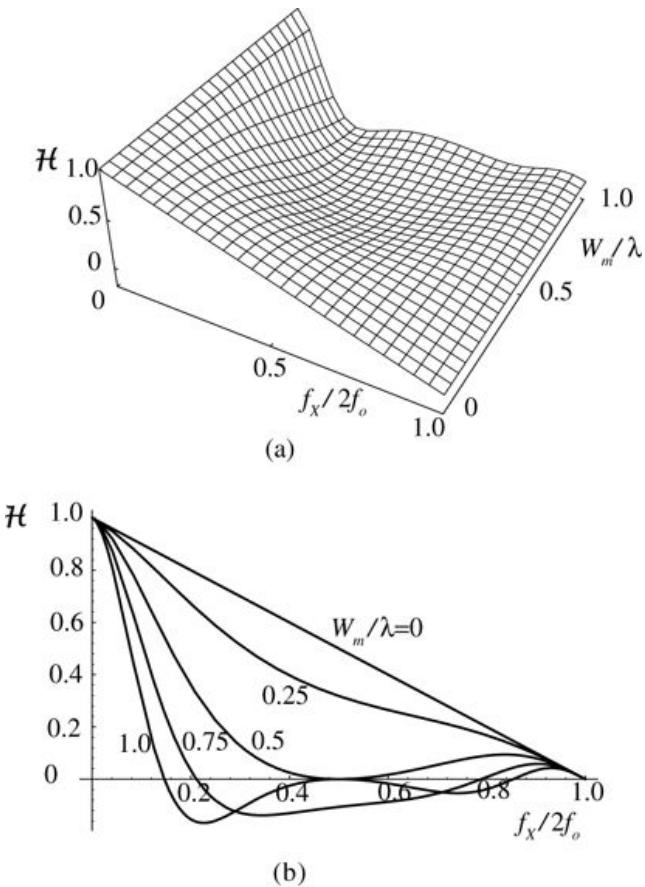
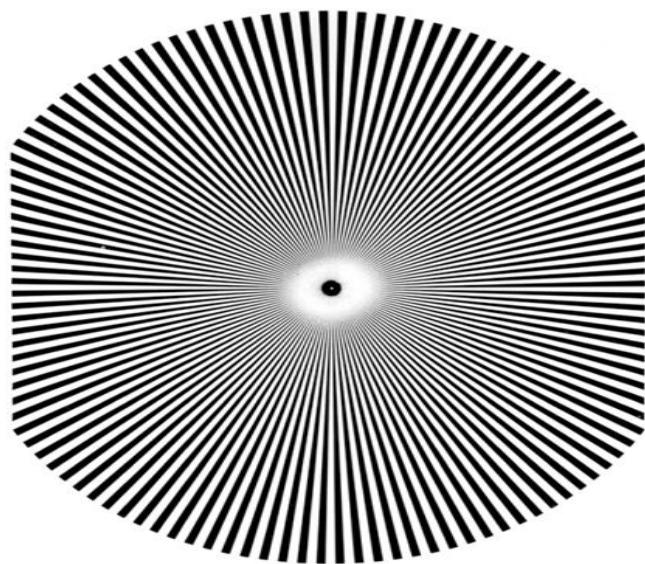


Figure 7.11

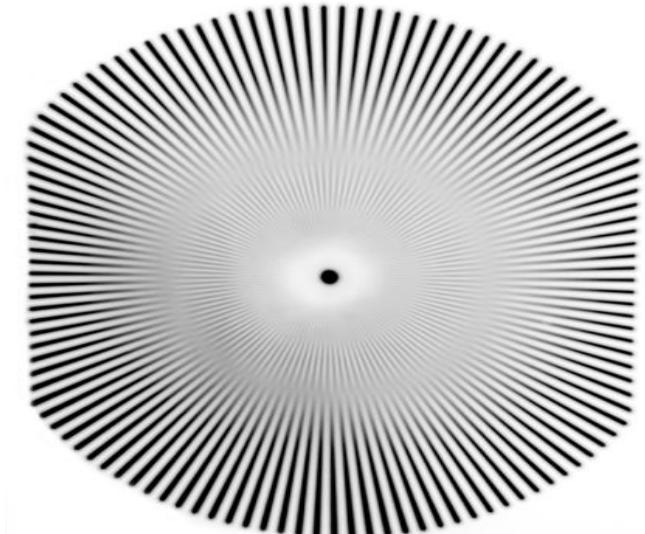
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 7.11 OTF for a focusing error in a system with a square pupil. (a) Three-dimensional plot with $f_X/2f_o$ along one axis and W_m/λ along the other axis. (b) Cross section along the f_X axis with W_m/λ as a parameter. Note that only when $W_m/\lambda > 0.5$ does the OTF go negative over a certain frequency range.

The illustration shows a 3 dimensional projection on a square horizontal plane whose adjoining sides are axes marked from 0 to +1; the axes represent $f_{\text{subscript } X} / 2f_{\text{subscript } o}$ and $W_{\text{subscript } m} / \lambda$. The vertical axis H is marked from 0 to 1. The projection resembles two mountain slopes at a right angle with a common valley. A side view of the projection is a right angled triangle whose perpendicular sides are the H axis and the side representing $f_{\text{subscript } X} / 2f_{\text{subscript } o}$. The graph shows horizontal axis $f_{\text{subscript } X} / 2f_{\text{subscript } o}$ and vertical axis H , both plotting values from 0 to 1. Four lines connect the 1 mark on the vertical axis to the 1 mark on the horizontal axis. The one that is a straight line represents the $W_{\text{subscript } m} / \lambda$ value of 0. Below the straight line are the U shaped curves that plot from right to left $W_{\text{subscript } m} / \lambda$ values of 0.25, 0.5, 0.75, and 1.0. The curve for 0.5 touches the horizontal axis and rises to the right in a gentle slope. The curves for 0.75 and 1.0 lunge a little below the horizontal axis and then rise to the right.



(a)



(b)

Figure 7.12
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 7.12 (a) Focused and (b) misfocused images of a spoke target.

Images a and b both are a series of tightly packed identical and uniformly spaced spokes radiating from the outside to a central point marked by a dot. In image a, the dot is distinct with a ring of fuzzy appearance around it. In image b, the ring of fuzzy appearance covers almost half the space.

When the system is out of focus, a gradual attenuation of contrast and a number of contrast reversals are obtained for increasing spatial frequency, as illustrated in [Fig. 7.12\(b\)](#).

Finally, consider the form of the OTF when the focusing error is very severe (that is, when $W_m \gg \lambda$). In such a case, the frequency response drops towards zero for relatively small

values of $f_X/2f_o$ and $f_Y/2f_o$. We may therefore write

$$1 - |f_X|/2f_o \approx 11 - |f_Y|/2f_o \approx 1,$$

$$1 - \frac{|f_X|}{2f_o} \approx 11 - \frac{|f_Y|}{2f_o} \approx 1,$$

and the OTF reduces to

$$\mathcal{H}(f_X, f_Y) \approx \text{sinc}\left[8\frac{W_m}{\lambda}\left(\frac{f_X}{2f_o}\right)\right] \text{sinc}\left[8\frac{W_m}{\lambda}\left(\frac{f_Y}{2f_o}\right)\right].$$

(7-45)

The interested reader can verify that this is precisely the OTF predicted by geometrical optics. Geometrical optics predicts a point-spread function that is the geometrical projection of the exit pupil onto the image plane, and therefore the point-spread function should be uniformly bright over a square and zero elsewhere (see Fig. 7.13). The Fourier transform of such a spread function yields the OTF of (7-45). More generally, *when aberrations of any kind are severe, the geometrical optics predictions of the intensity point-spread function may be Fourier-transformed to yield a good approximation to the OTF of the system.* The fundamental reason for this behavior lies in the fact that, when severe aberrations are present, the point-spread function is determined primarily by geometrical-optics effects, and diffraction plays a negligible role in determining its shape.

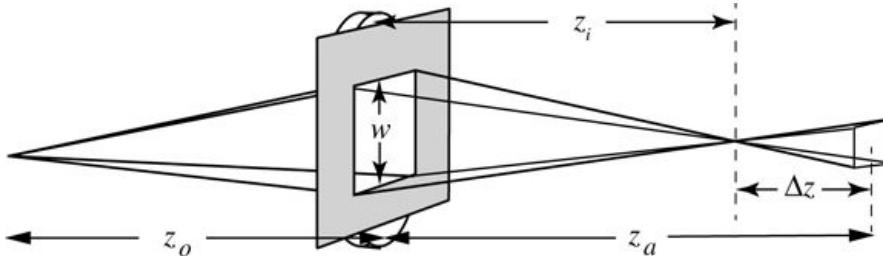


Figure 7.13

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 7.13 Geometrical optics prediction of the point-spread function of a system having a square pupil function and a severe focusing error.

The illustration shows rays from a point on the object plane on the left diverging and entering a rectangular entrance pupil of height w located at a distance of z_o to the right. The rays emerge on the other side at converge at a point on the image plane located at a distance of z_i . The rays, on their straight path, diverge and travel a distance of Δz to fall on plane that is at a distance z_a from the exit pupil.

7.4.5 Apodization and Its Effects on Frequency Response

The point-spread function of a diffraction-limited imaging system generally has side-lobes or side-rings of noticeable strength. While such extraneous responses may be of little concern in many imaging problems, they are of concern in a certain class of situations, such as when we wish to

resolve a weak point-source next to a stronger point-source. Such a problem is of considerable importance in astronomy, where the presence or absence of dim planets next to a bright star may often be of interest (see [Section 8.3](#)).

In an attempt to reduce the strength of side-lobes or side-rings, methods known as *apodization* have been developed. The word *apodize* is taken from the Greek language, and literally means “to remove the feet.” The “feet” being referred to are in fact the side-lobes and side-rings of the diffraction-limited impulse response. Similar techniques are well known in the field of digital signal processing, where they are known by the term *windowing* (see, for example, [96], [Section 3.3](#)).

Generally speaking, apodization amounts to the introduction of attenuation in the exit pupil of an imaging system, attenuation that may be insignificant at the center of the pupil but increases with distance away from the center. Thus it amounts to a “softening” of the edges of the aperture through the introduction of an attenuating mask. Remembering that diffraction by an abrupt aperture can be thought of as coming from edge waves originating around the rim of the aperture, a softening of the edge has the effect of spreading the origin of these diffracted waves over a broader area around the edges of the pupil, thereby suppressing ringing effects caused by edge waves with a highly localized origin. [Figure 7.14\(a\)](#) shows a plot of the unapodized and apodized intensity transmissions through a square pupil with and without a Gaussian intensity apodization

that falls to $(1/e)^2$ at the edge of the aperture. Part (b) of the figure shows cross sections of the intensity point-spread functions for the two cases. The logarithm of intensity is plotted vertically in order to emphasize the side-lobes, and the intensity normalization is proportional to the total integrated intensity passed by the pupil in each case. Note that the side-lobes have been significantly suppressed by the apodization. Also note that the width of the main lobe is increased somewhat by apodization, and that the maximum intensity is also reduced due to extra absorption in the pupil.

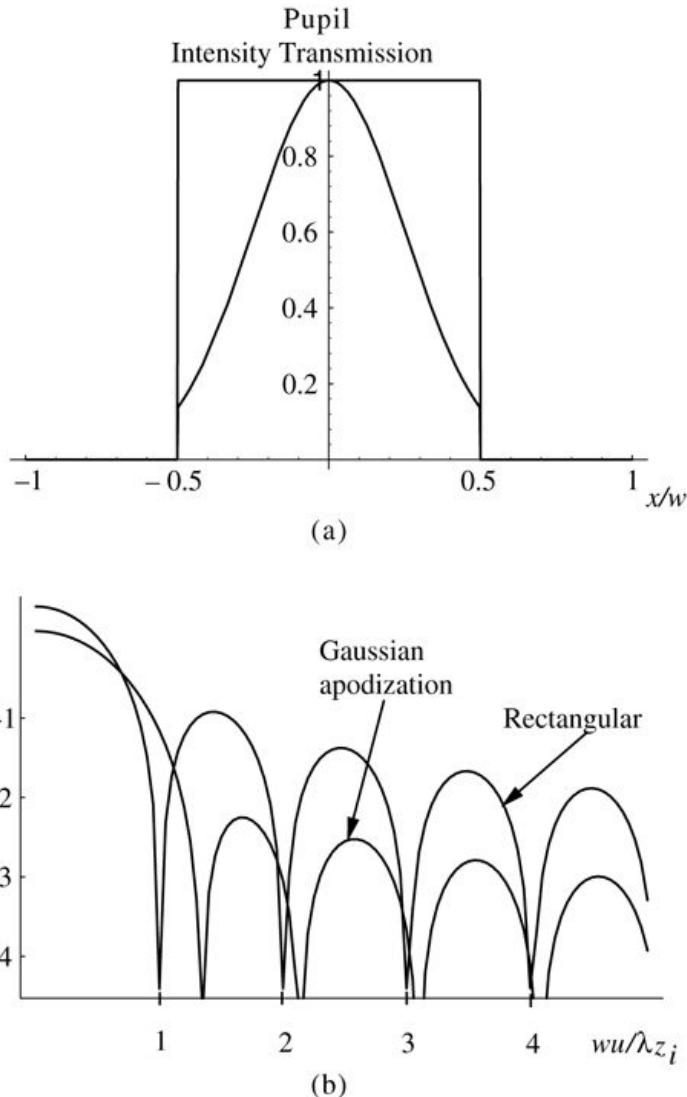


Figure 7.14
Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 7.14 Apodization of a rectangular aperture by a Gaussian function. (a) Intensity transmissions with and without apodization. (b) Point-spread functions with and without apodization.

Graph a plots a horizontal axis from minus 1 to +1 and a vertical axis from 0 to 1. The graph is a series of three horizontal lines and two vertical lines that connect the following points in the given order: (minus 1, 0), (minus 0.5, 0), (minus 0.5, 1), (0.5, 1), (0.5, 0), and (1, 0). A bell shaped symmetric curve extends from (minus 0.5, 0.15) to (0, 1) to (0.5, 0.15). Graph b plots $w u / \lambda z_i$ along the horizontal axis marked from 1 to 5 and $\log(I/I_o)$ along the vertical axis marked approximately from 0 to minus 4.5. The curve labeled Rectangular begins with half an arch beginning at the top left corner near (0, 0) and arching down to (1, minus 4.5). Thereafter, there is a continuous series of arches ending and then beginning the next at the points marked 2, 3, and 4 on the horizontal axis, progressively growing smaller as we move to the right. The curve labeled Gaussian apodization also begins with half an arch beginning at the top left corner near (0, 0.5) and arching down to (1.4, minus 4.5). Thereafter, there is a discontinuous series of arches ending at points that are to the right of 2, 3, and 4 on the horizontal

axis, progressively growing smaller as we move to the right. Every subsequent arch begins at a point slightly to the right of where the previous one ends.

The effects of apodization on the frequency response of both coherent and incoherent imaging systems are also of interest. In the coherent case the answer is straightforward due to the direct correspondence between the pupil and the amplitude transfer function. Attenuation that increases with distance from the center of the pupil results in an amplitude transfer function that falls off more rapidly with increasing frequency than it would in the absence of apodization. In the incoherent case, the less direct relationship between the OTF and the pupil makes the effects more subtle. [Figure 7.15](#) shows a plot of cross sections of the apodized and unapodized OTFs of a system with a rectangular pupil, where the apodization is of the Gaussian form described above. As can be seen, the effect of the apodization has been to boost the relative importance of midrange and low frequencies, while diminishing the strength of high frequencies.

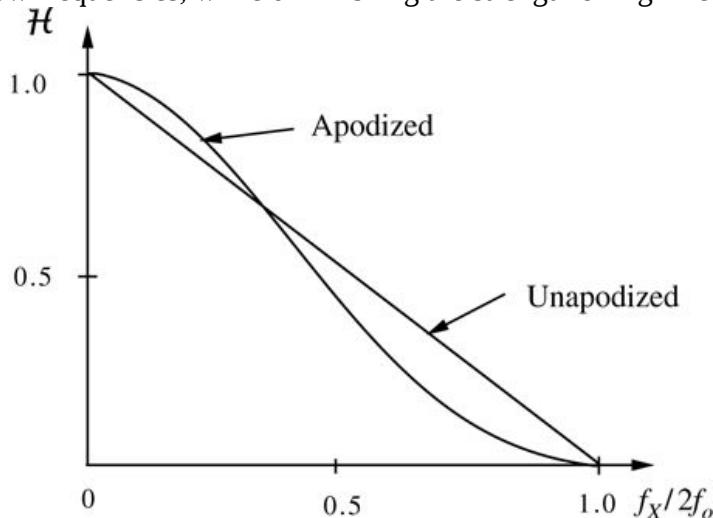


Figure 7.15

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 7.15 Optical transfer functions with and without a Gaussian apodization.

The graph plots $f_X/2f_o$ values along the horizontal axis and H values along the vertical axis. The line labeled Unapodized runs straight from $(0, 1)$ to $(1, 0)$. The curve labeled Apodized begins at $(0, 1)$ and curves slightly above the Unapodized straight line, which it intersects around $(0.35, 0.7)$ and then curves below it to meet at $(1, 0)$.

While the term *apodization* originally meant a tapering of the transmittance through the pupil near its edges in order to suppress side-lobes of the point-spread function, over time the term has come to be used to describe *any* introduction of absorption into the pupil, whether it lowers or raises the side-lobes. Perhaps a better term for weightings that increase the sidelobes of the point-spread function would be “inverse” apodization. [Figure 7.16](#) shows the amplitude transmittance through the pupil with and without a triangular amplitude weighting that gives extra emphasis to portions of the pupil near the edges, and de-emphasizes the importance of the center of the pupil. Also shown are cross sections of the OTF with and without this weighting. Note that this type of weighting emphasizes the importance of high frequencies relative to low frequencies.

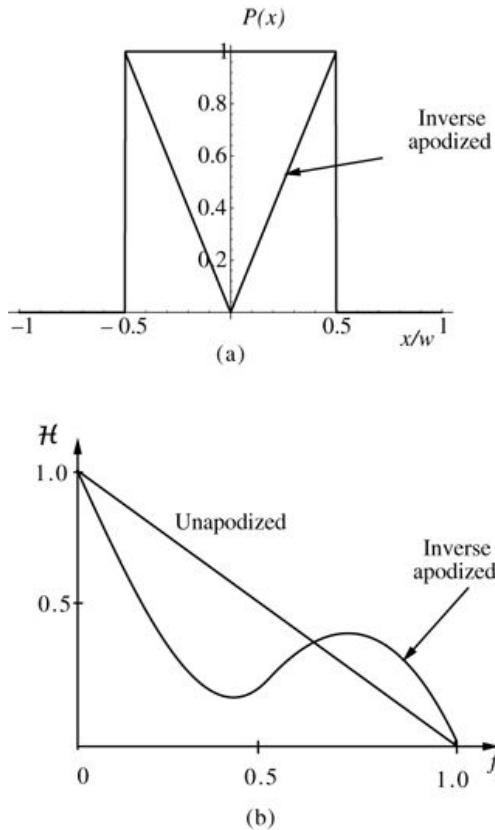


Figure 7.16
Goodman, Introduction to Fourier Optics, 4e,
 © 2017 W. H. Freeman and Company

Figure 7.16 Pupil amplitude transmittance and the corresponding OTF with and without a particular “inverse” apodization.

Graph a plots a horizontal axis from minus 1 to +1 and a vertical axis from 0 to 1. The graph is a series of three horizontal lines and two vertical lines that connect the following points in the given order: (minus 1, 0), (minus 0.5, 0), (minus 0.5, 1), (0.5, 1), (0.5, 0), and (1, 0). A V shaped symmetric curve, labeled Inverse apodized, extends from (minus 0.5, 1) to (0, 0) to (0.5, 1). Graph b plots $f_X / 2f_o$ values along the horizontal axis and H values along the vertical axis. The line labeled Unapodized runs straight from (0, 1) to (1, 0). The curve labeled Inverse apodized begins at (0, 1) and moves downward in a steep slope up to a point near (0.45, 0.7). Thereafter, it turns upward to cross the unapodized straight line at a point near (0.65, 0.35). The line then curves to join at (1, 0).

As a closing remark regarding this subject, note that while the OTF of a system with or without apodization always has the value unity at the origin, nonetheless it is *not* true that the amount of light transmitted to the image is the same in the two cases. Naturally the introduction of absorbing material in the pupil diminishes the light that reaches the image, but the normalization of the OTF suppresses this fact. Note also that, unlike the case of aberrations, inverse apodization *can* raise the value of the OTF at certain frequencies, as compared with its unapodized values.

See [Section 8.3.2](#) for more discussion of apodization.

7.5 Comparison of Coherent and Incoherent Imaging

As seen in previous sections, the OTF of a diffraction-limited system extends to a frequency that is twice the cutoff frequency of the amplitude transfer function. It is tempting, therefore, to conclude that incoherent illumination will invariably yield “better” resolution than coherent illumination, given that the same imaging system is used in both cases. As we shall now see, this conclusion is in general *not* a valid one; a comparison of the two types of illumination is far more complex than such a superficial examination would suggest.

A major flaw in the above argument lies in the direct comparison of the cutoff frequencies in the two cases. Actually, the two are not directly comparable, since the cutoff of the amplitude transfer function determines the maximum frequency component of the image *amplitude* while the cutoff of the optical transfer function determines the maximum frequency component of image *intensity*. Surely any direct comparison of the two systems must be in terms of the same observable quantity, image intensity.

Even when the quantity to be compared is agreed upon, the comparison remains a difficult one for an additional fundamental reason: the term *better* has not been defined. Thus we have no universal quality criterion upon which to base our conclusions. A number of potential criteria might be considered (e.g. the least-mean-square difference between the object and image intensities), but unfortunately the interaction of a human observer is so complex and so little understood that a truly meaningful criterion is difficult to specify.

In the absence of a meaningful quality criterion, we can only examine certain limited aspects of the two types of images, realizing that the comparisons so made will probably bear little direct relation to overall image quality. Nonetheless, such comparisons are highly instructive, for they point out certain fundamental differences between the two types of illumination.

7.5.1 Frequency Spectrum of the Image Intensity

One simple attribute of the image intensity which can be compared in the two cases is the *frequency spectrum*. Whereas the incoherent system is linear in intensity, the coherent system is highly nonlinear in that quantity. Thus some care must be used in finding the spectrum in the latter case.

In the incoherent case, the image intensity is given by the convolution equation

$$I_i = |h|^2 * I_g = |h|^2 * |U_g|^2.$$

$$I_i = |h|^2 * I_g = |h|^2 * |U_g|^2.$$

On the other hand, in the coherent case, we have

$$I_i = |h * U_g|^2.$$

$$I_i = |h * U_g|^2.$$

Let the symbol \star represent the autocorrelation integral

$$X(f_X, f_Y) \star X(f_X, f_Y) = \int_{-\infty}^{\infty} \int X(p, q) X^*(p - f_X, q - f_Y) dp dq.$$

$$X(f_X, f_Y) \star X(f_X, f_Y) = \int_{-\infty}^{\infty} \int X(p, q) X^*(p - f_X, q - f_Y) dp dq. \quad (7-46)$$

Then we can directly write the frequency spectra of the image intensities in the two cases as

$$\text{Incoherent: } \mathcal{F}[I_i] = H \star HG_g \star G_g \text{ Coherent: } \mathcal{F}[I_i] = HG_g \star HG_g,$$

$$\begin{aligned} \text{Incoherent: } \mathcal{F}[I_i] &= (H \star H)(G_g \star G_g) \\ \text{Coherent: } \mathcal{F}[I_i] &= (HG_g) \star (HG_g), \end{aligned} \quad (7-47)$$

where G_g is the spectrum of U_g and H is the amplitude transfer function.

The general result (7-47) does not lead to the conclusion that one type of illumination is better than the other in terms of image frequency content. It does, however, illustrate that the frequency content can be quite different in the two cases, and furthermore it shows that the results of any such comparison will depend strongly on both the intensity and *phase* distributions across the object.

To emphasize this latter point, we now consider two objects with the *same* intensity transmittance but different phase distributions, one of which can be said to be imaged better in coherent light and the other better in incoherent light. For simplicity, we suppose that the magnification of the system is unity, so that we may work in either the object or the image space at will without introducing a normalizing factor. Let the intensity transmittance of the object in both cases be

$$t(\xi, \eta) = \cos 2\pi f \xi$$

$$\tau(\xi, \eta) = \cos^2 2\pi \tilde{f} \xi$$

where to make our point we will assume that

$$f_o < \tilde{f} < f_o,$$

$$\frac{f_o}{2} < \tilde{f} < f_o,$$

f_o being the cutoff frequency of the amplitude transfer function. The amplitude transmittances of the two objects are taken to be

$$A: t_A(\xi, \eta) = \cos 2\pi f \xi \quad B: t_B(\xi, \eta) = |\cos 2\pi f \xi|.$$

$$\begin{aligned} A: \quad t_A(\xi, \eta) &= \cos 2\pi \tilde{f} \xi \\ B: \quad t_B(\xi, \eta) &= |\cos 2\pi \tilde{f} \xi|. \end{aligned}$$

Thus the two objects differ only by a periodic phase distribution.

[Figure 7.17](#) illustrates the various frequency-domain operations that lead to the image spectrum for object A A . In all cases the imaging system is assumed to be diffraction-limited. Note that the contrast of the image intensity distribution is *poorer* for the incoherent case than for the coherent case. Thus object A A is imaged better in coherent light than in incoherent light.

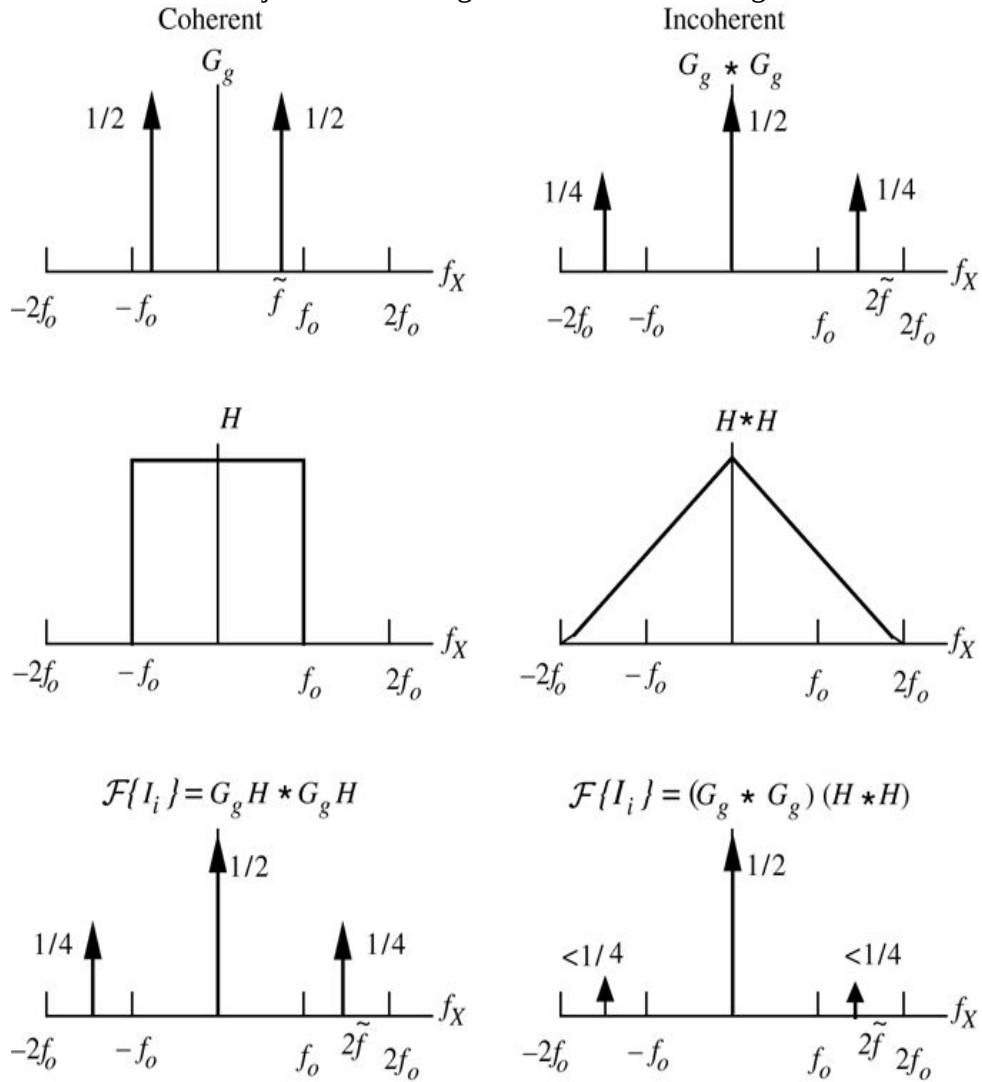


Figure 7.17

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 7.17 Calculation of the spectrum of the image intensity for object A A .

In all the graphs the horizontal axis plots values minus $2 f$ subscript o to $+ 2 f$ subscript o. In the first pair, the coherent case plots G subscript g along the vertical axis. Two upward pointing arrows labeled one half are perpendicular on the horizontal axis, one between the origin and f subscript o, at fundamental frequency f tilde, and the other is between the origin and minus f subscript o. In the incoherent case G subscript g asterisk G subscript g is plotted along the vertical axis. Two upward pointing arrows labeled one fourth are perpendicular on the horizontal axis, one between f subscript o and $2 f$ subscript o, at fundamental frequency $2f$ tilde, and the other is

between minus f subscript o and minus $2f$ subscript o . A third and taller arrow labeled one half is perpendicular at the origin. In the second pair, the coherent case plots H along the vertical axis. Two perpendiculars of equal length are dropped at f subscript o and minus f subscript o . Their top extremes are connected by a horizontal straight line. In the incoherent case, H asterisk H is plotted along the vertical axis. From a point on the vertical axis, two sloping straight lines are plotted, one to $2f$ subscript o and the other to minus $2f$ subscript o . In the third pair, the coherent case plots along the vertical axis $F \{ I \text{ subscript } i \} = G \text{ subscript } g \times H \text{ asterisk } G \text{ subscript } g \times H$. Two upward pointing arrows labeled one fourth are perpendicular on the horizontal axis, one between f subscript o and $2f$ subscript o , at fundamental frequency $2f$ tilde, and the other is between minus f subscript o and minus $2f$ subscript o . A third and taller arrow labeled one half is perpendicular at the origin.

In the incoherent case plots along the vertical axis $F \{ I \text{ subscript } i \} = (G \text{ subscript } g \times H \text{ asterisk } G \text{ subscript } g) (H \text{ asterisk } H)$. Two upward pointing arrows labeled one fourth are perpendicular on the horizontal axis, one between f subscript o and $2f$ subscript o , at fundamental frequency $2f$ tilde, and the other is between minus f subscript o and minus $2f$ subscript o . A third and taller arrow labeled one half is perpendicular at the origin.

The corresponding comparison for object B^B requires less detail. The object amplitude distribution is now periodic with fundamental frequency $2\tilde{f}$. But since $2\tilde{f} > f_o$, no variations of image intensity will be present for the coherent case, while the incoherent system will form the same image it did for object A^A . Thus for object B^B , incoherent illumination must be termed *better* than coherent illumination.

In summary, then, which particular type of illumination is better from the point of view of image spectral content depends very strongly on the detailed structure of the object, and in particular on its phase distribution. It is *not* possible to conclude that one type of illumination is preferred in all cases. The comparison is in general a complex one, although simple cases, such as the one illustrated above, do exist. For a second example, the reader is referred to [Prob. 7-10](#).

7.5.2 Two-Point Resolution

A second possible comparison criterion rests on the ability of the respective systems to resolve two closely spaced point sources. The two-point resolution criterion has long been used as a quality factor for optical systems, particularly in astronomical applications where it has a very real practical significance.

According to the so-called *Rayleigh criterion* of resolution, two incoherent point sources are “barely resolved” by a diffraction-limited system with a circular pupil of diameter w^W when the center of the Airy intensity pattern generated by one point source falls exactly on the first zero of the Airy pattern generated by the second. The minimum resolvable separation of the geometrical images is therefore

$$\delta x = 1.22 \lambda z_i / w = 1.22 \lambda F^{\#},$$

$$\delta x = 1.22 \lambda z_i / w = 1.22 \lambda F^{\#},$$

(7-48)

where $F^{\#} F^{\#}$ represents the F-number of the system, defined by

$$F\# = z_i/w.$$

$$F^\# = z_i / w.$$

(7-49)

The corresponding result in the nonparaxial case when the image plane is immersed in a material with index of refraction n can be shown to be

$$\delta x = 1.22 \lambda / 2n \sin \theta = 1.22 \lambda / 2NA$$

$$\delta x = 1.22 \frac{\lambda}{2n \sin \theta} = 1.22 \frac{\lambda}{2NA}$$

(7-50)

where θ represents the half-angle subtended by the exit pupil when viewed from the image plane, and NA is the *numerical aperture* of the optical system, defined by $NA = n \sin \theta$. If the exit pupil is square of width w instead of circular, the above results hold provided 1.22 is replaced by unity.

[Figure 7.18](#) illustrates the intensity distribution in the image of two equally bright incoherent point sources separated by the Rayleigh resolution distance. The central dip is found to fall about 27% below peak intensity.

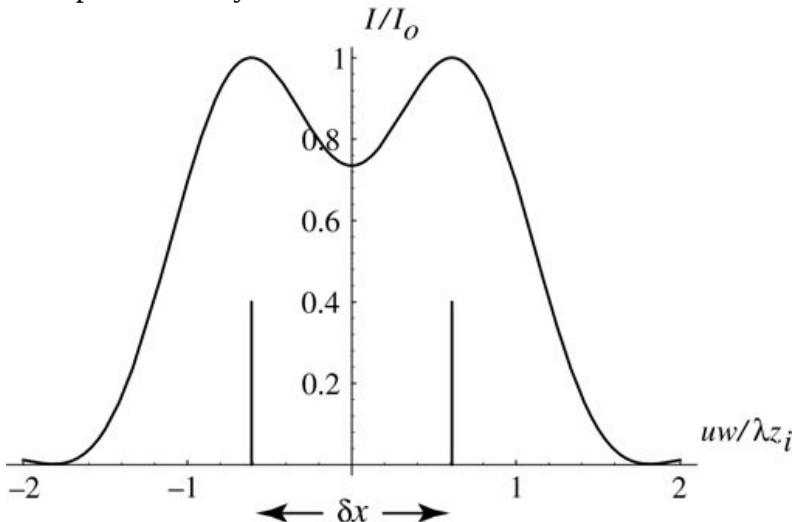


Figure 7.18

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 7.18 Image intensity for two equally bright incoherent point sources separated by the Rayleigh resolution distance. Circular aperture assumed. The vertical lines show the locations of the two sources.

The graph plots $uw/\lambda z_i$ along the horizontal axis marked from minus 2 to +2, and I/I_o along the vertical axis marked 0 to 1. Two perpendiculars are dropped at minus 0.6 and + 0.6; the distance between them is marked δx . The curve rises steeply from (minus 2,

0) and reaches (minus 0.6, 1) and then slopes downward to reach around (0, 0.74). The curve s path thus far is reflected on to the other side across the vertical axis.

We can now ask whether the two point-source objects, separated by the same Rayleigh distance δ , would be easier or harder to resolve with coherent illumination than with incoherent illumination. This question is academic for astronomical objects, but is quite relevant in microscopy, where the illumination is usually closer to coherent than incoherent, and where in some cases it is possible to control the coherence of the illumination.

As in the previous examples, the answer to this question is found to depend on the *phase distribution* associated with the object. A cross section of the image intensity can be directly written, in normalized image coordinates, as

$$I(x) = 2J_1[\pi(x-0.61)]\pi(x-0.61) + e^{j\phi} 2J_1[\pi(x+0.61)]\pi(x+0.61)$$

$$I(x) = \left| 2 \frac{J_1[\pi(x - 0.61)]}{\pi(x - 0.61)} + e^{j\phi} 2 \frac{J_1[\pi(x + 0.61)]}{\pi(x + 0.61)} \right|^2$$

where ϕ is the relative phase between the two point sources. [Figure 7.19](#) shows the distributions of image intensity for point sources in phase ($\phi=0$ radians), in quadrature ($\phi=\pi/2$ radians), and in phase opposition ($\phi=\pi$ radians). When the sources are in quadrature, the image intensity distribution is identical to that resulting from incoherent point sources. When the sources are in phase, the dip in the image intensity is absent, and therefore the two points are not as well resolved as with incoherent illumination. Finally, when the two objects are in phase opposition, the dip falls all the way to zero intensity (a 100% dip) at the point midway between the locations of the two points, so the two points must be said to be better resolved with coherent illumination than with incoherent illumination. Thus there can again be no generalization as to which type of illumination is preferred for two-point resolution.

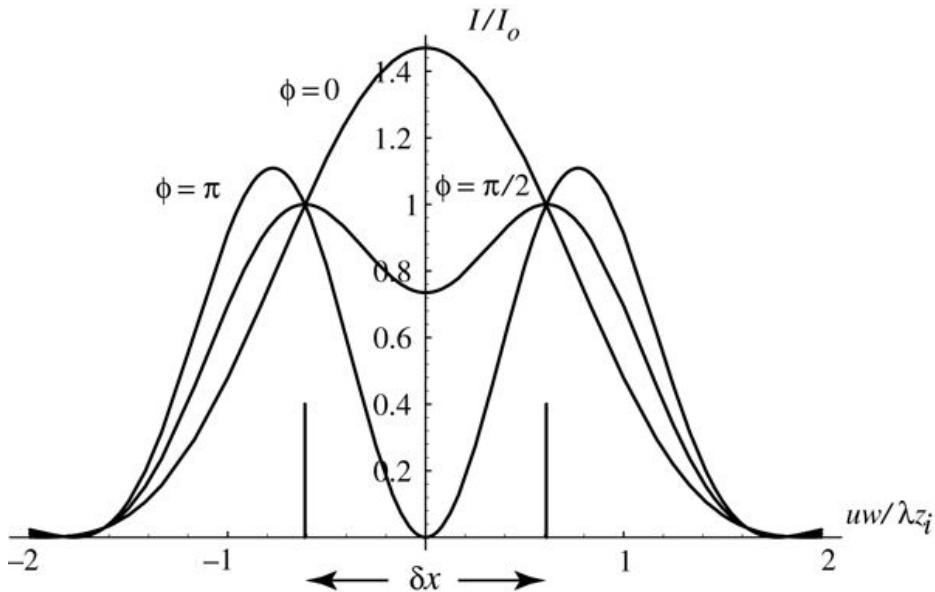


Figure 7.19

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 7.19 Image intensities for two equally bright coherent point sources separated by the Rayleigh resolution distance, with the phase difference between the two sources as a parameter. Circular aperture assumed. The vertical lines show the locations of the two point sources.

The graph plots $uw/\lambda z_i$ along the horizontal axis marked from minus 2 to +2, and I/I_o along the vertical axis marked 0 to 1. Two perpendiculars are dropped at minus 0.6 and + 0.6; the distance between them is marked δx . There are three curves, all symmetric about the vertical axis. The first curve rises steeply from (minus 2, 0) and reaches (minus 0.6, 1) and then slopes downward to reach around (0, 0.74) on the vertical axis. The curve's path thus far is reflected across the vertical axis. The second curve for $\phi = 0$ rises steeply up to (0, 1.48) and is reflected across the vertical axis. The third curve for $\phi = \pi$ rises steeply up to (minus 0.8, minus 1.12) and then turns downward to reach the origin and extend rightward reflecting the path thus far across the vertical axis.

7.5.3 Other Effects

There are certain other miscellaneous properties of images formed with coherent light that should be mentioned in any comparison with incoherent images [77]. First, the responses of incoherent and coherent systems to sharp edges are notably different. [Figure 7.20](#) shows the theoretical responses of a system with a circular pupil to a step function object, i.e. an object with amplitude transmittance

$$t_A(\xi, \eta) = \begin{cases} 0 & \xi < 0 \\ 1 & \xi \geq 0. \end{cases}$$

$$t_A(\xi, \eta) = \begin{cases} 0 & \xi < 0 \\ 1 & \xi \geq 0. \end{cases}$$

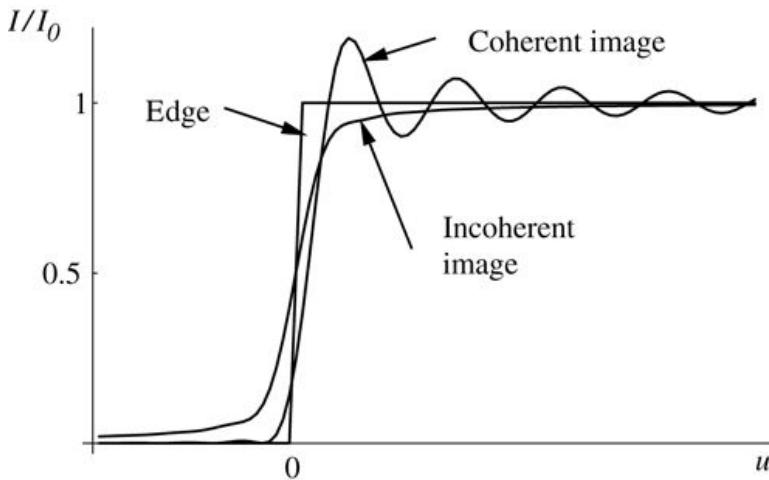


Figure 7.20

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 7.20 Images of a step in coherent and incoherent light. Circular pupil assumed.

The graph plots u along the horizontal axis and I/I_0 along the vertical axis located to the left of 0 on the horizontal axis.

In the graph, the edge is plotted as a horizontal line extending along the horizontal axis up to 0 and then rising up almost perpendicularly up to a point near $(0, 1)$ and then running horizontally to the right.

The curve for coherent image overlaps the edge line but begins an upward slope a little before reaching 0 on the horizontal axis, extending upward beyond the horizontal line of the edge curve and then sloping down to below the horizontal line and then rising again to form a wave overlapping the horizontal of the edge curve. The incoherent curve begins only slightly above where the edge curve begins. After a short rightward run, it begins its upward slope before reaching 0 on the horizontal axis, intersects the near vertical line of the edge curve and then, before reaching the edge horizontal line, turns right and progressively gets closer to it.

[Figure 7.21](#) shows actual photographs of the image of an edge in the two cases. The coherent system is seen to exhibit rather pronounced “ringing.” This property is analogous to the ringing that occurs in video amplifier circuits with transfer functions that fall too abruptly with frequency. The coherent system has a transfer function with sharp discontinuities, while the falloff of the OTF is much more gradual. Another important property of the coherent image is that it crosses the location of the actual edge with only 1/4 of its asymptotic value of intensity, whereas the incoherent image crosses with a value of 1/2 of its asymptotic value. If we were to assume that the actual location of the edge is at the position where the intensity reaches half its asymptotic value, we would arrive at a correct estimate of the position of the edge in the incoherent case, but in the coherent case we would err in the direction of the bright side of the edge. This fact can be important, for example, in estimating the widths of lines on integrated circuit masks.

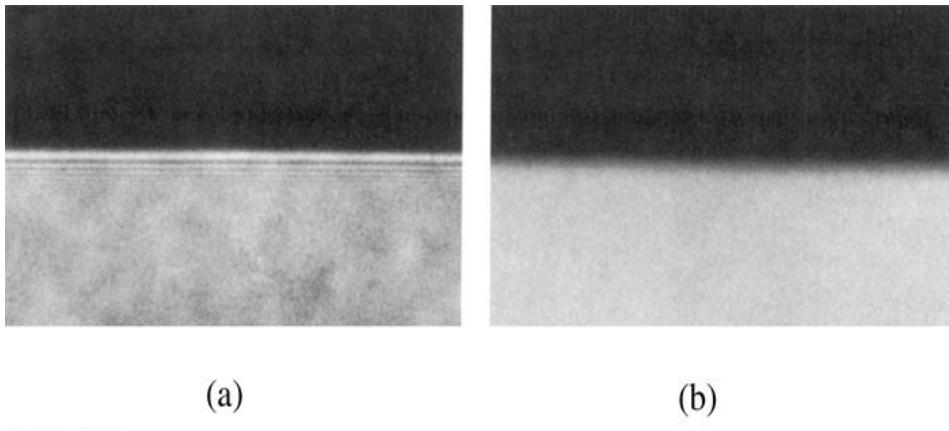


Figure 7.21
Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 7.21 Photographs of the image of an edge in (a) coherent and (b) incoherent illumination. [From [\[77\]](#). Copyright 1966 by the Optical Society of America, Inc., reprinted with permission.]

Both photos show a dark horizontal band in the top half and a gray horizontal band in the bottom half. Photo a is clearer with thin gray strips running across the length between the two bands. Photo b is fuzzy, with both bands almost uniformly shaded.

In addition, we must mention the so-called *speckle effect* that is readily observed with highly coherent illumination. While we shall consider this effect in the context of optical imaging, it has also proven to be a problem in certain other nonoptical imaging modalities, such as microwave side-looking radar and medical ultrasound imaging. [Figure 7.22](#) shows photographs of a transparency object, illuminated through a diffuser (e.g. a piece of ground glass), taken in coherent light and incoherent light. The granular nature of the coherent image is a direct consequence of the complex, random perturbation of the wavefront introduced by the diffuser, together with the coherence of the light. For background on the speckle effect, see, for example, [\[268\]](#), [\[132\]](#), and [\[83\]](#). The granularity in the image arises from interference between closely spaced and randomly phased scatterers within the diffuser. The size of the individual *speckles* can be shown [\[318\]](#) to be roughly the size of a *resolution cell* on the image. In the case of incoherent illumination, such interference cannot take place, and speckle is missing from the image. Thus when a particular object of interest is near the resolution limit of an optical system, the speckle effect can be quite bothersome if coherent light is used. Much of this problem can be eliminated by moving the diffuser during the observation, with the result that the coherence of the illumination is at least partially destroyed and the speckles “wash out” during the measurement process. Unfortunately, as we will see in a later chapter, motion of the diffuser is not possible in conventional holography, which by its very nature is almost always a coherent imaging process, so speckle remains a particular problem in holographic imaging. The subject is discussed further in that context in [Section 11.10.4](#).

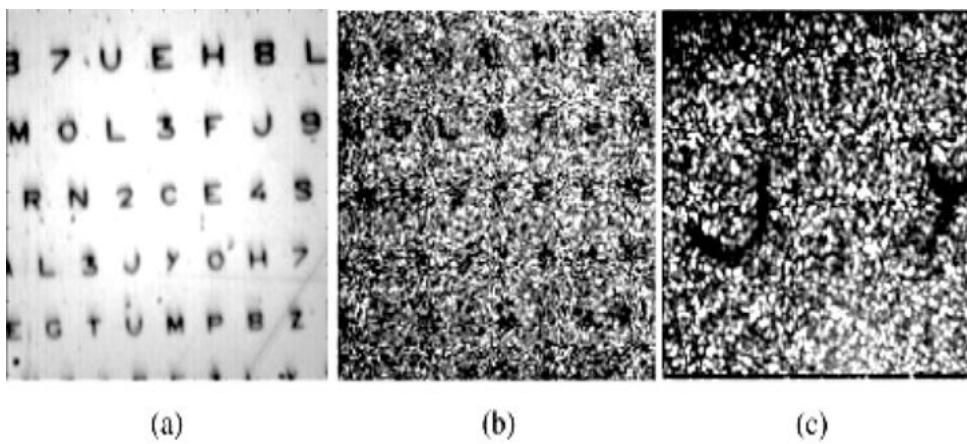


Figure 7.22
Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 7.22 Images illustrating the speckle effect. The object is a transparency illuminated through a diffuser. (a) Image in incoherent light. (b) Image in coherent light. (c) Close-up image of a particular letter in coherent light. [Photo courtesy of P. Chavel and T. Avignon, Institut d'Optique.]

Photo a shows rows of letters of the English alphabet with an occasional digit in place of a letter. The text is dark and set on a very bright white background. Photo b shows the same transparency but the background is dark and very granular with the letters barely legible. Photo c is an extreme close-up of the transparency. It shows just the letters J and Y set in background much more granular than that in photo b. Photo a shows rows of letters of the English alphabet with an occasional digit in place of a letter. The text is dark and set on a very bright white background. Photo b shows the same transparency but the background is dark and very granular with the letters barely legible. Photo c is an extreme close-up of the transparency. It shows just the letters J and Y set in background much more granular than that in photo b. Photo a shows rows of letters of the English alphabet with an occasional digit in place of a letter. The text is dark and set on a very bright white background. Photo b shows the same transparency but the background is dark and very granular with the letters barely legible. Photo c is an extreme close-up of the transparency. It shows just the letters J and Y set in background much more granular than that in photo b. Photo a shows rows of letters of the English alphabet with an occasional digit in place of a letter. The text is dark and set on a very bright white background. Photo b shows the same transparency but the background is dark and very granular with the letters barely legible. Photo c is an extreme close-up of the transparency. It shows just the letters J and Y set in background much more granular than that in photo b.

Finally, highly coherent illumination is particularly sensitive to optical imperfections that may exist along a path to the observer. For example, tiny dust particles on a lens may lead to very pronounced diffraction patterns that will be superimposed on the image. One fundamental reason for the importance of such effects in coherent imaging is the so-called “interference gain” that occurs when a weak undesired signal interferes with a strong desired signal (see [Prob. 7-17](#)).

A reasonable conclusion from the above discussion would be that one should choose incoherent illumination whenever possible, to avoid the artifacts associated with coherent illumination. However, there are many situations in which incoherent illumination simply can not be realized or can not be used for a fundamental reason. These situations include high-resolution microscopy, coherent optical information processing, and holography.

7.6 Confocal Microscopy

The confocal microscope is an imaging system that, by virtue of its geometry, is capable of resolving details that are finer than possible with a conventional microscope using the same objective lens. The invention of the confocal microscope is usually attributed to Marvin Minsky, who filed a patent application on the idea in 1957 [253]. Detailed studies of this imaging modality were carried out by T. Wilson and C. Sheppard in the late 1970s and beyond (see, for example, [374]). For detailed discussions of confocal microscopes, see [78] and [250]. The confocal microscope is a scanning microscope, and can be constructed in either a reflection or a transmission mode. The geometry of a reflection microscope is illustrated in Fig. 7.23. Note that the same objective lens is used for both illumination and imaging. Scanning can be accomplished either by moving the sample or by deflecting the illumination and detection beams. Light, usually supplied by a laser, is incident on the object and reflected, then separated from the incident light by a beam splitter, and finally detected after it has passed through a small pinhole. Thus the detector records the light intensity transmitted through the pinhole for each focal point on the object.

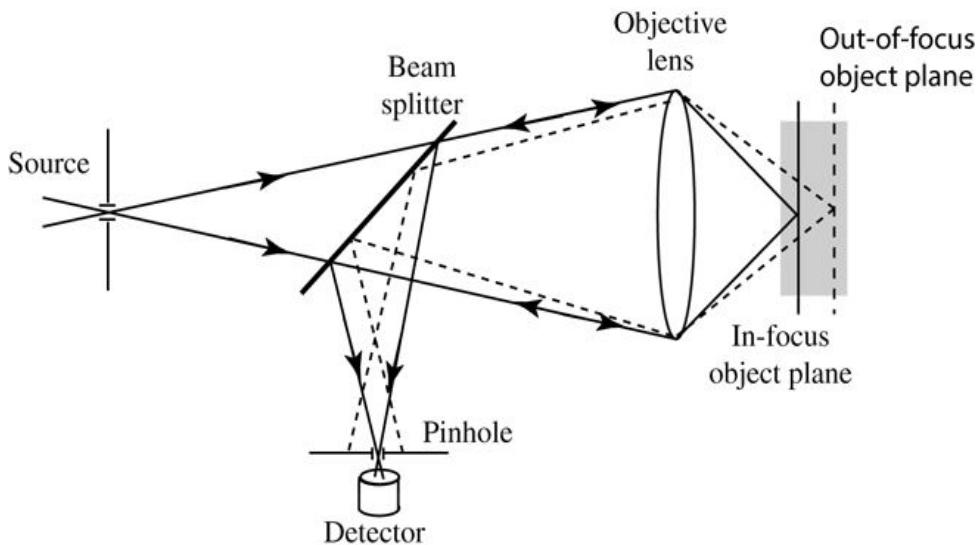


Figure 7.23

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 7.23 Confocal reflection microscope geometry. The solid lines represent rays incident on and reflected from an in-focus object point. The dashed lines represent the rays reflected from an out-of-focus object point.

The illustration shows a Source in the left extreme. Through an aperture at its center rays enter and move rightward, their path marked by an upward sloping ray reaching the top end of a biconvex lens and a downward sloping ray reaching the lower end of the lens. An upward sloping thick line representing a beam splitter lies between the source and the lens. From the two points where the splitter intersects the rays to the lens, rays move downward to pass through a pinhole and reach a detector. The rays between the splitter and the lens are marked with both rightward and leftward

arrowheads. Rays from the lens move rightward to converge on a vertical line labeled “In-focus object plane.”

A pair of dotted lines from the top and bottom ends of the lens marks the convergence of rays from the lens to a dotted vertical line representing out-of-focus object plane located to the right of the in-focus object plane. Dotted lines from either ends of the lens extend up to points on the splitter away from where rays from the source intersect it. The dotted rays then go downward and then intersect well before reaching opposite sides of the pinhole.

The detailed behavior of this microscope depends on whether the object volume simply reflects or backscatters the incident coherent light, in which case the system is coherent, or whether the illumination is used to excite a fluorescent sample, in which case the light collected by the imaging system is incoherent. Consider the coherent case first, then the incoherent case.

7.6.1 Coherent Case

In both the coherent and incoherent cases, the resolution of the confocal microscope differs from that of a conventional microscope with the same objective lens by virtue of the fact that both the sample illumination and the imaging portions of the system provide spatial discrimination. The first effect is due to illumination: diffraction by the finite aperture of the objective lens results in a diffraction-limited amplitude point-spread function illuminating the object point of interest. The second effect is caused by the detector pinhole: arguing from reciprocity, if it is sufficiently small, the detector pinhole creates a weighting function on the object that, in the ideal limit of an infinitesimal pinhole, is an amplitude distribution that is identical with the amplitude distribution of the incident beam. Of course, in practice the pinhole can not be infinitesimal, since it must pass light to the detector, so we continue to distinguish the illumination and detector pinhole effects. Viewed from the detector, the *effective* amplitude pattern on the object is the product of the amplitude of the illumination pattern and the amplitude weighting function due to the detector pinhole. Thus if the transverse coordinates in the object space are (x, y) and the scanning spot is centered on coordinates (x_0, y_0) , the detected intensity must be

$$I(x_0, y_0) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h_i(x - x_0, y - y_0) h_d(x - x_0, y - y_0) r(x, y) dx dy,$$

$$I(x_0, y_0) = \left| \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h_i(x - x_0, y - y_0) h_d(x - x_0, y - y_0) r(x, y) dx dy \right|^2, \quad (7-51)$$

where $h_i(x, y)$ represents the amplitude of the illumination pattern, $h_d(x, y)$ represents the amplitude weighting function on the object caused by the detector pinhole, and $r(x, y)$ represents the amplitude reflectivity of the object in the plane of focus. Thus we can regard the imaging system as a coherent system that is linear and invariant in amplitude, producing a convolution with amplitude impulse response $h_{coh}(x, y)$ given by

$$h_{coh}(x, y) = h_i(x, y) h_d(x, y).$$

$$h_{coh}(x, y) = h_i(x, y) h_d(x, y).$$

(7-52)

A product in the space domain implies a convolution in the frequency domain, and therefore we conclude that the amplitude transfer function $H_{coh}(f_X, f_Y)$ of the coherent confocal microscope must be

$$H_{coh}(f_X, f_Y) = H_i(f_X, f_Y) * H_d(f_X, f_Y),$$

$$H_i(f_X, f_Y) = H_i(f_X, f_Y) * H_d(f_X, f_Y),$$

(7-53)

where $H_i(f_X, f_Y)$ is the Fourier transform of $h_i(x, y)$ and $H_d(f_X, f_Y)$ is the Fourier transform of $h_d(x, y)$. As usual, the asterisk represents convolution. From [Eq. \(7-21\)](#), we know that for a circular objective pupil of diameter w , the Fourier transform of h_i is a circle with cutoff ρ_o ,

$$H_i(\rho) = \text{circ}\left(\frac{\rho}{\rho_o}\right),$$

(7-54)

where $\rho_o = w / (2\lambda z_i)$ is the frequency cutoff radius, and z_i is the image distance. For a circular pinhole with a diameter that is sufficiently small, we have an identical expression for $H_d(f_X, f_Y)$, and we see that the amplitude transfer function of the coherent confocal microscope is identical with the OTF of a conventional microscope using the same objective lens but with an *incoherent* object. If for convenience we normalize this amplitude transfer function to be unity at the origin, we then have

$$H_{coh}(\rho) = \frac{2}{\pi} \left[\arccos\left(\frac{\rho}{2\rho_o}\right) - \left(\frac{\rho}{2\rho_o}\right) \sqrt{1 - \left(\frac{\rho}{2\rho_o}\right)^2} \right]$$

(7-55)

for $\rho < 2\rho_o$ and zero otherwise.

7.6.2 Incoherent Case

If the object is naturally fluorescent, or its molecules have been labeled with fluorophores, the incident light intensity stimulates the object to emit incoherent light, generally at a different wavelength than that of the illumination. In such a case, the imaging equation becomes

$$I(x_0, y_0) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h_i(x-x_0, y-y_0) 2h_d(x-x_0, y-y_0) 2R(x, y) dx dy,$$

$$I(x_0, y_0) = \int_{-\infty}^{\infty} \int |h_i(x - x_0, y - y_0)|^2 |h_d(x - x_0, y - y_0)|^2 R(x, y) dx dy, \quad (7-56)$$

where $R(x, y) = |r(x, y)|^2$ is the intensity reflectivity of the sample. Again the system is linear and invariant, but this time linearity is with respect to *intensity*. The point spread function of the incoherent system is given by

$$h_{\text{inc}}(x, y) = h_{\text{coh}}(x, y) = |h_i(x, y)|^2 |h_d(x, y)|^2.$$

$$h_{\text{inc}}(x, y) = |h_{\text{coh}}(x, y)|^2 = |h_i(x, y)|^2 |h_d(x, y)|^2. \quad (7-57)$$

If we again assume a circular objective pupil and a circular detector pinhole of sufficiently small size, we can use (7-55) for H_{coh} , and the result will be a properly normalized autocorrelation of two such functions. Thus the OTF of this system is the properly normalized autocorrelation function of the amplitude transfer function H_{coh} ,

$$\mathcal{H}_{\text{inc}}(f_X, f_Y) = H_{\text{coh}}(f_X, f_Y) \star H_{\text{coh}}(f_X, f_Y),$$

$$\mathcal{H}_{\text{inc}}(f_X, f_Y) = H_{\text{coh}}(f_X, f_Y) \star H_{\text{coh}}(f_X, f_Y), \quad (7-58)$$

where the \star represents autocorrelation. The autocorrelation can be performed numerically, with the results shown in Fig. 7.24, valid for incoherent emission. The figure shows the OTFs for both a conventional microscope and a confocal microscope. As can be seen from the figure, while the absolute cutoff of the OTF of the confocal microscope is twice that of the conventional microscope, in practice the extension of bandwidth is of the order of 1.5 or 1.6.

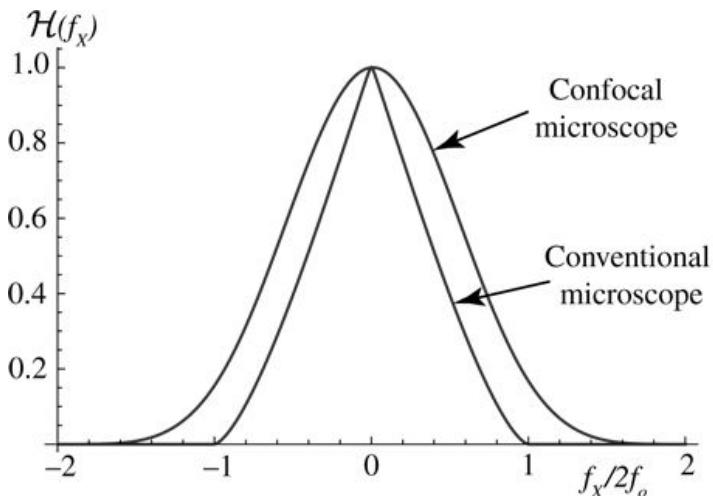


Figure 7.24

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 7.24 Optical transfer functions for a conventional microscope and a confocal microscope, assuming incoherent emission.

The graph plots $f_x / 2f_o$ along the horizontal axis marked from minus 2 to 2 and $H(f_x)$ along the vertical axis marked from 0 to 1. There are two curves, both symmetrical. The curve for conventional microscope runs from minus 2 to minus 1 on the horizontal axis and then rises in a steep upward slope to $(0, 1)$. After a sharp U turn, it slopes down to the 1 mark on the horizontal axis and extends till +2.

The curve for confocal microscope is bell shaped and wraps around the conventional microscope graph. It runs from minus 2 to minus 1.5 on the horizontal axis and then rises in a steep upward slope to $(0, 1)$. After a smooth U turn, it slopes down to the 1.5 mark on the horizontal axis and extends till +2.

7.6.3 Optical Sectioning

While the confocal microscope has a spatial frequency response that extends beyond that of a conventional microscope with the same objective lens, the most important property of this system is its enhanced ability to section an image in depth. For an object that extends in depth, a conventional microscope produces an in-focus image of the portion of the object that lies within the depth of focus of the objective lens, and out-of-focus images of planes that lie beyond the depth of focus. For a high-NA objective lens, the depth of focus is very shallow, so methods for removing the out-of-focus portions of the image are of great interest. The dashed rays in Fig. 7.23 illustrate the ray paths for a point-scatterer in an out-of-focus plane. As can be seen, for an out-of-focus plane that lies behind the in-focus plane, the image of the scatterer appears before the pinhole, and is spread substantially by the time the light reaches the pinhole. A similar effect occurs for object points in front of the depth of focus. Thus the effect of the pinhole is to reduce or eliminate much of the out-of-focus light, providing a sectioning capability not present with a conventional microscope. This sectioning capability holds for both the coherent case and the incoherent case, and is the primary reason for the great popularity of the confocal microscope.

Methods for scanning more than one point at a time and using a detector array rather than a single detector have been developed [282] [78], but will not be covered here.

Problems - Chapter 7

1. 7-1. The mask shown in Fig.P7.1 is inserted in the exit pupil of an imaging system. Light from the small openings interferes to form a fringe in the image plane.
1. Find the spatial frequency of this fringe in terms of the center-to-center spacing s of the two openings, the wavelength λ , and the image distance z_i .
 2. The openings are circular with diameter d . Specify the envelope of the fringe pattern caused by the finite openings in the pupil plane.

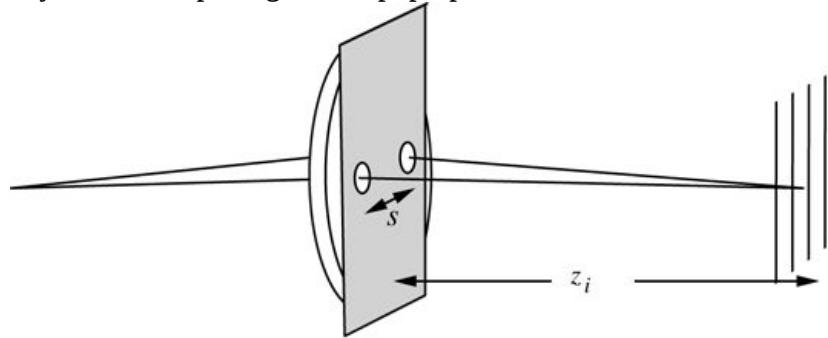


Figure P7.1

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure P7.1

The illustration shows rays starting at a point on the left extreme diverging and falling on an image system where a vertical mask is placed. The mask has two circular apertures of equal size and on the same horizontal level. The distance between them is s . Rays exiting from the apertures converge at an image plane represented by four vertical parallel lines on the right extreme. The image plane is at a distance of z subscript i from the mask.

2. 7-2. The *line-spread function* of a two-dimensional imaging system is defined to be the response of that system to a one-dimensional delta function passing through the origin of the input plane.

1. In the case of a line excitation lying along the x axis, show that the line-spread function l and the point-spread function p are related by

$$l(y) = \int_{-\infty}^{\infty} p(x, y) dx,$$

$$l(y) = \int_{-\infty}^{\infty} p(x, y) dx,$$

where l and p are to be interpreted as amplitudes or intensities, depending on whether the system is coherent or incoherent, respectively.

2. Show that for a line source oriented along the x axis, the (one-dimensional) Fourier transform of the line-spread function is equal to a slice through the (two-dimensional)

Fourier transform of the point-spread function, the slice being along the $f_Y f_Y$ axis. In other words, if $\mathcal{F}\{l\} = L$ and $\mathcal{F}\{p\} = P$, then $L(f) = P(0, f)$.

3. Find the relationship between the line-spread function and the step response of the system, i.e. the response to a unit step excitation oriented parallel to the $x x$ axis.
3. 7-3. An incoherent imaging system has a square pupil function of width $w w$. A square stop of width $w/2 w/2$ is placed at the center of the pupil, as shown in [Fig. P7.3](#).
 1. Sketch cross sections of the optical transfer function with and without the stop present.
 2. Sketch the limiting form of the optical transfer function as the size of the stop approaches the size of the full pupil.

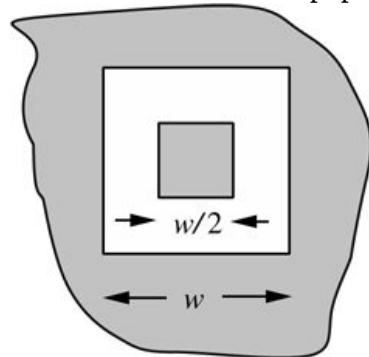


Figure P7.3
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

4. Figure P7.4 shows a circular pupil of diameter $w w$. A half-plane stop is inserted in the pupil, yielding the modified pupil shown in [Fig. P7.4](#). Find expressions for the optical transfer function evaluated along the $f_X f_X$ and $f_Y f_Y$ axes.

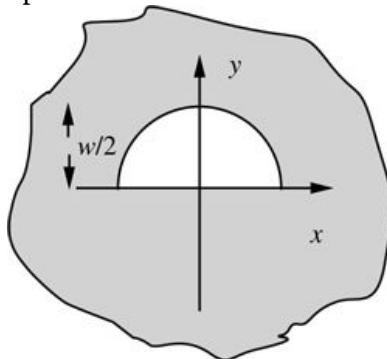


Figure P7.4
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure P7.4

5. 7-5. An incoherent imaging system has a pupil consisting of an equilateral triangle, as shown in [Fig. P7.5](#). Find the OTF of this system along the f_X and f_Y axes in the spatial



Figure P7.5
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

frequency domain.

- Figure P7.5**
6. 7-6. Sketch the f_X and f_Y cross sections of the optical transfer function of an incoherent imaging system having as a pupil function the aperture shown in [Fig. P7.6](#). Be sure to label the various cutoff frequencies and center frequencies on these sketches.

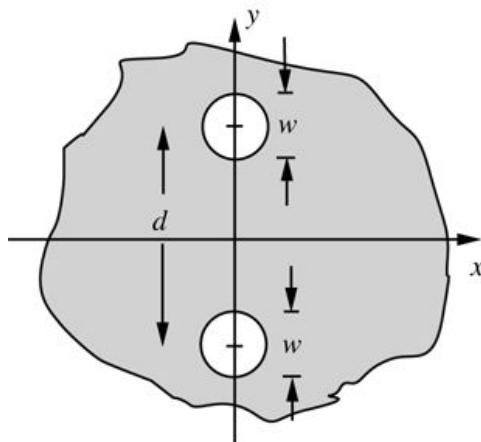


Figure P7.6
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

- Figure P7.6**
7. 7-7. Consider a *pinhole camera* shown in [Fig. P7.7](#).

Assume that the object is incoherent and nearly monochromatic, the distance z_o from the object is so large that it can be treated as infinite, and the pinhole is circular with diameter w

- Under the assumption that the pinhole is large enough to allow a purely geometrical-optics estimation of the point-spread function, find the optical transfer function of this camera. If we define the “cutoff frequency” of the camera to be the frequency where the first zero of the OTF occurs, what is the cutoff frequency under the above geometrical-

optics approximation? (Hint: First find the intensity point-spread function, then Fourier transform it. Remember the second approximation above.)

2. Again calculate the cutoff frequency, but this time assuming that the pinhole is so small that Fraunhofer diffraction by the pinhole governs the shape of the point-spread function.
3. Considering the two expressions for the cutoff frequency that you have found, can you estimate the “optimum” size of the pinhole in terms of the various parameters of the system? Optimum in this case means the size that produces the highest possible cutoff frequency.

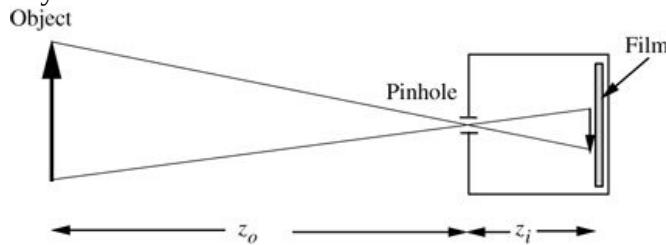


Figure P7.7

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure P7.7

On the left is an upright object represented by an upward pointing perpendicular arrow. On the right is a square cross section of a pinhole camera. At the center of side facing the object is a pinhole and on the opposite side inside the square is a film. The horizontal distance of the pinhole from the object is z_o and that between the pinhole and the film is z_i . An image less than half the length of the object is represented by a downward pointing arrow at the center of the film.

8. 7-8. Consider the OTF of (7-45), as predicted for a system having square pupil and a focusing error. It is hypothesized that the point-spread function of this system is the convolution of the diffraction-limited point-spread function with the point-spread function predicted by geometrical optics. Examine the validity of this claim.
9. 7-9. A quantity of considerable utility in determining the seriousness of the aberrations of an optical system is the *Strehl definition DS* D_s , which is defined as the ratio of the light intensity at the maximum of the point-spread function of the system with aberrations to that same maximum for that system in the absence of aberrations. (Both maxima are assumed to exist on the optical axis.) Prove that D_s is equal to the normalized volume under the optical transfer function of the aberrated imaging system; that is, prove

$$D_s = \frac{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathcal{H}(f_X, f_Y) \text{with } df_X df_Y}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathcal{H}(f_X, f_Y) \text{without } df_X df_Y},$$

$$D_s = \frac{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathcal{H}(f_X, f_Y) \text{with } df_X df_Y}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathcal{H}(f_X, f_Y) \text{without } df_X df_Y},$$

where the notations “with” and “without” refer to the presence or absence of aberrations, respectively.

10. 7-10. An object with a square-wave amplitude transmittance (shown in Fig. P7.10) is imaged by a lens with a circular pupil function. The focal length of the lens is 10 cm, the fundamental frequency of the square wave is 100 cycles/mm, the object distance is 20 cm, and the wavelength is $1 \mu\text{m}$. What is the minimum lens diameter that will yield *any variations* of intensity across the image plane for the cases of

1. Coherent object illumination?
2. Incoherent object illumination?

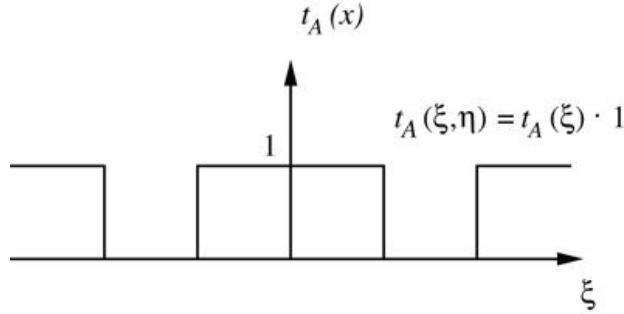


Figure P7.10
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure P7.10

11. 7-11. An object has an intensity transmittance given by

$$\tau(\xi, \eta) = 121 + \cos 2\pi f \xi$$

$$\tau(\xi, \eta) = \frac{1}{2}(1 + \cos 2\pi \tilde{f} \xi)$$

and introduces a constant, uniform phase delay across the object plane. This object is placed at distance $2f$ in front of a positive lens of focal length f , and the image is examined in a plane $2f$ behind the lens. Compare the maximum frequencies \tilde{f} transmitted by the system for the cases of coherent and incoherent illumination.

12. 7-12. A sinusoidal amplitude grating with transmittance

$$t_A(\xi, \eta) = 121 + \cos 2\pi f \xi$$

$$t_A(\xi, \eta) = \frac{1}{2}(1 + \cos 2\pi \tilde{f} \xi)$$

is placed in front of a thin, positive lens (circular with diameter w , focal length f) and obliquely illuminated by a monochromatic plane wave traveling at angle θ to the z axis in the (ξ, z) plane, as shown in Fig. P7.12.

1. What is the Fourier transform of the amplitude distribution transmitted by the object?

2. Assuming $z_i = z_o = 2f$, what is the maximum angle θ for which *any* variations of intensity will appear in the image plane?
3. Assuming that this maximum angle is used, what is the intensity in the image plane, and how does it compare with the corresponding intensity distribution for $\theta=0$?
4. Assuming that the maximum angle θ is used, what is the maximum grating frequency f that will yield variations of intensity in the image plane? How does this frequency compare with the cutoff frequency when $\theta=0$?

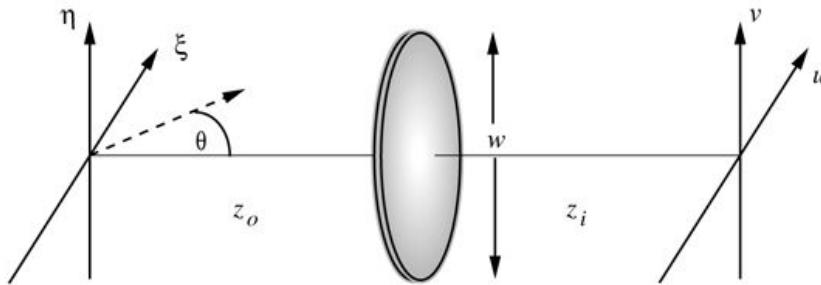


Figure P7.12
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure P7.12

The illustration shows horizontal axis z . On its left extreme is vertical axis η and third axis ξ . On the right extreme is vertical axis v and third axis u . A lens of diameter w is placed between the two planes such that it is at a distance of z subscript i from the axis v and z subscript o from axis η .

13. 7-13. The F -number of a lens with a circular aperture is defined as the ratio of the focal length to the lens diameter. Show that when the object distance is infinite, the cutoff

frequency for a coherent imaging system using this lens is given by $f_o = \frac{1}{2\lambda F \#}$, where $F \#$ represents the F-number.

14. 7-14. The *Sparrow resolution criterion* states that two equally strong incoherent point sources are barely resolved when their separation is the maximum separation for which the image of the pair of points shows no dip at the midpoint. This condition can be equivalently stated as one for which the curvature of the total intensity at the midpoint between the centers of the individual spread functions vanishes.

1. Show that, for a spread function that is an even function of u , such a condition occurs when the separation (in the u direction) between the centers of the spread functions is twice the value of u that satisfies the equation

$$\partial^2 h(u, 0) / \partial u^2 = 0$$

$$\frac{\partial^2 |h(u, 0)|^2}{\partial u^2} = 0$$

where $|h|^2$ is the intensity point-spread function of the system.

2. What is the value of the Sparrow separation (in the image space) for a system with a square aperture of width w^W , where an edge of the aperture runs parallel to the direction of separation of the two sources?
15. 7-15. Consider the step responses of two different imaging systems, one with a circular aperture of diameter w^W and the second with a square aperture of width w^W , with one edge of the aperture parallel with the edge of the step. All other aspects of the two systems are identical.
1. Show that, with coherent illumination, the step responses of the two systems are identical.
 2. Show that, with incoherent illumination, the step responses of the two systems are not identical.
 3. Describe how you would numerically calculate the step responses in both cases.
16. 7-16. Show that the intensity image of a step-object (edge along the η^H axis) formed by a coherent imaging system having a square pupil (width w^W) with edges parallel to and orthogonal to the direction of the step can be expressed as
- $$I_i(u, v) = c \left| \frac{\pi}{2} + \text{Si}\left(\frac{\pi w u}{\lambda z_i}\right) \right|^2$$
- where $\text{Si}(z)$ is known as the “sine integral” and is defined by
- $$\text{Si}(z) = \int_0^z \frac{\sin t}{t} dt,$$
- and c^C is a constant. Note: The function $\text{Si}(z)$ is a tabulated function and is known to many mathematical software packages.
17. 7-17. Consider the addition of a strong desired field of amplitude A^A with a weak undesired field of amplitude a^a . You may assume that $A \gg a$.
1. Calculate the relative perturbation $\Delta I / |A|^2$ to the desired intensity caused by the presence of the undesired field when the two fields are mutually coherent.
 2. Repeat for the case of mutually incoherent fields.
18. 7-18. Using the definition of mutual intensity, show that any purely monochromatic wave is fully coherent spatially and therefore must be analyzed as a system that is linear in amplitude.

8 Point-Spread Function and Transfer Function Engineering

In the sections that follow in this chapter, we briefly discuss several imaging techniques that involve design of a point-spread function or a transfer function for a special purpose. In some cases the technique rests on designing a system with a unique PSF or MTF that has properties that allow it to yield improved images, often after computer manipulation of the image. In other cases, the imaging system is intentionally altered to gain object information not otherwise available. Finally, in the case of fluorescence microscopy, the object itself is modified in such a way that computer post-processing allows impressive improvements of resolution.

8.1 Cubic Phase Mask for Increased Depth of Field

The cubic phase mask represents one of several techniques that allow an imaging system to achieve greater depth of field than would otherwise be possible with the unmodified system. To understand benefits of such systems, it is first necessary to define the concepts of depth of focus and depth of field.

8.1.1 Depth of Focus

The depth of focus of an imaging system is defined as the distance an image plane can be moved about the plane of best focus while maintaining essentially unchanged resolution in the image space. The images of object planes within the depth of focus remain in focus, while those outside of the depth of focus become blurred. Thus depth of focus refers to the distance an image sensor can move axially in the image plane while maintaining full resolution in the image of an object that is at a fixed distance in front of the entrance pupil.

We can derive an expression for the depth of focus using a combination of wave optics and geometrical optics. The geometry is illustrated in [Fig. 8.1\(a\)](#). Here, the depth of focus, defined to encompass one side of the true image plane, is represented by Δz_i . As a criterion for defining “full resolution”, we require that the transverse image blur predicted by geometrical optics in the out-of-focus plane be identical with the image resolution predicted by diffraction in the in-focus plane. The point spread function predicted by geometrical optics in the out-of-focus plane is a projection of the exit pupil onto the image plane. Assuming the exit pupil to be square of width w , simple geometry shows that the width X of the projected exit pupil is given by

$$X = w \Delta z_i / z_i.$$

(8-1)

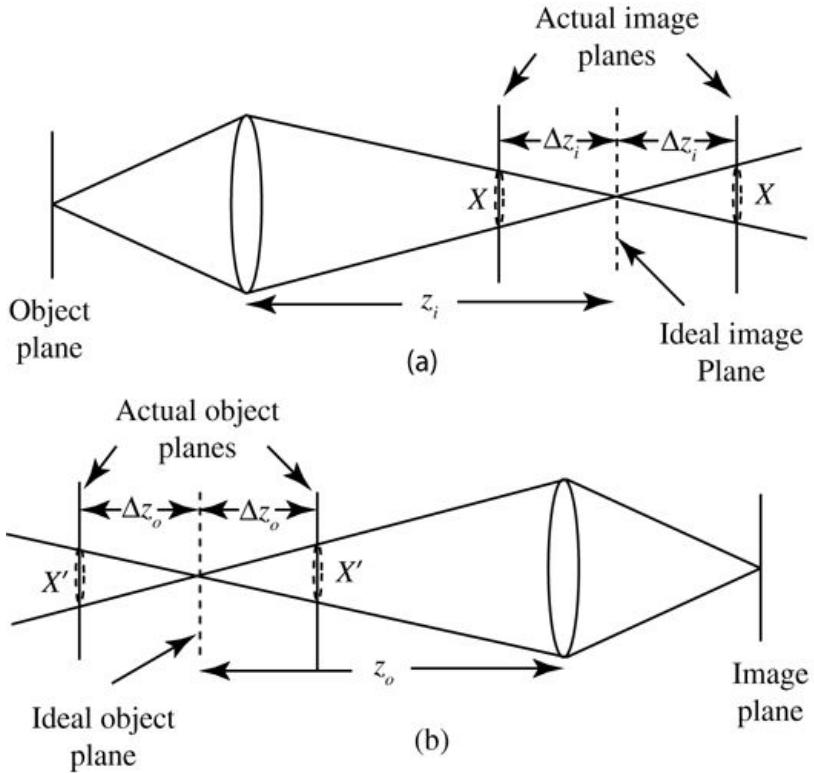


Figure 8.1

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 8.1 Geometries for calculating (a) depth of focus, (b) depth of field.

Illustration a shows an object plane on the left extreme. From its center rightward rays diverge to a biconvex lens, on the other side of which the rays converge to a point on the ideal image plane represented by a vertical dotted line. The rays pass through the plane and diverge again. The distance between the plane and the lens is z_i . On either side of the ideal image plane, at a distance of Δz_i from it, is an actual image plane. The part of the actual image plane that lies between the converging or diverging rays is a dotted outline of a slight bulge labeled X. Illustration b shows an image plane on the right extreme. From its center leftward rays diverge to a biconvex lens, on the other side of which the rays converge to a point on the ideal object plane represented by a vertical dotted line. The rays pass through the plane and diverge again. The distance between the plane and the lens is z_o . On either side of the ideal object plane, at a distance of Δz_o from it, is an actual object plane. The part of the actual object plane that lies between the converging or diverging rays is a dotted outline of a slight bulge labeled X dash.

For a square pupil of width w , we know that the transverse resolution, as defined by the Rayleigh criterion and limited by diffraction, is

$$\Delta x_i = \lambda F_i^{\#}.$$

$$\Delta x_i = \lambda F_i^{\#} .$$

(8-2)

Equating X with $\Delta x_i \Delta z_i$ and solving for Δz_i , we have

$$\Delta z_i = \lambda(F_i^{\#})^2.$$

$$\Delta z_i = \lambda(F_i^{\#})^2.$$

(8-3)

Note that for a high-resolution system, for which $F_i^{\#}$ is of the order of unity, the depth of focus becomes very small, of the order of a wavelength.

8.1.2 Depth of Field

The depth of field is defined to be the one-sided distance over which an object can be moved while maintaining essentially fixed resolution in a fixed image plane. The geometry illustrating the depth of field is shown in [Fig. 8.1\(b\)](#). The calculation of the depth of field is based on the lens law, the expression for depth of focus, and a small increment approximation. Let the lens law be written

$$1/z_o + \Delta z_o + 1/z_i + \Delta z_i + 1/f = 1/z_o(1 + \Delta z_o/z_o) + 1/z_i(1 + \Delta z_i/z_i) + 1/f \\ \approx 1/z_o(1 - \Delta z_o/z_o) + 1/z_i(1 - \Delta z_i/z_i) + 1/f = 0,$$

$$\frac{1}{z_o + \Delta z_o} + \frac{1}{z_i + \Delta z_i} + \frac{1}{f} \\ = \frac{1}{z_o(1 + \Delta z_o/z_o)} + \frac{1}{z_i(1 + \Delta z_i/z_i)} + \frac{1}{f} \\ \approx \frac{1}{z_o}(1 - \Delta z_o/z_o) + \frac{1}{z_i}(1 - \Delta z_i/z_i) + \frac{1}{f} = 0,$$

(8-4)

where we have assumed that $\Delta z_i \ll z_i$ and $\Delta z_o \ll z_o$. Using the lens equation once more, we can express Δz_o in terms of Δz_i as follows:

$$\Delta z_o = -z_o z_i / 2 \Delta z_i = -\Delta z_i |M|^2 = -\lambda(F_i^{\#})^2 |M|^2,$$

$$\Delta z_o = -\left(\frac{z_o}{z_i}\right)^2 \Delta z_i = -\frac{\Delta z_i}{|M|^2} = -\frac{\lambda(F_i^{\#})^2}{|M|^2},$$

(8-5)

where M is the magnification. Thus for a fixed image F -number, a system with a magnification larger than unity has a depth of field that is smaller than the depth of focus, while a system that has a magnification smaller than unity has a depth of field that is larger than the depth of focus.¹

8.1.3 The Cubic Phase Mask

It was originally shown by [Dowski and Cathey \[92\]](#) that the insertion of a cubic phase mask in the pupil of an incoherent optical system, combined with digital post-processing, can extend the depth of field of an imaging system significantly. For simplicity, we limit attention to a one-dimensional system. By a cubic phase mask we mean a transmitting object with amplitude transmittance

$$tA(x) = \exp(j2\pi W_{m3} \lambda x^3 / (w/2)^3),$$

(8-6)

where W_{m3} is the path length error at the edge of a rectangular pupil of width w . To understand the details of this technique, we must return to a general expression for the OTF when aberrations are present.

The OTF of an imaging system with both a focusing error and a cubic phase mask in the exit pupil can be written as (cf. [\(7-43\)](#))

$$\begin{aligned} \mathcal{H}(f_X) &= 1/w \int -|f_X|/2 to |f_X|/2 \exp[j\theta_2(f_X, x) + j\theta_3(f_X, x)] dx, \\ \mathcal{H}(f_X) &= \frac{1}{w} \int_{-\frac{|f_X|}{2}}^{\frac{|f_X|}{2}} \exp[j\theta_2(f_X, x) + j\theta_3(f_X, x)] dx, \end{aligned}$$

(8-7)

where

$$\theta_2(f_X, x) = 16\pi W_{m2} \lambda z_i f_X x / w^2$$

$$\theta_2(f_X, x) = 16\pi \frac{W_{m2} \lambda z_i f_X}{w^2} x = 16\pi \frac{W_{m2}}{\lambda} \left(\frac{f_X}{2f_o} \right) \left(\frac{x}{w} \right)$$

(8-8)

represents the effect of the focusing error alone, W_{m2} being the maximum path-length error at the edge of the pupil due to the focusing error, while

$$\begin{aligned} \theta_3(f_X, x) &= 16\pi \frac{W_{m3}}{w^3} \left[\left(x + \frac{\lambda z_i f_X}{2} \right)^3 - \left(x - \frac{\lambda z_i f_X}{2} \right)^3 \right] \\ &= 4\pi \frac{W_{m3}}{\lambda} \left(\frac{\lambda z_i f_X}{w} \right)^3 + 48\pi \frac{W_{m3}}{\lambda} \left(\frac{\lambda z_i f_X}{w^3} \right) x^2 \\ &= 4\pi \frac{W_{m3}}{\lambda} \left(\frac{f_X}{2f_o} \right)^3 + 48\pi \frac{W_{m3}}{\lambda} \left(\frac{f_X}{2f_o} \right) \left(\frac{x}{w} \right)^2 \end{aligned}$$

(8-9)

represents the effect of the cubic phase mask alone, where W_{m3} represents the maximum path-length error at the edge of the pupil due to the cubic phase mask. As usual, $fo = (w/2)/(\lambda z_i)$

$f_o = (w / 2) / (\lambda z_i)$. With these results we can write an expression for the optical transfer function when a cubic phase mask is inserted in the pupil and an arbitrary amount of defocus is present:

$$\begin{aligned} \mathcal{H}(f_X) &= \exp[j4\pi W_m 3\lambda(fX^2 f_0)^3] w^2 - w^2(1 - |fX|^2 f_0) w^2(1 - |fX|^2 f_0) \exp[j48\pi W_m 3\lambda(fX^2 f_0) \\ &\quad (xw)^2] \times \exp[j16\pi W_m 2\lambda(fX^2 f_0)(xw)] dx. \\ \mathcal{H}(f_X) &= \frac{\exp\left[j4\pi \frac{W_{m3}}{\lambda} \left(\frac{f_X}{2f_0}\right)^3\right]}{w} \int_{-\frac{w}{2}\left(1 - \frac{|f_X|}{2f_0}\right)}^{\frac{w}{2}\left(1 - \frac{|f_X|}{2f_0}\right)} \exp\left[j48\pi \frac{W_{m3}}{\lambda} \left(\frac{f_X}{2f_0}\right) \left(\frac{x}{w}\right)^2\right] \\ &\quad \times \exp\left[j16\pi \frac{W_{m2}}{\lambda} \left(\frac{f_X}{2f_0}\right) \left(\frac{x}{w}\right)\right] dx. \end{aligned} \quad (8-10)$$

This integral is a formidable one, but it can be solved by *Mathematica*. The resulting OTF is given by

$$\begin{aligned} \mathcal{H}(f_X) &= (-1)^{3/4} \exp\left[j8\pi\Delta^3 W_{m3} - j\frac{4\pi\Delta W_{m2}^2}{3W_{m3}}\right] \\ &\quad \times \left(\operatorname{erfi}\left[\frac{(1+j)\sqrt{2\pi\Delta/3(W_{m2}-3W_{m3}+3W_{m3}|\Delta|)}}{\sqrt{W_{m3}}}\right] \right. \\ &\quad \left. - \operatorname{erfi}\left[\frac{(1+j)\sqrt{2\pi\Delta/3(W_{m2}+3W_{m3}-3W_{m3}|\Delta|)}}{\sqrt{W_{m3}}}\right] \right), \end{aligned} \quad (8-11)$$

where $\Delta = fX/(2f_0)$, $\Delta = f_X / (2f_0)$ and $\operatorname{erfi}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt$.

$$\operatorname{erfi}(z) = \sqrt{\frac{2}{\pi}} \int_0^z \exp(-t^2) dt.$$

(8-12)

The imaginary error function can also be expressed in terms of Fresnel integrals,

$$\operatorname{erfi}(z) = (1-j)C(1+j)z\pi - jS(1+j)z\pi.$$

$$\operatorname{erfi}(z) = (1-j) \left(C\left[\frac{(1+j)z}{\sqrt{\pi}}\right] - jS\left[\frac{(1+j)z}{\sqrt{\pi}}\right] \right).$$

(8-13)

For an expression for $\mathcal{H}(f_X) \mathcal{H}(f_X)$ in terms of Fresnel integrals, see [325].

The MTF can now be stated as

$$|\mathcal{H}(f_X)| = 183\Delta W_m^3 \times \operatorname{erfi}(1+j)2\pi\Delta/3(W_m^2 - 3W_m^3 + 3W_m^3|\Delta|)W_m^3 - \operatorname{erfi}(1+j)2\pi\Delta/3(W_m^2 + 3W_m^3 - 3W_m^3|\Delta|)W_m^3.$$

$$\begin{aligned} |\mathcal{H}(f_X)| &= \frac{1}{8\sqrt{3\Delta W_m^3}} \\ &\times \left| \operatorname{erfi} \left[\frac{(1+j)\sqrt{2\pi\Delta/3}(W_m^2 - 3W_m^3 + 3W_m^3|\Delta|)}{\sqrt{W_m^3}} \right] \right. \\ &\quad \left. - \operatorname{erfi} \left[\frac{(1+j)\sqrt{2\pi\Delta/3}(W_m^2 + 3W_m^3 - 3W_m^3|\Delta|)}{\sqrt{W_m^3}} \right] \right|. \end{aligned}$$

(8-14)

It is possible to show (see [92]) that, in the mid-frequency range of the MTF, for W_m^3 sufficiently large the MTF can be approximated as

$$|\mathcal{H}(f_X)| \approx 112W_m^3\lambda|f_X|2f_0.$$

$$\left| \mathcal{H}(f_X) \right| \approx \frac{1}{\sqrt{12\frac{W_m^3}{\lambda}\left(\frac{|f_X|}{2f_0}\right)}}.$$

(8-15)

[Figure 8.2](#) illustrates the effect of introducing the cubic phase mask in a defocused system. These are plots of the *exact* MTFs for a variety of cases. See the caption for more details.

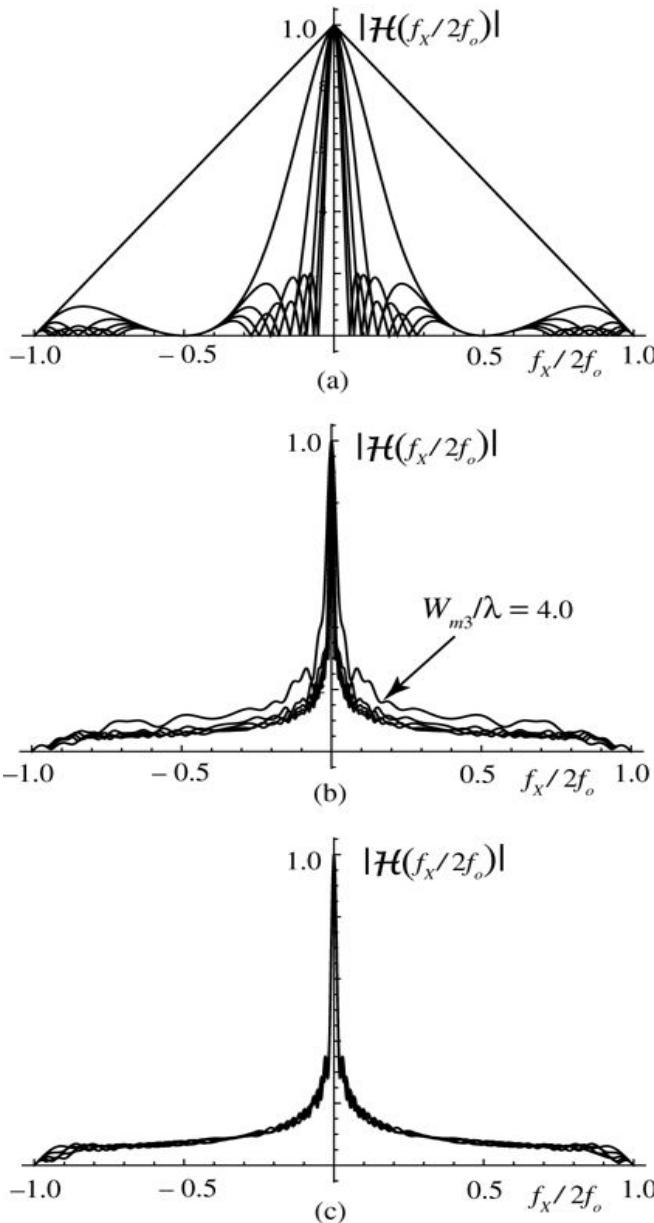


Figure 8.2
Goodman, Introduction to Fourier Optics, 4e,
 © 2017 W. H. Freeman and Company

Figure 8.2 Exact MTFs with various levels of defocus and a cubic phase aberration. (a) MTFs with no cubic phase mask, defocus parameter $W_{m2}/\lambda = 0, 0.5, 1.0, 1.5, 2.0$, and 2.5 . (b) MTFs with defocus parameter fixed at $W_{m2}/\lambda = 2.0$, $W_{m2}/\lambda = 2.0$, with the cubic phase parameter $W_{m3}/\lambda = W_{m3}/\lambda = 4.0, 8.0, 10.0, 12.0$, and 14.0 . While the curves for $W_{m3}/\lambda > 4.0$ look as if they are approaching an asymptote in the mid-frequency range, in fact they are approaching the approximate values given by (8-15) and are inversely proportional to $W_{m3}\sqrt{W_{m3}}$. (c) MTFs for a fixed cubic phase parameter $W_{m3}/\lambda = 8$, with variable amounts of defocus. The defocus parameter W_{m2}/λ for these curves takes the values $0.5, 1.0, 1.5, 2.0$, and 2.5 . The curves are almost indistinguishable, confirming that in the mid-frequency range, the MTF is, to a good approximation, independent of the amount of defocusing.

All three graphs plot $f_x/2f_o$ values along the horizontal axis marked from minus 1 to +1 and $|H(f_x/2f_o)|$ values along the vertical axis marked from 0 to 1. Each graph shows several, repeatedly intersecting curves that are symmetric about the vertical axis.

In graph a, there is a straight line connecting (0, 1) on the vertical axis to (1, 0) on the horizontal axis. Another curve beginning at (0, 1) follows a steep downward slope and reaches (0.5, 0) and then rises gently and falls extending up to (1, 0). There are a few more curves that similarly begin at (0, 1), fall steeply, and then form a wavelike pattern of rise and fall below the second curve described above. These curves are all reflected on to the other side of the vertical axis. Graph b shows several densely packed curves running very close to the vertical axis, extending from (0, 1) to around (0, 0.3) and then continuing in a slight wavelike pattern up to (1, 0) in a gentle downward slope. One of the curves is plotted slightly but distinctly higher than the rest. A callout pointing at it reads $W_{m3}/\lambda = 4.0$. These curves are all reflected on to the other side of the vertical axis. Graph c shows several densely packed curves running very close to the vertical axis, extending from (0, 1) to around (0, 0.3) and then continuing in a slight wavelike pattern up to (1, 0) in a gentle downward slope. The extent of overlap between the curves is such that they appear almost like a single thick line. These curves are all reflected on to the other side of the vertical axis.

[Figure 8.3](#) shows the behavior of the *phase* of the exact OTF when the defocus parameter $W_{m2}/\lambda = 2$ and the cubic phase parameter $W_{m3}/\lambda = 8$. This phase shift must be compensated for in the processing of the image spectrum. Note that a linear approximation to this phase function corresponds to image shift. If this image shift is removed, the phase function to be compensated is somewhat simpler.

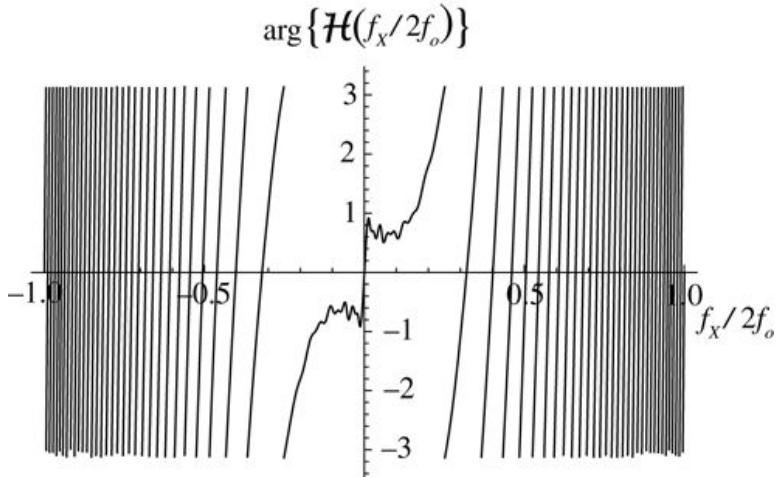


Figure 8.3
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

[Figure 8.3](#) Phase structure of the OTF when $W_{m2}/\lambda = 2.5$ and $W_{m3}/\lambda = 8$.

The graphs plots $f_x/2f_o$ values along the horizontal axis marked from minus 1 to 1 and values of argument $\{H(f_x/2f_o)\}$ along the vertical axis marked from minus 3 to +3. Between 1 and 0.3 and between minus 1 and minus 0.3, there are vertical lines extending from top to bottom across the plane. The lines closer to +1 and minus 1 are perpendicular to the horizontal axis and as we move toward the vertical axis the lines marginally

become more and more slanted such that the lower end of the line nearest to the vertical axis on the right side is closer to the vertical axis than the upper end is. It is the other way round on the left side. A curve that completely stands out of this pattern begins in the third quadrant near (0.3, minus 3) and arches in a slight zigzag pattern to almost reach (0, minus 1) and then shoots up to intersect the origin and reflect the curve thus far across the origin on to the first quadrant.

We should note that the cubic phase plate need not be a separate structure that is placed against the lens. Rather, a single imaging element can be designed to incorporate both the normal focusing power of a lens and a cubic phase aberration. In fact, there are other methods for modifying the pupil function to extend the depth of field that use entirely different approaches to pupil modification. The reader interested in pursuing this subject may wish to consult [63] for an example.

Finally, since the MTF is suppressed to a level given by (8-15) in the mid-frequency range, a restoration filter (for example, see [Section 10.6.2](#)) will be required to restore the MTF to its unaberrated levels. Increased exposure times, as compared with the case of no focusing error, will be required to obtain a signal-to-noise ratio adequate for the restoration filtering to work well.

8.2 Rotating Point-Spread Functions for Depth Resolution

A point-spread function that rotates as an object point source moves away from the plane of best focus (while approximately maintaining its transverse coordinates) is potentially useful for determining the location of that point source in three-dimensional space. A significant number of publications on such rotating point-spread functions exist. See, for example, [309], [206], [193], [283], [147], and [279]. Most of these publications have used the fact that the Laguerre-Gaussian beams form a complete orthogonal set of modes with which any paraxially propagating beam can be represented, as well as the fact that for proper choice of parameters, the beam can be made to rotate as it propagates in the z -direction.

To explain this phenomenon, we rewrite the expression (4-97) for the general Laguerre-Gaussian beam as

$$\begin{aligned} U_{l,p}(r, \theta, z) &= G_{l,p}(r, z) \Theta_l(\theta) \Psi_{l,p}(z), \\ U_{l,p}(r, \theta, z) &= G_{l,p}(r, z) \Theta_l(\theta) \Psi_{l,p}(z), \end{aligned} \quad (8-16)$$

where

$$G_{l,p}(r, z) = A_{l,p} W_0 W(z) 2r W(z) |l| L_p |l| 2r^2 W^2(z) \exp(-r^2 W^2(z)) \times \exp(jkz + jkr^2 R(z) - j\psi(z)),$$

$$\begin{aligned} G_{l,p}(r, z) &= A_{l,p} \left[\frac{W_0}{W(z)} \prod_{n=1}^{|l|} \frac{\sqrt{2}r}{W(z)} \right] |l| L_p |l| \left(\frac{2r^2}{W^2(z)} \right) \exp \left[-\frac{r^2}{W^2(z)} \right] \\ &\times \exp \left[jkz + jk \frac{r^2}{2R(z)} - j\psi(z) \right], \end{aligned} \quad (8-17)$$

$$\begin{aligned} \Theta_l(\theta) &= \exp[jl\theta], \\ \Theta_l(\theta) &= \exp[jl\theta], \end{aligned} \quad (8-18)$$

and

$$\Psi_{l,p}(z) = \exp[-j(|l|+2p)\psi(z)].$$

$$\begin{aligned} \Psi_{l,p}(z) &= \exp[-j(|l|+2p)\psi(z)]. \\ (8-19) \end{aligned}$$

As a reminder, the various important parameters in these equations are given by:

(r, θ)	Radius and angle in a transverse plane,
(l, p)	The angular and radial indices, respectively, associated with the modes
z	Axial position, with $z=0$ being the position of the beam waist
W_0	Radius of the beam waist at focus, i.e. at $z=0$
$2z_0$	The Rayleigh range, a measure of the depth of focus
$W(z)$	The beam width at axial coordinate z , $W(z)=W_0\sqrt{1+\left(\frac{z}{z_0}\right)^2}$
$R(z)$	Radius of curvature of the wavefront at axial coordinate z ,
	$R(z)=z_0^2/z^2$,
$\psi(z)$	The Gouy phase at axial position z , $\psi(z)=\tan^{-1}\frac{z}{z_0}$
$A_{l,p}$	The mode amplitude normalization constant, $2p!(1+\delta_{0l})\pi(l +p)!\sqrt{\frac{2p!}{(1+\delta_{0l})\pi(l +p)!}}$, where δ_{0l} is a kronecker delta. The modes are normalized to have equal total power.

We can identify two conditions that must be satisfied if we are to realize a rotating point-spread function:

1. The point-spread function must not be circularly symmetric about the z axis, and
2. The point-spread function must rotate as a function of z .

To consider initially the first of these requirements, we examine the $z=0$ plane where $\Psi_{l,p}(0)=1$. The reader may wonder how the circularly symmetric intensity patterns shown in Fig. 4.22 can be superimposed to yield a non-circularly-symmetric pattern of intensity. The answer lies in the fact that while the modal intensities are circularly symmetric for every index pair (l, p) , the modal *phases* are not circularly symmetric and can be superimposed to yield non-symmetric patterns. Note first that as z departs from $z=0$, the term $G_{l,p}(r, z)$ changes only by virtue of an increasing beam width, a decreasing radius of curvature (within the Rayleigh range), a linear phase shift kz , and a Gouy phase term. The linear phase shift, the quadratic-phase factor and the Gouy phase term in $G_{l,p}$ affect all modes identically, and therefore do not affect the conditions required for rotation.

For the moment, consider the plane of best focus ($z=0$) and fix the radial index $p=p_0$. Consider only the sum over the set \mathcal{L} of azimuthal indices l ,

$$U_{p_0}(r, \theta, 0) = \sum_{\mathcal{L}} G_{l, p_0}(r, 0) \exp[jl\theta].$$

$$U_{p_0}(r, \theta, 0) = \sum_{\mathcal{L}} G_{l, p_0}(r, 0) \exp[jl\theta].$$

(8-20)

This equation is reminiscent of a Fourier series in angle θ , with period 2π and Fourier coefficients $G_{l, p_0}(r, 0)$. Thus in the $z=0$ plane, a periodic function of angle can be synthesized, with the form of one period in angle determined by the choice of Fourier coefficients that depend on Laguerre polynomials with index (l, p_0) . In the more general case with an arbitrary z but still with $p=p_0$, we can write the field as

$$U_{p_0}(r, \theta, z) = \sum_{\mathcal{L}} G_{l, p_0}(r, z) \exp[jl\theta] \exp[-j(|l| + 2p_0)\psi(z)].$$

$$U_{p_0}(r, \theta, z) = \sum_{\mathcal{L}} G_{l, p_0}(r, z) \exp[jl\theta] \exp[-j(|l| + 2p_0)\psi(z)].$$

(8-21)

In this case of a fixed index $p=p_0$, the limited subset of the Laguerre-Gaussian modes is not a complete set of basis functions, and therefore the field profiles that can be synthesized are not arbitrary. Nonetheless, because of the linear dependence of the exponent $j\theta - (|l| + 2p_0)\psi(z)$ on $|l|$ in (8-21), the distribution can be made to rotate as z changes. The final exponential in the previous equation depends on the radial coefficient p_0 . For each $|l|$ this adds a constant phase shift, but the value of the phase shift depends on which radial mode we are considering.

The total field $U(r, \theta, z)$ at axial distance z is formed by summing over a yet-to-be-determined subset \square of the indices (l, p) . We index that set by an integer n ; each member of the set has an associated pair (l_n, p_n) , and we assume that there are a finite number N members of the set. Then

$$U(r, \theta, z) = \sum_{n=1}^N G_{l_n, p_n}(r, z) \Theta_{l_n}(\theta) \psi_{l_n, p_n}(z) = \sum_{n=1}^N G_{l_n, p_n}(r, z) \exp(jl_n\theta) \exp[-j(|l_n| + 2p_n)\psi(z)],$$

$$\begin{aligned} U(r, \theta, z) &= \sum_{n=1}^N G_{l_n, p_n}(r, z) \Theta_{l_n}(\theta) \psi_{l_n, p_n}(z) \\ &= \sum_{n=1}^N G_{l_n, p_n}(r, z) \exp(jl_n\theta) \exp[-j(|l_n| + 2p_n)\psi(z)] \\ &= \sum_{n=1}^N G_{l_n, p_n}(r, z) \exp\left[jl_n\left(\theta \mp \left(1 + \frac{2p_n}{|l_n|}\right)\psi(z)\right)\right], \end{aligned}$$

(8-22)

where $p_n \geq 0$ and the top sign in the last line above is to be used when $l_n > 0$ and the bottom sign when $l_n < 0$. The task remains to find the set \square over which we sum.

We make the following observations:

1. The term $\exp[j\ln\theta \mp (1+2pn|\ln|)\psi(z)]$ is responsible for rotation of the amplitude pattern; the amount of rotation at distance $z z$ is $\pm 1+2pn|\ln|\psi(z) \pm (1 + \frac{2p_n}{|l_n|})\psi(z)$.
2. The indices (l_n, p_n) should be chosen so that $2pn/|\ln| 2p_n / |l_n|$ is a constant for all added modes, thus ensuring that they will all rotate by the same amount.
3. Since the Gouy phase $\psi(z)$ varies from $-\pi/2$ to $\pi/2$ when $z z$ varies from $-\infty -\infty$ to $\infty \infty$, for a given ratio $2pn/|\ln| 2p_n / |l_n|$, the angle of the point-spread function amplitude pattern will rotate from $-1+2pn|\ln|(\pi/2) - (1 + \frac{2p_n}{|l_n|})(\pi/2)$ to $+1+2pn|\ln|(\pi/2) + (1 + \frac{2p_n}{|l_n|})(\pi/2)$ as $z z$ ranges from $-\infty -\infty$ to $\infty \infty$.
4. If $2pn/|\ln|=1 2p_n / |l_n| = 1$, the point-spread function will rotate from $-\pi$ to π as $z z$ passes from $-\infty -\infty$ to $\infty \infty$; that is, it will make a single full rotation of 360 degrees.
5. If $2pn/|\ln|=1 2p_n / |l_n| = 1$, the point-spread function will rotate $(K+1)/2 (K+1)/2$ times 360 degrees as $z z$ passes from $-\infty -\infty$ to $\infty \infty$.
6. Since only a subset of the available modes can be used if rotation is to be achieved, it is not possible to realize arbitrary point-spread function shapes, since the set of modes that can be used is not a complete set.
7. In many applications it is the intensity point-spread function that is of interest. If the amplitude point-spread function rotates by an angle θ_0 , so too will the intensity point-spread function.

[Figure 8.4](#) shows density plots of the intensity point-spread function predicted by the above results. In part (a), the four modes $(l, p)=(2,1),(4,2)(6,3)$ and $(8,4)$ are added, while in part (b), the four modes $(l, p)=(1,1),(2,2)(3,3)$ and $(4,4)$ are added. Note that in this figure the expansion of the width of the beam as $z/z_0 z/z_0$ grows has been neglected.

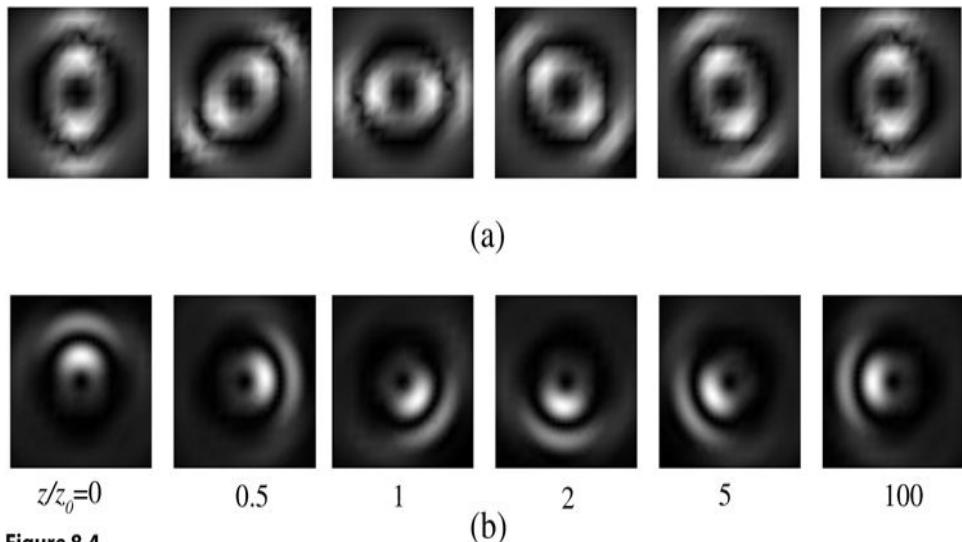


Figure 8.4

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W.H. Freeman and Company

Figure 8.4 Rotation of the intensity point-spread function as z/z_0 grows from 0 to 100. $z/z_0 = 0$ is the plane of best focus, and $z/z_0 = \pm 1$ are the boundaries of the Rayleigh range. (a) In this example, $2p/|l|=1$, and the four modes $(l,p)=(2,1),(4,2)(6,3)$ and $(8,4)$ are added. The total rotation represented by these density plots is 180 degrees, so over $z=\pm\infty$ the total rotation is 360 degrees. However, the symmetry of the point-spread function restricts the unambiguous range to $z/z_0=\pm\tan(\pi/4)$.
 (b) In this example, $2p/|l|=2$, and the four modes $(l,p)=(1,1),(2,2),(3,3),$ and $(4,4)$ are added. The total rotation represented by these density

plots is approximately 270 degrees, so over $z/z_0=\pm\infty$, the total rotation is $\frac{1}{2}$ times 360 degrees.
 Rotation can be restricted to 360 degrees if z is restricted to the range $z/z_0=\pm\tan(\pi/3)$.

The six images in each series correspond to z/z_0 values of 0, 0.5, 1, 2, 5, and 100. The images in series a show a dark spot at the center surrounded by a bright ring surrounded by a dark ring. They are all blurry images. The six images in series b show the same pattern of rings but they appear clearer and smooth and rounded.

In microscopy applications in which a dilute set of point sources (i.e. point sources with non-overlapping point-spread functions) located at various depths are to be imaged, the depth of a point source can be determined from the orientation of the point-spread functions it generates, provided that the object volume of interest is restricted to eliminate any ambiguity caused by rotation greater than 360 degrees, or by symmetries of the point-spread function. For an example of an application, see [280].

The point-spread functions are generated by use of an appropriate mask in the Fourier plane of the microscope. Because the throughput of any mask formed by sums of Laguerre-Gaussian modes is limited by absorption, suitable pure phase masks with high optical throughput have been found by means of iterative techniques that reinforce high throughput in the Fourier plane and reinforce the desired form of the rotating the point-spread function in various axial planes [279].

See [problems 8-1](#) and [8-2](#) for further exploration of the properties of sums of Laguerre-Gaussian modes.

8.3 Point-Spread Function Engineering for Exoplanet Discovery

Imaging exoplanets (i.e. planets in other solar systems) orbiting around distant stars is a subject of great interest in astronomy, and many ingenious methods have been invented to make this possible. For an excellent review of this field, see [349]. The ratio of the amount of light collected from a star to the amount of light collected from a planet orbiting that star (i.e. the “contrast”) varies from 10^{11} to 10^6 , with the smaller contrasts encountered in the infrared. The separation of the planet from its star, as observed from Earth, may be in the range 10^{-2} to 10 arcsec, again depending on the particular planet/star system being imaged. Imaging from Earth’s surface in general requires adaptive optics to remove the effects of atmospheric turbulence, ideally yielding near diffraction-limited performance.

The central problem in imaging exoplanets is suppressing the light from the star so that the light from the planet can be observed. Many unique methods for suppressing the starlight have been invented by astronomers, and exoplanets have indeed been successfully imaged. In many cases, these methods involve point-spread function engineering. We review just two of these here, but provide references to other approaches.

8.3.1 The Lyot Coronagraph

The term “coronagraph” refers to a multitude of types of instruments that at least partially block the light from the star while passing light from the planet, thus improving the contrast of an image of the planet. The original coronagraph was invented by a French astronomer, Bernard Lyot, and was introduced in 1931. The purpose of the invention at that time was to image the corona of the Sun; the contrast of the Sun/corona system is of the order of 10^6 . The geometry of the Lyot coronagraph is best understood with the help of Fig. 8.5. The telescope pupil is in plane P_1 . The darker rays passing through the pupil represent light arriving from a bright star and the lighter rays represent light arriving from a faint planet orbiting that star. The angular offset of the star is not known *a priori* because neither the position of the planet in its orbit nor the angular separation of the planet from its star are known. The telescope is pointed at the star, and its image forms on-axis in the plane P_2 . An occulting mask is placed in plane P_2 to block as much as possible of the light from the star while simultaneously passing as much as possible of the light from the off-axis planet. Assuming that the effects of the atmosphere have been cancelled by an adaptive-optic system, the image of the star is a diffraction limited Airy pattern, some sidelobes of which are not blocked by the stop. The light from the planet also forms an Airy pattern, most of which passes around the image stop. Note that the larger the stop, the more the light from the star is blocked, but the further away from the star the planet must be to pass by the stop with minimum attenuation. The occulting mask in plane P_2 acts as a spatial filter, blocking the low-spatial frequencies of the light from the star, and passing primarily only the high spatial frequencies that have been generated by the edge of the telescope pupil. Thus in plane P_3 , a filtered image of the telescope pupil is formed, with most of the light at the edges of the pupil, and a second stop,

called the Lyot stop is placed to block that light. Finally, the imaging lens forms an image in the camera plane, where the contrast of the planet image has been improved.²

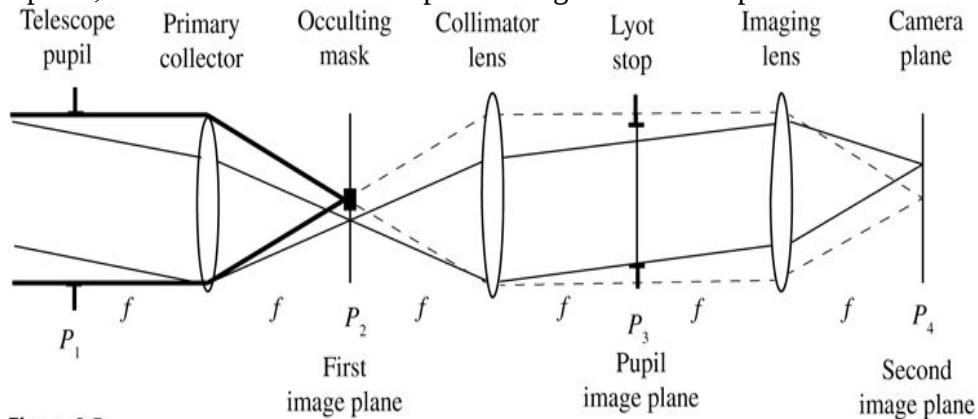


Figure 8.5

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 8.5 Geometry of the Lyot Coronagraph. The heavy lines on the left represent rays from the star, while the lighter solid lines represent rays from the planet. The light dotted lines to the right of the occulting stop represent paths the light from the star would have taken if the occulting stop were not present.

The illustration is as follows, described moving in the rightward direction. In the left extreme are two thick parallel lines of length f leading up to the extremes of the primary collector. The vertical space between the lines at a particular point is the telescope pupil P_1 . To the right of the primary collector, the thick parallel lines converge at the center of the first image plane, P_2 , labeled “Occulting mask.” The converging thick lines continue in their path beyond the mask, now shown as dotted lines diverging to reach a biconvex lens labeled “Collimator lens.” Thereafter the dotted rays run horizontal towards the biconvex imaging lens, beyond which they converge to the center of second image plane, P_4 , labeled “camera plane.” Between the collimator lens and the imaging lens is the pupil image plane, P_3 , labeled Lyot stop. Another set of ray movement is shown as follows moving in the rightward direction. A downward slanted line begins at the top left corner of the telescope and extends up to a point somewhere in the middle of the upper half of the primary collector. A line parallel to it extends up to the lower end of the primary collector. These two lines converge at a point below the center of the first image plane, P_2 . Beyond P_2 , the rays diverge as they continue on their path, one to the lower end of the collimator lens and the other to the middle of its upper half. The lines extend beyond the collimator in an upward slope, one reaching a little below the upper end of the imaging lens and the other reaching around the middle of its lower half. From the imaging lens the lines converge to a point above the center of the second image plane, P_4 .

See Fig. 8.6 for illustrations of the light intensity incident on and passed by several planes, obtained by simulation. In this simulation, if the normalized maximum of the star Airy disk incident on the stop is taken to be unity, then the maximum of the planet Airy disk in the same plane is 10^{-6} . However, in the camera plane at the output of the coronagraph, the maximum intensity of the star image is found to be 4.6×10^{-2} and the maximum intensity of the planet image is 1.5×10^{-4} . Thus in this simulation the coronagraph has changed the planet/star intensity ratio from 10^{-6} to 3.4×10^{-3} in this particular example.³

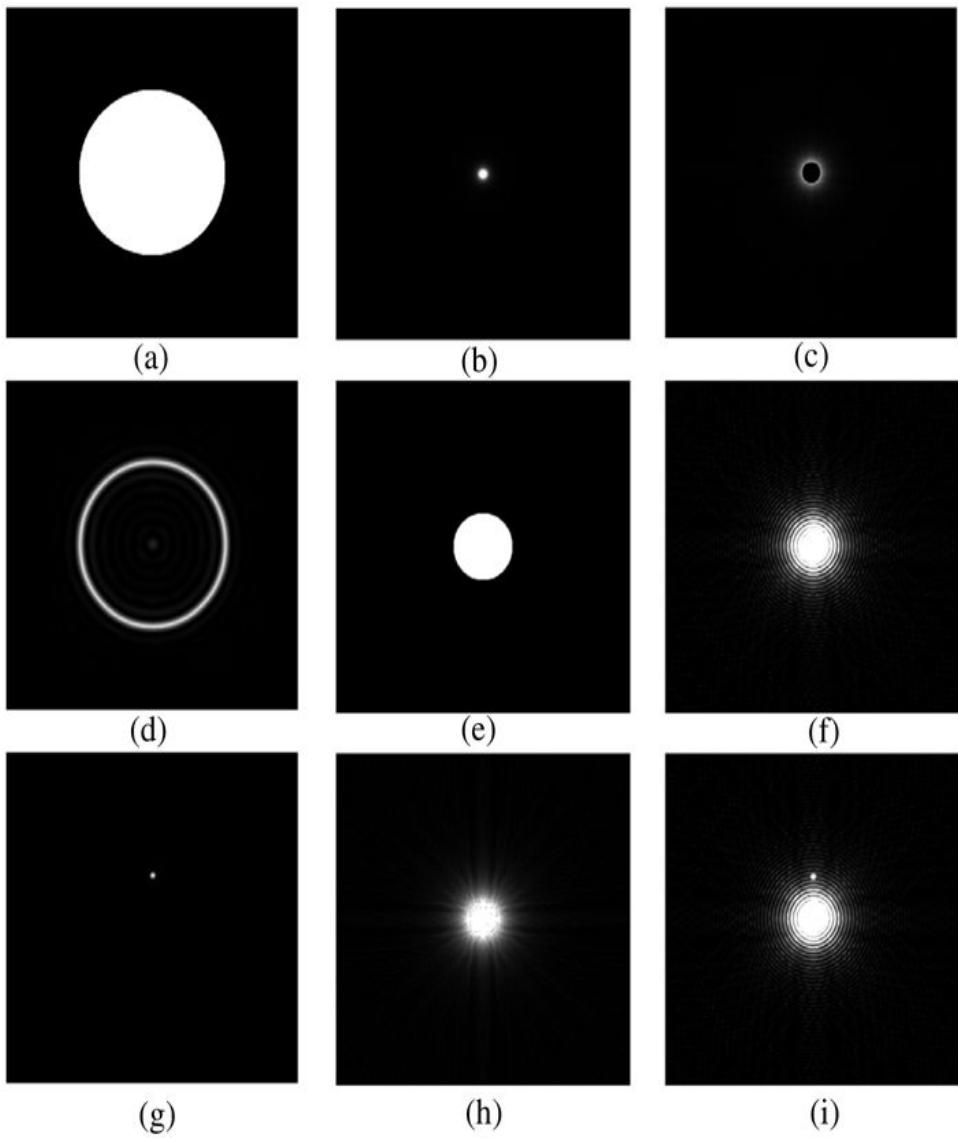


Figure 8.6

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 8.6 Simulation of the effects of the Lyot coronagraph. The planet/star intensity ratio is 10^{-6} in this example. (a) The telescope pupil (plane P_1) when starlight alone is incident. The simulation uses 400×400 pixels and the pupil has a diameter of 200 pixels. (b) Intensity distribution of the starlight incident on the initial image plane P_2 . (c) Starlight passed by the stop of diameter of 26 pixels in plane P_2 . (d) The resulting image of the pupil plane P_3 . The presence of the central stop in the image plane has diffracted light to the edges of the pupil. (e) The Lyot stop, having a clear opening with a diameter of 80 pixels. (f) An overexposed image of the star alone in plane P_4 . (g) Image of the planet alone in plane P_4 . The planet is assumed offset from the star by 50 image pixels. (h) Image of the star and the planet in plane P_2 . The planet is not detectable. (i) Image of the star and the planet in plane P_4 . The planet can be seen next to an overexposed image of the star. Note that in practice, the planet would be much closer to the star than the 50 image pixels assumed here, and the image of the planet would be in the midst of sidelobes of the image of the star, but for illustration purposes here we have chosen a larger separation.

The images are as follows. Image a: a large white circle. Image b: a tiny white spot. Image c: a spot of brightness hidden behind a dark circle such that only the edge of the brightness is in view. Image d: concentric circles that are almost indistinct from the background, the largest of which is, however, bright and distinct.

Image e: a small white circle. Image f: a small bright spot with ripples of light around it. Image g: a white spot tinier than that in image b. Image h: a very bright spot with rays radiating from it. Image i: same as image f.

A large number of variants of the original Lyot coronagraph exist. The image stop can take many other forms, including some that use a pure phase mask instead of an opaque stop. See [159] and [300] for examples. An additional example is the so-called “vortex phase filter” which, when placed in the location of the Airy pattern from the star, deflects light away from the center of the pupil to be blocked by the Lyot stop but allows most of the light from the planet to pass to the final image plane [115] [185].

As a final note, the Lyot coronagraph may be considered an example of point-spread function engineering in that achieves a point-spread function that is angle-variant, attenuating on-axis point-source objects but passing off-axis point-source objects.

8.3.2 Apodization for Starlight Suppression

An obvious approach to making a faint planet orbiting around a bright star more visible is apodization of the telescope pupil. We have discussed apodization in [Section 7.4.5](#) and seen that it can be used to reduce the strength of the sidelobes of a point-spread function. A Gaussian apodization was illustrated there. However, because the pupil is finite, a Gaussian apodization always results in a residual step of transmittance at the edge at the pupil (see [Fig. 7.14](#)), and this step is largely responsible for the residual sidelobes. The narrower the Gaussian apodization, the lower the residual step at the edge of the pupil and the lower the sidelobes, but at the same time a narrower Gaussian apodization results in less light transmitted through the pupil and a broader main lobe of the PSF. The tradeoff between transmitted light and lower sidelobes is common in apodization problems.

While a number of apodization functions with continuous, circularly-symmetric graylevel amplitude transmittance might be considered, the manufacturing tolerances for such masks are in practice very severe if suppression of starlight to the required levels is to be achieved. As a consequence, attention has been focused on binary masks, which can be made with great accuracy using the methods of microlithography. Since the planet is orbiting around its star, if the orbit is properly situated with respect to the telescope, it can appear anywhere in the region surrounding the main lobe of the PSF of the star. As a consequence, an alternative approach to apodization would be to suppress the sidelobes only in a certain finite region around the main lobe, allowing the sidelobes to be larger outside of this region. The planet may then be detectable as it orbits through the region of suppressed sidelobes, while not detectable in other regions around the center of the PSF caused by the star. Rotation of the apodizing mask to a series of angles allows exploration of a larger angular region. Ideally one would like to have an apodization of the pupil that presents no hard edge whatsoever in the direction in which the sidelobes are to be suppressed. In practice, if the apodization suppresses the starlight in only a certain angular region around the star, the mask can be rotated and additional images can be taken to find the planet in its orbit.

Such an apodization can be provided by the prolate spheroidal wave functions [320], [215], [260], which in one dimension are eigenfunctions $\psi_n(c, x)$ of the equation

$$\int -11 \operatorname{sinc}(x-\xi) \pi(x-\xi) \psi_n(c, \xi) d\xi = \mu_n(c) \psi_n(c, x),$$

$$\int_{-1}^1 \frac{\sin(c(x - \xi))}{\pi(x - \xi)} \psi_n(c, \xi) d\xi = \mu_n(c) \psi_n(c, x),$$

(8-23)

where $\mu_n(c)$ are the eigenvalues, and c is a parameter related to the space-bandwidth product. In one dimension, such functions can be shown to have the properties that (1) they are zero outside of a finite region which we take to be $(-1, 1)$, and (2) their Fourier transforms have maximum possible concentration of energy in the main lobe and minimum possible concentration in the sidelobes. These are properties ideally suited to apodization. Two-dimensional versions of the prolate spheroidal wavefunctions also exist and have similar properties [319].

To introduce the use of prolate spheroidal functions for apodization, we present a simplified one-dimensional analysis based on the results of [Prob. 2-18](#), which we repeat here. (The reader may also wish to consult [3]). Suppose we insert in an aperture a binary apodization mask with an amplitude transmittance defined by

$$t_A(x, y) = 1 \text{ for } -g(x) \leq y \leq +g(x) \quad 0 \text{ otherwise,}$$

$$t_A(x, y) = \begin{cases} 1 & \text{for } -g(x) \leq y \leq +g(x) \\ 0 & \text{otherwise,} \end{cases}$$

(8-24)

where $g(x)$ is a given function of x , to be chosen. Then, as implied by the projection-slice theorem, the spectrum $G(f_X, f_Y)$ evaluated along the f_X axis is given by a one-dimensional Fourier transform,

$$G(f_X, 0) = 2 \int_{-\infty}^{\infty} g(x) \exp(-j2\pi x f_X) dx.$$

$$G(f_X, 0) = 2 \int_{-\infty}^{\infty} g(x) \exp(-j2\pi x f_X) dx.$$

(8-25)

If $g(x)$ is a properly chosen one-dimensional prolate spheroidal wavefunction, then along the axis $u = \lambda z_i f_X$ in the image plane the sidelobes should be minimized, at the cost of a broader main lobe and reduced power transmission. Here z_i is the image distance from the exit pupil, usually the effective focal length of the telescope, and λ is the wavelength.

Let $g(x)$ be chosen as a zero-order prolate spheroidal wave function, $\psi_0(c, x)$ and consider the behavior of its normalized form $\psi_0(c, x)/\psi_0(c, 0)$ as the parameter c takes on three values, $c=2, 10, 50$. The results are shown in [Fig. 8.7\(a\)](#) and the derivatives of these functions with respect to x are shown in part (b). Note that a hard edge of $g(x)$ parallel to the vertical axis (an infinite value of $dg(x)/dx$) would result in the deflection of light into sidelobes along the u axis. These 3 functions all are

confined to the interval $(-1, 1)$ and do not have a hard edge parallel to the vertical axis, but as can be seen in part (b) of the figure, when the chosen value of c is too large (50) or too small (2), the value of the derivative becomes larger than for $c=10$. A large derivative near the end-points of $g(x)$ narrows the range about the x axis for which the sidelobes will be reduced. In addition, the larger the value of c , the less optical power will be transmitted to the image plane. A good choice of c is therefore in the vicinity of 10, and we use this choice in what follows. For such a choice, the fraction of power transmitted to the image is about 1/2 the power that would be transmitted by the full circular aperture.

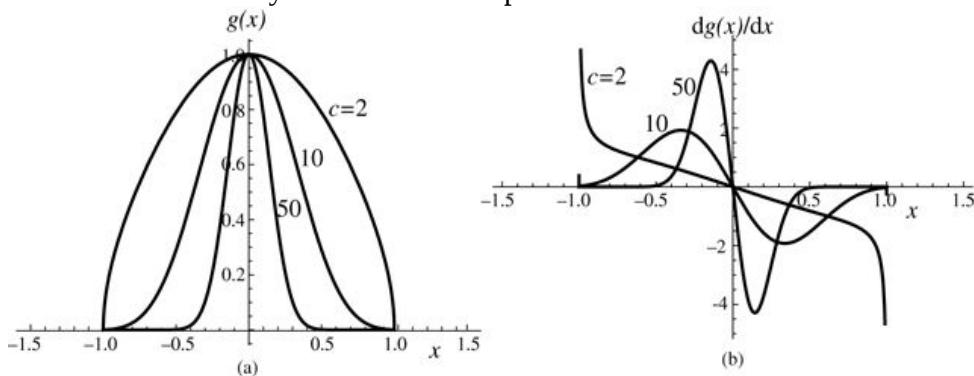


Figure 8.7

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 8.7 (a) Plots of $g(x) = \psi_0(c, x) / \psi_0(c, 0)$ for $c=2, 10, 10$, and 50 . (b) Corresponding plots of the derivative of $g(x)$ with respect to x for the same choices of c . A large derivative corresponds to a hard edge that is nearly vertical.

Graph a plots x along the horizontal axis marked from minus 1.5 to +1.5 and $g(x)$ along the vertical axis marked from 0 to 1. There are 3 curves, all three are symmetric. The curve for $c = 2$ is dome shaped, beginning at minus 1 and rising to reach $(0, 1)$ and then reflecting the path thus far on the other side of the vertical axis. The curve for $c = 10$ is bell shaped, beginning at minus 1 and rising steeply to reach $(0, 1)$ and then reflecting the path thus far on the other side of the vertical axis. The curve for $c = 50$ is also bell shaped but narrower, it begins at minus 0.5 and ends at +0.5. Graph b plots x along the horizontal axis marked from minus 1.5 to +1.5 and $dg(x)/dx$ along the vertical axis marked from minus 4 to +4. Three curves extend across the second and the fourth quadrants; all three are symmetrical across the origin. The curve for $c = 2$ is a steep downward slope extending from around $(-1, 4)$ to around $(-0.9, 2)$. Thereafter it continues in a gentle slope toward the origin. The path thus far is reflected across the origin onto the fourth quadrant.

The curve for $c = 10$ begins at $(-1, 0)$ and rises upward up to around $(-0.35, 2)$, where it begins to slope downward and reach the origin. The path thus far is reflected across the origin onto the fourth quadrant.

The curve for $c = 50$ begins at $(-0.5, 0)$ and rises steeply upward up to around $(-0.15, 4.25)$, where it begins to slope downward and reach the origin. The path thus far is reflected across the origin onto the fourth quadrant.

Figure 8.8 shows the apodization resulting from a choice of $c=10$. The white circle with radius 1 represents the edge of the pupil that is subject to apodization. This form of apodizing

mask is sometimes referred to as a “cat’s-eye” mask.

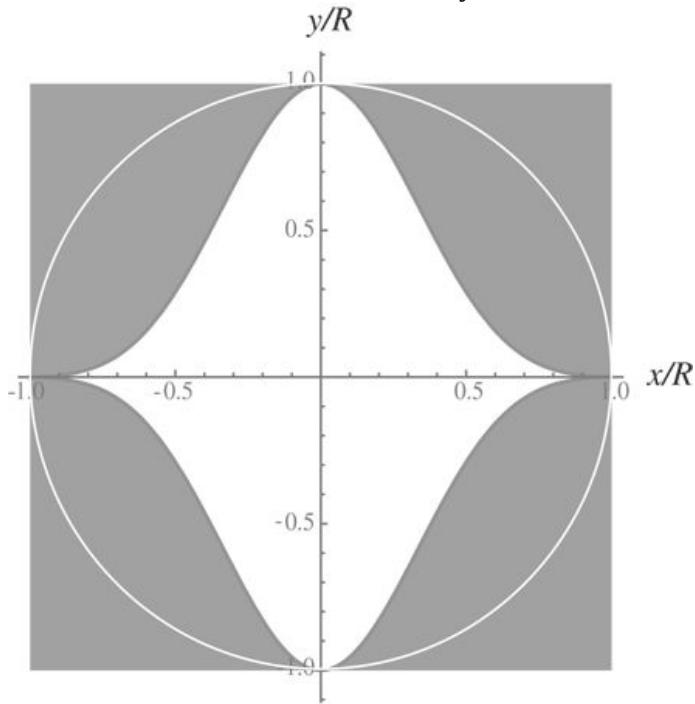


Figure 8.8

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 8.8 The “cat’s eye” apodizing mask when $g(x) = \psi_0(10, x) / \psi_0(10, 0)$. Note there are no hard edges parallel to the vertical axis. R is the radius of the circular aperture.

The graph plots x/R along the horizontal axis and y/R along the vertical axis, both axes marked from minus 1 to +1. A circle is drawn passing through $(0, 1)$, $(1, 0)$, $(0, -1)$, and $(-1, 0)$. An upward sloping smooth curve extends from $(1, 0)$ to $(0, 1)$. The curve is reflected across the axes onto the second and the fourth quarters and across the origin onto the third quadrant. The area of the plane outside the curves is shaded.

[Figure 8.9](#) shows plots of the point-spread function intensity $I_i(u, 0)$ along the u axis when cat’s-eye masks with $c=2$, 10 , and 50 are inserted in the pupil. Also shown in the same plot is the intensity distribution when there is no apodizing mask in the pupil. As can be seen, as the parameter c increases, the width of the main lobe of the point-spread function increases and the sidelobe level decreases, at the price of less total transmitted power in the image plane. Note that for $c=10$, the height of the sidelobes has been reduced by a factor in the range of 10^5 to 10^6 , depending on the sidelobe range of interest. [Figure 8.10](#) shows a density plot of the intensity distribution in the full two-dimensional point-spread function.

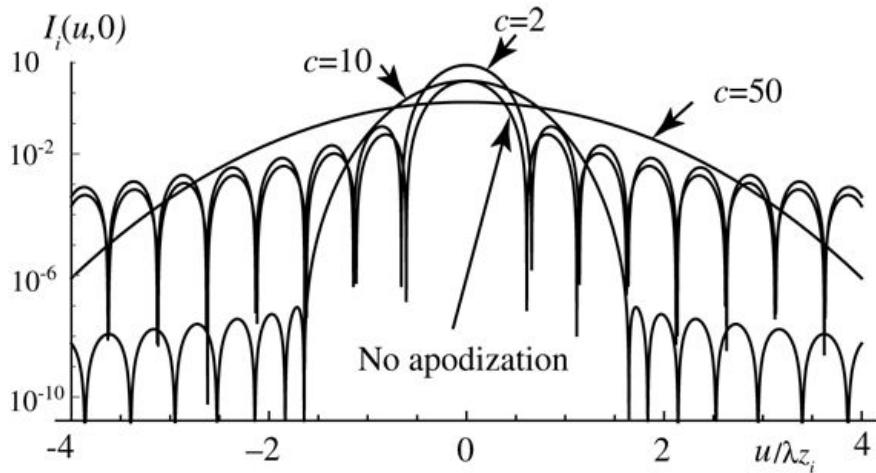


Figure 8.9

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 8.9 Intensity point-spread functions for no apodization and for cats-eye apodizations with $c=2$, 10 , 10 , and 50 . The width of the main lobe when $c=10$ is about 2.7 times the width of the main lobe when there is no apodization.

The graph plots $u/\lambda z_i$ values along the horizontal axis and $I_i(u, 0)$ along the vertical axis. The horizontal axis is marked from minus 4 to +4 and the vertical axis is marked from 10^{-10} to 10^0 . The curve for $c = 50$ is an arch that begins at the 10^{-6} mark on the vertical axis and extends up to $(4, 10^{-6})$. The highest point on the arch is near $(0, 5)$.

The curve for $c = 10$ is a continuous series of 6 small arches on the horizontal axis from minus 4 to a little beyond minus 2, where a single tall arch begins and extends almost up to +2, rising almost up to $(0, 10)$; thereafter the initial pattern of six arches is repeated symmetrically. The smaller arches rise to a height that is approximately between the levels of the 10^{-10} mark and the 10^{-6} mark on the vertical axis.

The curve for $c = 2$ is a continuous series of 7 small arches sitting atop the arches for $c = 10$. They extend from minus 4 on the horizontal axis to almost minus 0.5, where a single larger arch begins and extends a little beyond +0.5, rising up to $(0, 10)$; thereafter the initial pattern of seven arches is repeated symmetrically. The smaller arches rise to a height that range approximately between the levels a little below the 10^{-2} mark and a little above.

A callout pointing at the area where the arches intersect around $(0, 10)$ reads, “No apodization.”

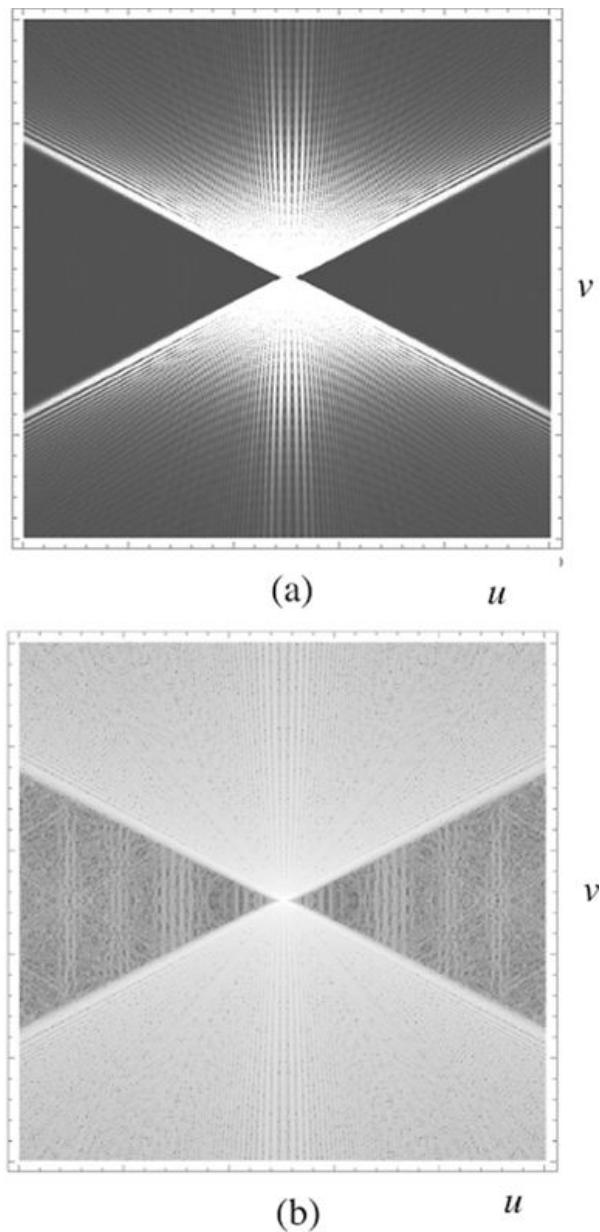


Figure 8.10
Goodman, Introduction to Fourier Optics, 4e,
 © 2017 W. H. Freeman and Company

Figure 8.10 (a) Linear density plot of the full two-dimensional point-spread function of a system with a cat's eye apodizing mask. The center of the PSF is strongly overexposed in this depiction. (b) Logarithmic plot of the same function, revealing some of the structure in the region where the point-spread function of the star would be suppressed. The intensity dynamic range represented in this figure is about 10^8 .

Each of the two square images is set along a horizontal axis u and vertical axis v . Each has a downward sloping line and an upward sloping line intersecting at their centers. The lines being equal, they are symmetric. On either side of the intersection is a triangle. And above the center is a down arrowhead shaped like an irregular pentagon. Below the center is the mirror image of that. In image a, the triangles are uniformly dark, while the rest is very bright at the center. Streaks of light are shown rising upward and downward from the center. In addition, waves of light are shown,

progressively dimming as they move from the center to the extremes. Image b is same as image a but it is bright with light. So the triangle are a light shade of gray and the other parts are even lighter.

Note that if the cats-eye mask were shrunk in the vertical direction, the slopes of the aperture edges would be reduced, and the region about the horizontal axis in which the sidelobes are reduced would be broadened in angle. The light throughput would be reduced and the resolution would be reduced in the v^y (vertical) direction. These effects can be partially reduced by opening additional apertures half-cat's-eye above and below the cat's eye, provided the apertures are chosen to have no hard edges in the x^x direction.

An improvement in apodizing mask design is represented by the work reported in [51], in which apodizing masks are found by an optimization procedure that uses constraints of (1) a binary pupil mask, (2) a targeted region in which the sidelobe level is reduced to a desired value, and (3) a specified amount of throughput of the mask. The reader should consult the cited article for more details.

8.4 Resolution beyond the Classical Diffraction Limit

For many years it was believed that the Rayleigh limit to resolution is absolute, and that no method would improve the resolution of an imaging system beyond this limit. Thus, referencing (7-50) and neglecting the 1.22 factor, an imaging system with an NA equal to the refractive index n , can resolve detail only to the limit $\lambda/2n$, which for light of 500 nm wavelength yields a resolution limit of 250 nm in air. In microscopy, immersion of the object in a liquid with refractive index n increases the maximum possible numerical aperture by a factor n^2 and for a medium with refractive index 1.5 reduces the resolution limit to 167 nm.

The first hint that the conventional diffraction limit might not be the ultimate limit of resolution came in 1952 in a paper by the Italian physicist [Toraldo di Francia \[347\]](#). In this paper he showed that, by means of a clever choice of pupil function, the width of the central lobe of the point-spread function could be made arbitrarily small and separated from sidelobes at the price of less and less light ending up in the central lobe.

In the subsections to follow, we describe a number of approaches to exceeding the conventional diffraction limit to resolution. The first approach, analytic continuation, which is closely related to Toraldo di Francia's observations, is mathematically elegant but has proven to be impractical for reasons to be explained. It is included here for completeness. The four approaches that follow this section, synthetic-aperture Fourier holography, coherent spectral multiplexing, incoherent structured illumination and Fourier ptychography, are useful primarily in extending the resolution of an imaging system so that it more closely approaches the ultimate resolution limit, that is, a transverse resolution of $\lambda/2n$. The final approaches, described in [Section 8.4.6](#) under the title "Super-resolution Fluorescent Microscopy," are capable of extending resolution well beyond the limit of $\lambda/2n$, and have been sufficiently important to warrant awarding of the 2014 Nobel Prize in Chemistry to their inventors.

8.4.1 Analytic Continuation

We shall show in this section that, for the class of *spatially bounded* objects, in the absence of noise it is *in principle* possible to resolve infinitesimally small object details.

Underlying Mathematical Fundamentals

There exist very fundamental mathematical reasons why, in the absence of noise and for the cited class of objects, resolution beyond the classical diffraction limit should be possible. These reasons rest on two basic mathematical principles, which we list here as theorems. For proofs of these theorems, see, for example, [151].

1. **Theorem 1.** The two-dimensional Fourier transform of a spatially bounded function is an *analytic* function in the (f_X, f_Y) plane.
2. **Theorem 2.** If an analytic function in the (f_X, f_Y) plane is known exactly in an arbitrarily small (but finite) region of that plane, then the entire function can be found (uniquely) by means of *analytic continuation*.

Now for any imaging system, whether coherent or incoherent, the image information arises from only a finite portion of the object spectrum (i.e. a portion of the spectrum of object amplitude in the coherent case, or a portion of the spectrum of object intensity in the incoherent case), namely, that portion passed by the transfer function of the imaging system. If this finite portion of the object spectrum can be determined exactly from the image, then, for a bounded object, the *entire* object spectrum can be found by analytic continuation. If the entire object spectrum can be found, then the exact object present can be reconstructed with arbitrary precision. As we shall see shortly, this conclusion is only valid in the total absence of noise, which is never the case in practice.

Intuitive Explanation of Bandwidth Extrapolation

A plausibility argument that super-resolution might be possible for a spatially limited object can be presented with the help of a simple example. For this example we assume that the object illumination is incoherent, and for simplicity we argue in one dimension rather than two. Let the object be a cosinusoidal intensity distribution of finite extent, with a frequency that exceeds the incoherent cutoff frequency, as illustrated in [Fig. 8.11](#). Note that the cosinusoidal intensity necessarily rides on a rectangular background pulse, ensuring that intensity remains a positive quantity.

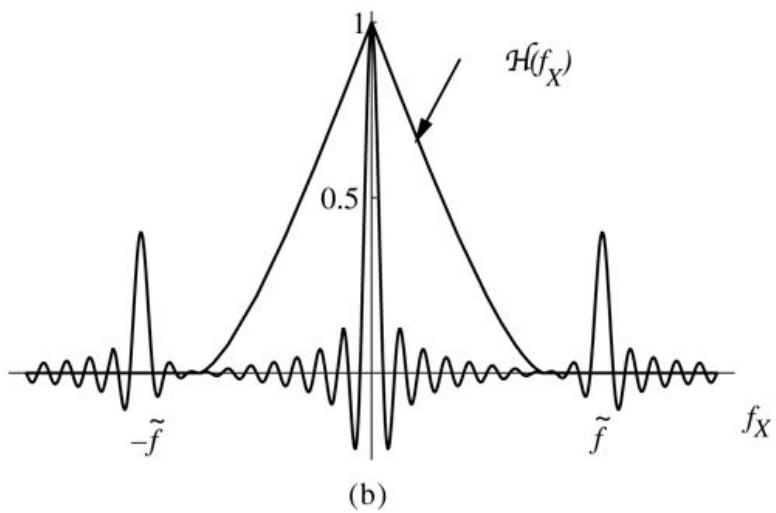
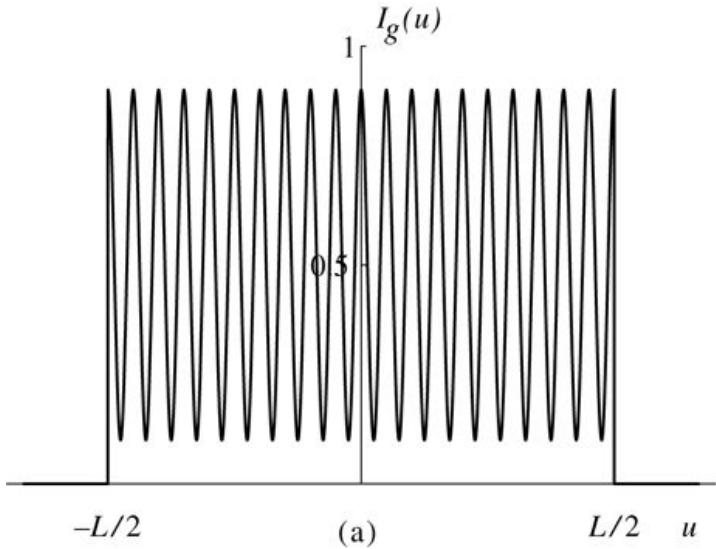


Figure 8.11

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 8.11 (a) Object intensity distribution, and (b) object spectrum and the OTF.

Graph a shows horizontal axis u plotting values from minus $L/2$ to $+L/2$ and a vertical axis plotting values of $I_g(u)$ from 0 to 1. The graph line begins with a perpendicular line at $-L/2$ that rises almost up to $(-L/2, +0.9)$. Thereafter it follows a very steep downward slope up to a point near the horizontal axis and then almost immediately rises in a very steep upward slope. This pattern repeats rightward, 10 times to the left of the vertical axis and 10 times to the right, up to a perpendicular at $+L/2$. The general appearance is that of a fine tooth comb symmetric about the vertical axis, all sloping lines being of equal length and the two perpendiculars also being equal. Graph b plots f_x along the horizontal axis and values 0 to 1 along the vertical axis.

The curve for $H(f_x)$ begins at the left extreme of the horizontal axis and overlaps the axis up to the minus $f_{\tilde{x}}$ mark, where it follows an upward slope up to $(0, 1)$. Continuing to the

right, the curve reflects the path thus far.

The other curve begins at the left extreme where the first curve begins. In a wavelike movement intersects the axis 8 times before reaching the minus f tilde mark where the wave shoots up distinctly higher, nearly level with the 0.5 mark on the vertical axis, and then falls to the axis and continues the wavelike movement of gradual widening. Just before the vertical axis it shoots up sharply toward (0, 1). Thereafter, continuing to the right, the curve reflects the path thus far.

The finite-length cosine itself can be expressed as the following intensity distribution:

$$I_g(u) = 121 + m \cos(2\pi f \tilde{u}) \operatorname{rect}(u/L).$$

$$I_g(u) = \frac{1}{2} [1 + m \cos(2\pi f \tilde{u})] \operatorname{rect}\left(\frac{u}{L}\right).$$

It follows that the (suitably normalized) spectrum of this intensity distribution is

$$\mathcal{G}_g(f_X) = \operatorname{sinc}(Lf_X) + m^2 \operatorname{sinc}[L(f_X - \tilde{f})] + m^2 \operatorname{sinc}[L(f_X + \tilde{f})],$$

$$\mathcal{G}_g(f_X) = \operatorname{sinc}(Lf_X) + \frac{m}{2} \operatorname{sinc}[L(f_X - \tilde{f})] + \frac{m}{2} \operatorname{sinc}[L(f_X + \tilde{f})].$$

as shown in part (b) of the figure, along with the assumed OTF of the imaging system. Note that

the frequency \tilde{f} lies beyond the cutoff of the OTF. The critical point to note from this figure is that the finite width of the cosinusoid has spread its spectral components into sinc functions, and

while the frequency \tilde{f} lies beyond the limits of the OTF, nonetheless the tails of the sinc

functions centered at $f_X = \pm \tilde{f}$ extend *below* the cutoff frequency into the observable part of the spectrum. Thus, within the passband of the imaging system, there does exist information that originated from the cosinusoidal components that lie outside the passband. To achieve super-resolution, it is necessary to retrieve these extremely weak components and to utilize them in such a way as to recover the signal that gave rise to them.

While the fundamental mathematical principles are most easily stated in terms of analytic continuation, there are a variety of specialized procedures that have been applied to the problem of bandwidth extrapolation. These include an approach based on the sampling theorem in the frequency domain [161], an approach based on prolate spheroidal wave-function expansions [15], and an iterative approach suitable for digital implementation that successively reinforces constraints in the space and space-frequency domains [130], [277].

Unfortunately, to achieve any meaningful expansion of bandwidth, the exceedingly weak components originating outside the passband must be detected within the passband while accompanied by ever-present noise (any optical measurement must be made with a finite amount of energy, and as a consequence the image always contains at least photon noise, and usually noise of thermal origin as well). Successful extrapolation requires exceedingly accurate measurement of the noise-free components within the passband, and this required accuracy has proven virtually impossible to achieve in practice. For discussions of the noise sensitivity of these methods, see, for example, [304], [348], [57] and [312].

Because this method has not proven successful in practice, we do not pursue it further here, but rather turn to methods that have been used with more success.

8.4.2 Synthetic Aperture Fourier Holography

Synthetic aperture Fourier holography is an approach to improving resolution using a sequence of changing angles of coherent object illumination, holographic recording of the complex amplitudes in the Fourier plane for each member of the sequence, and subsequent combining of the Fourier spectra to obtain a composite Fourier spectrum, the width of which exceeds the width of the spectrum recorded with a single on-axis illumination. We will discuss holographic recording at considerable length in [Chapter 11](#), but for our purposes here, it suffices to know that using such techniques it is possible to determine both the amplitude and phase of the light incident on a detector array. The geometry of the experiment is shown in [Fig. 8.12](#). An expanded Gaussian beam from a laser is split into a reference arm and an object arm. The expanded beams are represented by single lines in this figure. The reference is incident on a pixelated detector at an angle with respect to the optical axis. The object beam strikes a rotatable mirror and illuminates the object at an angle that can be changed between hologram recordings. The lens Fourier transforms the complex wavefront transmitted by the object, and at the detector the reference beam interferes with the complex field representing the Fourier transform of the object field. Because in each exposure the object is illuminated at a different angle, in the Fourier plane the spectrum is, step-by-step, rotated past the detector, allowing spectral components that would otherwise not pass through the aperture of the Fourier transforming lens to be captured by the detector. Each exposure thus records a different portion of the spectrum of the object, although overlap of spectral regions is needed to ensure that the same phase reference is used for all regions. The recorded complex spectra are then digitally moved to their proper center frequencies, taking account of the tilt of the object illumination in each case, and their phases are equalized in the overlap regions.

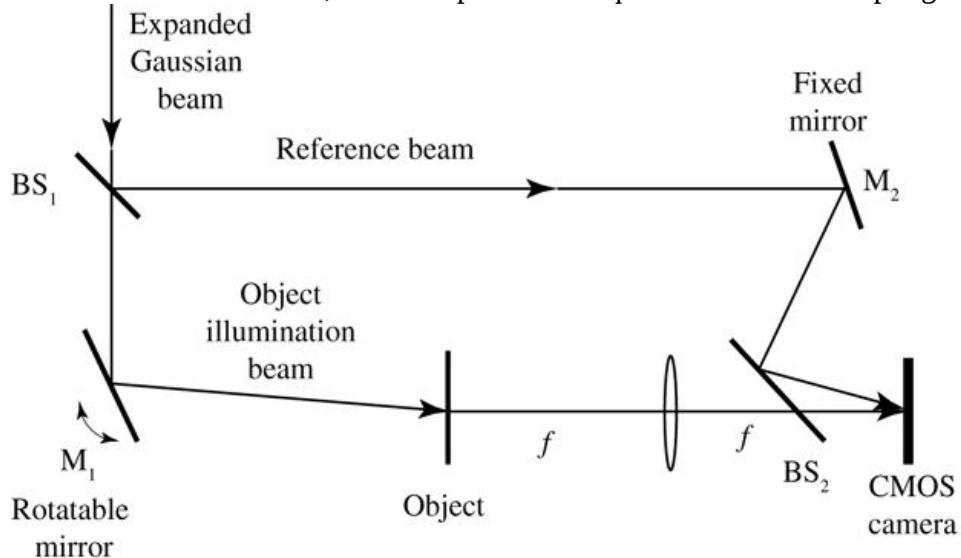


Figure 8.12

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 8.12 Geometry for synthetic aperture holography.

The illustration shows in the top left corner a vertical downward pointing ray representing an expanded Gaussian beam passing through the center of a downward sloping beam splitter BS_1 and reaching a downward sloping rotatable mirror M_1 in the bottom left corner. From BS_1 a reference beam runs horizontally rightward to the left surface of a downward sloping fixed mirror M_2 on the right extreme. From Mirror M_1 an object illumination beam runs rightward to a vertical object roughly at the center of the horizontal path. From the object the ray passes through a lens

and then via a downward sloping beam splitter 2 it reaches the CMOS camera. The reference beam above incident on M2 falls on the right surface of BS2 and also reaches the CMOS camera.

In this fashion, an imaging system with a low NA becomes a system with higher NA, while retaining the large field of view and depth of field of the lower NA system. A synthetic aperture of gigapixel size ($32,000 \times 32,000$) has been assembled and processed into an image in [114].

8.4.3 Fourier Ptychography

In [Section 8.4.2](#) we have seen that by illuminating an object with a set of coherent beams with different incidence angles, various portions of the Fourier spectrum of the object are brought into the observable region of Fourier space. In that section, the complex-valued spectrum of each image was obtained by holography. An alternative approach, called *Fourier ptychography* [385], replaces holographic detection with simpler detection of the low-resolution image intensity distributions obtained for each of a sequence of illumination angles, and uses iterative computational techniques to obtain a high-resolution image. A useful reference for this subject is found in [384].

This method uses an array of point sources (generally an LED array) to generate a two-dimensional set of plane waves at different angles and to sequentially illuminate an object with those plane waves. The low-resolution (limited by the low NA of the imaging lens) image intensities produced by each such illumination angle are detected. Recognizing that each low-resolution intensity image results from the optical system passing known regions in Fourier space with bandlimited support, an iterative algorithm can be devised that uses the known set of intensity images and the known limited Fourier spectrum support to recover and stitch together overlapping spectral regions. A much wider coverage of Fourier space can be obtained and a high-resolution image results. The resolution of the image obtained is that of the extended NA system, but the field of view of the image is that afforded by the lower NA of the unaided imaging system. Therefore the field of view is much greater than would be obtained for a conventional imaging system with the higher NA.⁴

Let $i_{h0} e^{j\phi_h}$ represent the complex amplitude distribution of the high-resolution image we are ultimately interested in obtaining. Let i_{mk} represent the intensity distribution measured when the object is illuminated by the k th incident plane wave arriving at angle α_k , where k ranges from 1 to K . Let $i_{\ell k}^{(p)}$ represent the low resolution image intensity obtained at the end of the p th iteration when the k th illumination angle is assumed. In general, $i_{\ell k}^{(p)} \neq i_{mk}$, but will be replaced by i_{mk} in the $(p+1)$ st step of the iterative process. The procedure for achieving an increase of resolution beyond that afforded by a low-NA imaging lens is related to techniques used for phase recovery (see [Section 2.7](#)). The steps in one particular iterative procedure can be summarized as follows:

1. Begin with an initial guess for the high-resolution image amplitude distribution, $i_{h0} e^{j\phi_{h0}}$. A suitable guess for i_{h0} is the intensity distribution of any one of the low-resolution measured images, represented by i_{m0} , perhaps the one taken with normally-

incident object illumination. An initial phase distribution can be taken to be $\phi_{h0} = 0$.

Now $i_{h0} = i_{m0}$ should be upsampled⁵ to an array length that will be appropriate for the high-resolution image rather than the lower length required for the low-resolution image. Thus if there are M samples in the low-resolution image, when this image is upsampled by a factor N/M , there will be N samples in the new sequence. A Fourier transform of this upsampled amplitude image yields a length- N spectrum with the length M spectrum of i_{m0} centered at its proper location in Fourier space, surrounded by zeros.

2. Choose a subregion of this expanded spectrum that corresponds to a different angle of illumination α_k and which overlaps with the spectral segment found in step 1. This subregion is bounded by the (usually circular) amplitude transfer function of the coherent imaging system, centered on the spatial frequency corresponding to the angle of illumination α_k . Inverse Fourier transform this spectral subregion to produce a new low-resolution image amplitude distribution $i_{\ell k} e^{j\phi_{\ell k}}$.
3. Replace $i_{\ell k}$ by i_{mk} , where i_{mk} is the actual measured image obtained for angle of illumination α_k . Now Fourier transform this new amplitude distribution $i_{mk} e^{j\phi_{\ell k}}$, thus expanding the non-zero region of the spectrum.
4. Repeat steps 2 and 3 for all K angles of illumination. The angles of illumination should be chosen in a two-dimensional array such that there is spectral overlap between the various subregions generated by the different angles of illumination.
5. Repeat steps 2 through 4 until there is consistency between common portions of overlapping adjacent spectral regions, in which case an extended Fourier spectrum has been obtained. The inverse Fourier transform of this extended spectrum is the amplitude distribution in the high-resolution image that is sought.

The algorithm outlined above is not the only possible algorithm nor is it necessarily the best algorithm, but it has been demonstrated to work [385]. For discussion of other algorithms, see, for example, [380]. The images resulting from this approach have both an extremely high resolution and an extremely large field of view.

8.4.4 Coherent Spectral Multiplexing

In 1966, W. Lukosz proposed [235] (see also [236]) a coherent multiplexing technique that modulated the system input with a high-frequency grating near the object plane, and demodulated with a similar grating in plane where the grating is imaged, near the image plane. The effect of the input grating was to generate multiple shifted copies of the object diffraction pattern in the entrance pupil of the system, resulting in different portions of the object diffraction pattern being aliased or multiplexed by overlapping at the entrance pupil. The second grating then demodulated the image, separating the multiplexed portions of the diffraction pattern and placing them in the proper relative positions to effectively enhance the effective numerical aperture (NA) of the system. This approach introduced ghost images at various positions away from the primary image. To avoid overlap of the primary image with the ghost images, the field of view was necessarily restricted; in effect, the total number of degrees of freedom of the image remained constant.

To achieve a wide field of view it is necessary to introduce temporal degrees of freedom which are used to ultimately increase the spatial degrees of freedom. We describe such a method here [371] [372]. Like the Fourier synthetic-aperture approach, this method requires measurement of the *amplitude* distribution in a sequence of K images. Since optical sensors respond to intensity, it is again necessary to encode complex amplitude in a measured intensity distribution, which can be done using holography, to be discussed in detail in [Chapter 11](#). For the purposes of this discussion, it suffices to simply assume that we are able to extract the complex amplitude distributions of the detected images. The images may be collected by a system with a low NA, but the ultimate synthesized image has the resolution of a high NA system, while retaining the large field of view and depth of field of a low NA system.

The approach described here consists of placing a grating near the object and shifting it in a sequence of K equal-increment steps. K images are obtained, one for each position of the grating. The grating can be imaged onto the object, or placed in near contact with the object, preceding or following it. Coherent detection via digital holography allows linear signal processing to be used to de-alias the transmitted spectrum and reconstruct images with a significant resolution gain.⁶

The grating amplitude transmittance is a periodic function with period L , and for the k th image we represent it (in one dimension for simplicity) by a complex function $P_k(x)$. This function can be expanded in a complex Fourier series,

$$P_k(x) = \sum_{n=-\infty}^{\infty} p_{k,n} e^{-j2\pi nx/L},$$

$$P_k(x) = \sum_{n=-\infty}^{\infty} p_{k,n} \exp(-j2\pi nx/L),$$

(8-26)

where the Fourier coefficients are in general complex-valued. Furthermore, we assume that the grating has been fabricated such that it possesses, for each image, a finite set of Fourier coefficients $p_{k,n}$ that are approximately equal in magnitude, meaning that all the plane wave components illuminating the object are of approximately equal magnitude, while all the other higher-order $p_{k,n}$ are close to zero. If $t_o(x)$ represents the complex amplitude transmittance of the object, which is the quantity we wish to recover, the field leaving the sandwiched object and grating for the k th image is

$$u_k(x) = t_o(x)P_k(x) = t_o(x)\sum_{n=-N}^{N} p_{k,n} e^{-j2\pi nx/L},$$

$$u_k(x) = t_o(x)P_k(x) = t_o(x) \sum_{n=-N}^{N} p_{k,n} \exp(-j2\pi nx/L),$$

(8-27)

where the grating has been assumed to have $2N+1$ significant grating orders. The spectrum of $t_o(x)P_k(x)$ is given by

$$U_k(fX) = T_o(fX) * \sum_{n=-N}^{N} p_{k,n} \delta(fX - n/L),$$

$$U_k(f_X) = T_o(f_X)^* \sum_{n=-N}^N p_{k,n} \delta(f_X - n/L),$$

(8-28)

where $T_o(f_X)$ is the object spectrum and, as usual, the asterisk represents convolution.

We explicitly assume that the grating is located before or in contact with the object. If the grating follows the object, then only the non-evanescent portion of the object spectrum for normal incidence should be considered. In the absence of the grating, the finite pupil of the system will restrict the light that passes through the pupil stop to a finite region of the spectrum, with cutoff frequencies we represent by $\pm f_p$. Consequently, important information lying at frequencies beyond $\pm f_p$ is lost, degrading the image resolution. The effect of the grating is to multiplex many different parts of the object spectrum into the pupil. The resulting image will not resemble the object, but with a series of $K \geq N$ images it will be shown to be possible to expand the spectrum by a factor $\leq N$, each image taken with an appropriate change in the grating Fourier coefficients $p_{k,n}$.

Now briefly consider the digital processing performed on a set of $2K+1$ measured fields when $2N+1$ grating orders are used (we use $2K+1$ measurements and $2N+1$ grating orders rather than K measurements and N orders for mathematical convenience, and attempt to expand the spectrum by a factor $\leq 2N+1$). Using a one-dimensional analysis, the Fourier transform of the k th detected image amplitude $A_k(f_X)$ can be written as

$$A_k(f_X) = \sum_{n=-N}^N p_{k,n} T_o(f_X - n/L) \operatorname{rect}(f_X/2f_p) \quad k = -K, \dots, K,$$

$$A_k(f_X) = \sum_{n=-N}^N p_{k,n} T_o(f_X - n/L) \operatorname{rect}(f_X/2f_p) \quad k = -K, \dots, K,$$

(8-29)

where again the subscript k indexes the detected images, while the subscript n indexes the grating orders, and the grating order amplitudes $p_{k,n}$ will change between images in a way to be determined. As before, the frequency f_p represents the cutoff frequency of the limited-NA optics.

We assume that the grating frequency is chosen such that $(\sigma f_p) = 1/L$, where σ is a factor between 0 and 1 that determines the degree of overlap of the multiplexed spectral regions. In this way, the ± 1 diffraction orders of the grating itself are located within the pupil, to a degree determined by σ , and the grating therefore produces aliasing with overlapping spectral regions. In practice, σ is taken to be between 0.75–0.90. We utilize the spectral overlap regions in signal processing, but this overlap is also needed when extending to two-dimensional

with a circular pupil so that complete coverage of the broadened spectral domain is obtained without gaps.

Equation (8-29) can be rewritten in vector form as

$$A \rightarrow (fX) = PT \rightarrow (fX),$$

$$\vec{A} (f_X) = \mathbf{P} \vec{T} (f_X),$$

(8-30)

where $A \rightarrow (fX) \vec{A} (f_X)$ and $T \rightarrow (fX) \vec{T} (f_X)$ are column vectors,

$$A \rightarrow (fX) = A \cdot K(fX) : AK(fX),$$

$$\vec{A} (f_X) = \begin{bmatrix} A_{-K}(f_X) \\ \vdots \\ A_K(f_X) \end{bmatrix},$$

(8-31)

$$T \rightarrow (fX) = T_o(fX + N/L) \text{rect}(fX/2fp) : T_o(fX - N/L) \text{rect}(fX/2fp),$$

$$\vec{T} (f_X) = \begin{bmatrix} T_o(f_X + N/L) \text{rect}(f_X/2fp) \\ \vdots \\ T_o(f_X - N/L) \text{rect}(f_X/2fp) \end{bmatrix},$$

(8-32)

and \mathbf{P} has $(2K+1)(2N+1)$ rows and $(2N+1)(2N+1)$ columns,

$$\mathbf{P} = p_{-K, -N} \dots p_{-K, N} \vdots \vdots \vdots p_{K, -N} \dots p_{K, N}.$$

$$\mathbf{P} = \begin{bmatrix} p_{-K, -N} & \cdots & p_{-K, N} \\ \vdots & \vdots & \vdots \\ p_{K, -N} & \cdots & p_{K, N} \end{bmatrix}.$$

(8-33)

Now if the number of image measurements is equal to the number of grating orders, and the matrix \mathbf{P} is nonsingular, then the matrix inverse \mathbf{P}^{-1} exists and the column vector $T \rightarrow (fX) \vec{T} (f_X)$ of spectral islands $T_o(f_X - n/L) \vec{T} (f_X - n/L)$ can be recovered as

$$T \rightarrow (fX) = \mathbf{P}^{-1} A \rightarrow (fX).$$

$$\vec{T} (f_X) = \mathbf{P}^{-1} \vec{A} (f_X).$$

(8-34)

More than $2N+1$ measurements can be made to improve the signal-to-noise ratio, in which case the matrix P is no longer square but can still be readily inverted with a

pseudoinverse operation. Once the vector $T \rightarrow (f_X)$ is known, the various spectral islands can be digitally moved to their proper center frequencies, and with the help of the overlapping portions of their spectra, can be stitched together to form a spectrum with a cutoff frequency that has been increased from f_p to $(\sigma N + 1)f_p$.

The question remains as to how the grating coefficients for each image should be chosen. We know that for the matrix P to be nonsingular, its rows must be orthogonal and its columns must be orthogonal. In addition, we want the largest possible grating-order amplitudes $p_{k,n}$ in order to maximize the signal-to-noise ratio in the detected images. A suitable choice is to use a phase grating with equal or nearly equal-strength order amplitudes (for example, a Dammann phase grating [87] has such orders). The spectral coefficients $p_{k,n}$ can be obtained by translating the grating between image captures by a fraction of the period, in particular by $\Delta x = L / (2K + 1)$. For a given grating order n , each such translation changes the phase of that component in the k th image by $\Delta\phi_{k,n} = 2\pi nk / (2K + 1)$

$$\Delta\phi_{k,n} = 2\pi nk / (2K + 1)$$

The rows and columns of P are then orthogonal, and assuming that $K \geq N$, an inverse or pseudoinverse of P can be found.

[Figure 8.13](#) shows a simplified diagram of the holographic recording geometry that allows the complex field to be measured for each member of the sequence of images. [Figure 8.14](#) shows the magnitude of demultiplexed complex spectra after placing different spectral islands in their proper positions, in this case with considerable overlap by virtue of a choice of σ that is less than unity. The resulting NA of the system has been increased from 0.063 to 0.164, an increase by a factor of 2.6. Finally, [Fig. 8.15](#) shows the image obtained through the system (a) without multiplexing, and (b) with multiplexing. The improvement of resolution in the horizontal direction is clear.

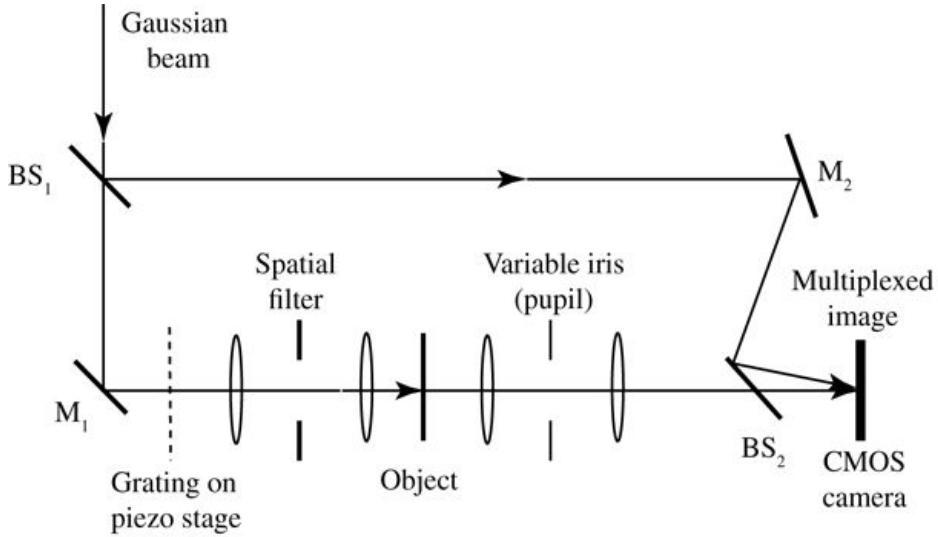


Figure 8.13

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 8.13 Holographic camera for recording complex-valued images. M_1 and M_2 are mirrors, while BS_1 and BS_2 are beam splitters. The Gaussian illumination beam is split into two paths, the upper path being the reference path and the lower path the object path. The light passes through a grating on a piezo stage that is used to accurately translate the grating. The spatial filter selects certain orders of the grating and equalizes their strengths. The object is illuminated by these orders, ($0, \pm 1, \pm 2$ orders in the example to be given). The variable iris represents the finite entrance pupil of the system, which in this example corresponds to an $NA = 0.063$. The multiplexed image falls on a CMOS detector, where it interferes with the tilted reference wave, yielding an interference pattern that encodes the phase. How to obtain the complex field from the hologram will be covered in [Chapter 11](#).

The illustration shows in the top left corner a vertical downward pointing ray representing a Gaussian beam passing through the center of a downward sloping beam splitter BS_1 and reaching a downward sloping rotatable mirror M_1 in the bottom left corner. From BS_1 a beam runs horizontally rightward to the left surface of a downward sloping mirror M_2 on the right extreme. From Mirror M_1 a beam runs rightward passing through grating on peizo stage represented by a vertical dotted line, a lens, and a spatial filter, in that order, to reach the object. The ray continues rightward, entering a lens, the variable iris (pupil), and another lens, in that order, before passing through the left surface of the beam splitter BS_2 and reaching the CMOS camera. The beam above incident on M_2 falls on the right surface of BS_2 and also reaches the CMOS camera.

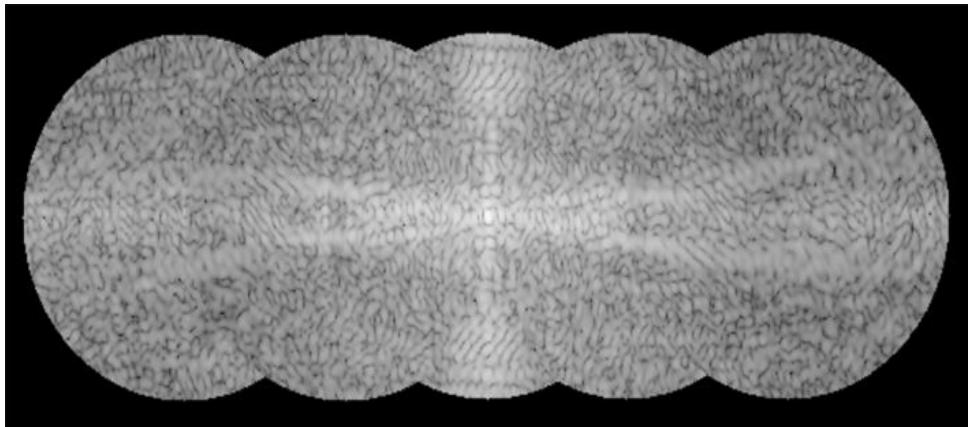


Figure 8.14

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 8.14 Reconstructed image spectrum log-magnitude when the $0, \pm 1$ and ± 2 orders of the grating illuminate the object. Coherent superresolution imaging via grating-based illumination, Jeffrey P. Wilde, Joseph W. Goodman, Yonina C. Eldar, and Yuzuru Takashima, *Appl. Opt.* 56(1), A79–A88 (2017).

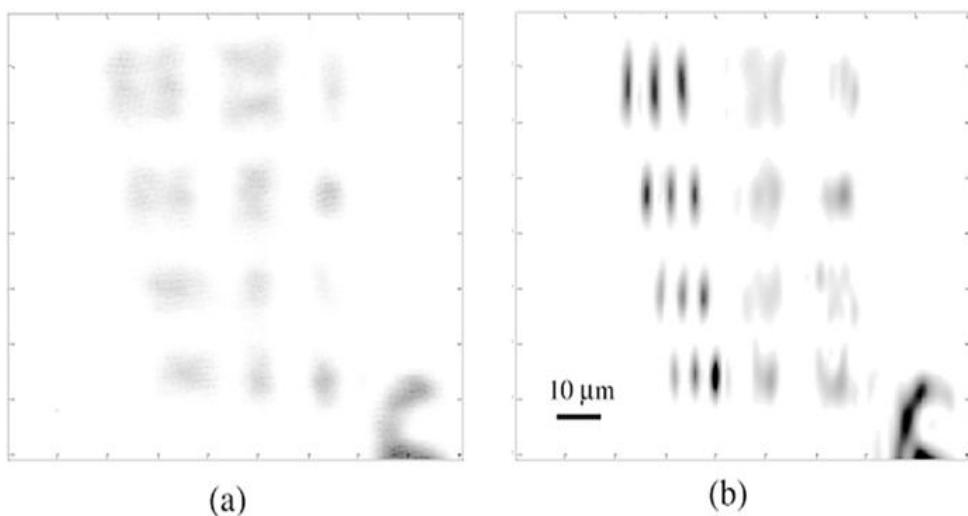


Figure 8.15

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 8.15 Images obtained (a) without multiplexing and (b) with multiplexing. Coherent superresolution imaging via grating-based illumination, Jeffrey P. Wilde, Joseph W. Goodman, Yonina C. Eldar, and Yuzuru Takashima, *Appl. Opt.* 56(1), A79–A88 (2017).

Finally, a few closing comments. First, the procedure described requires that the grating be moved in increments that correspond to the maximum resolution that the enhanced system can achieve. This is in principle not a significant limitation, since piezo stages with $< 10 \text{ nm}$ resolution are available. Second, the procedure can be generalized to two dimensions by using a grating with orders propagating at angles in two dimensions. Third, there are other possible choices for the matrix P , such as a Hadamard matrix, that can be realized if a dynamic spatial light modulator (see [Chapter 9](#)) is used to create a different appropriate complex grating for each recorded image. Fourth, the method described is only useful when the NA of the objective lens is

significantly less than unity and it is desired to increase the effective NA to something closer to unity. If the objective lens has an NA close to unity, little gain in resolution can be obtained by this method. Lastly, the coherent imaging method that has been described is closely related to the incoherent *structured illumination imaging* method to be described in the section that follows.

8.4.5 Incoherent Structured Illumination Imaging

The technique of incoherent structured illumination imaging has been pioneered by [M.G.L. Gustafsson and colleagues \[158\] \[156\]](#) (see also [\[165\]](#)). This method is most widely used for imaging fluorescent objects which emit incoherent light in response to an illumination intensity, where the illumination of the object can be coherent or incoherent. If the illumination intensity contains high-frequency fringes, various portions of the spatial spectrum of the incoherent light emitted by the object will be aliased into the optical transfer function of the incoherent imaging system.

Let the intensity transmittance or reflectance of the object be represented by $\tau_o(x)$ ($\tau_o(x)$ again for simplicity, we illustrate with one-dimensional mathematics), and let the intensity of the illumination incident on the object be represented by $I_{il}(x)$. Then assuming a linear response of fluorescent intensity to the intensity incident on the object, the light distribution to be imaged can be represented by the product of the illumination intensity and the intensity transmittance or reflectance of the object,

$$Io(x) = \kappa I_{il}(x) \tau_o(x),$$

$$I_o(x) = \kappa I_{il}(x) \tau_o(x),$$

(8-35)

where κ is a proportionality constant. The object spectrum is therefore

$$\mathcal{O}(f_X) = \kappa \mathcal{I}_{il}(f_X) * \mathcal{T}_o(f_X)$$

$$\mathcal{G}_o(f_X) = \kappa \mathcal{G}_{il}(f_X) * \mathcal{T}_o(f_X)$$

(8-36)

where $\mathcal{O}(f_X)$ and $\mathcal{I}_{il}(f_X)$ are the Fourier transforms of Io and I_{il} , respectively, while $\mathcal{T}_o(f_X)$ is the Fourier transform of $\tau_o(x)$. Passage of the object intensity to the image is modified by the OTF, $\mathcal{H}(f_X)$, of the imaging system, yielding

$$\mathcal{I}(f_X) = \mathcal{O}(f_X) \mathcal{H}(f_X) = \kappa \mathcal{H}(f_X) \mathcal{I}_{il}(f_X) * \mathcal{T}_o(f_X).$$

$$\mathcal{G}_i(f_X) = \mathcal{G}_o(f_X) \mathcal{H}(f_X) = \kappa \mathcal{H}(f_X) [\mathcal{G}_{il}(f_X) * \mathcal{T}_o(f_X)].$$

(8-37)

Now let the object be illuminated by two equal-amplitude plane waves that are equally but oppositely inclined to the optical axis, described by a complex field incident on the object

$$U_{il}(x) = A e^{j2\pi\alpha_1 x} + A e^{-j2\pi\alpha_2 x}.$$

$$U_{il}(x) = A \exp\left(j2\pi\frac{\alpha}{2}x\right) + A \exp\left(-j2\pi\frac{\alpha}{2}x\right).$$

(8-38)

The corresponding intensity of the illumination is

$$I_{il}(x) = U_{il}(x)\bar{U}_{il}(x) = 2A^2[1 + \cos(2\pi\alpha x)];$$

$$I_{il}(x) = |U_{il}(x)|^2 = 2A^2[1 + \cos(2\pi\alpha x)];$$

(8-39)

that is, the incident intensity is a cosinusoidal fringe of frequency α . The spectrum of the incident illumination intensity is accordingly

$$\mathcal{G}_{il}(f_X) = 2A^2[\delta(f_X) + \frac{1}{2}\delta(f_X - \alpha) + \frac{1}{2}\delta(f_X + \alpha)].$$

$$\mathcal{G}_{il}(f_X) = 2A^2\left[\delta(f_X) + \frac{1}{2}\delta(f_X - \alpha) + \frac{1}{2}\delta(f_X + \alpha)\right].$$

(8-40)

Now using (8-36), we find that the spectrum of the object with the illumination described above is given by

$$\mathcal{G}_o(f_X) = \kappa' \mathcal{G}_{il}(f_X) + 12\mathcal{G}_{il}(f_X - \alpha) + 12\mathcal{G}_{il}(f_X + \alpha),$$

$$\mathcal{G}_o(f_X) = \kappa \left[\mathcal{T}_o(f_X) + \frac{1}{2}\mathcal{T}_o(f_X - \alpha) + \frac{1}{2}\mathcal{T}_o(f_X + \alpha) \right],$$

(8-41)

where $\kappa' = \kappa A^2$.

Recall that the OTF of a circular aperture with radius w has a circular cutoff in frequency space at frequency $f_X = \pm 2f_0 = \pm w/(\lambda z_i)$. If the frequency α of the cosinusoidal illumination pattern is chosen to coincide with that frequency cutoff, then in the frequency domain we will have a situation as depicted in Fig. 8.16. As can be seen, the cosinusoidal fringe illumination has brought portions of the spectrum of the object into the passband of the system that previously were beyond the frequency cutoff.

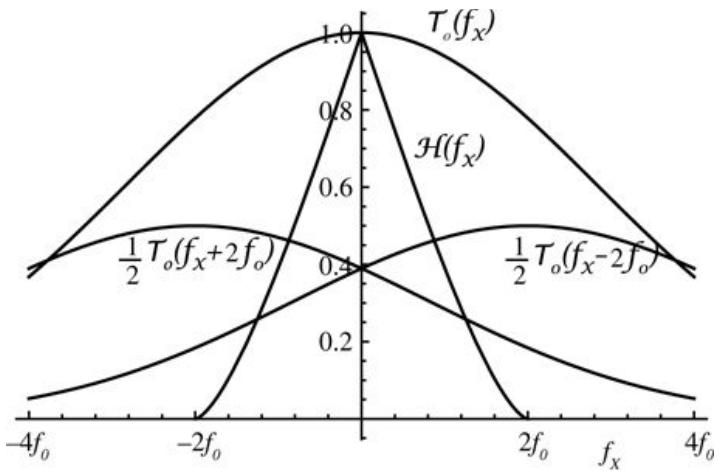


Figure 8.16

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 8.16 Frequency domain depiction of the three object spectral islands and the OTF.

The graph plots f_x along the horizontal axis marked from $-4f_0$ to $+4f_0$. The vertical axis is marked from 0 to 1. The curve for $H(f_x)$ is a steep upward slope from $-2f_0$ to $(0, 1)$, followed by a steep downward slope to $2f_0$. The curve for $T_o(f_x)$ begins a little below $(-4f_0, 0.4)$ and extends in a slightly bulging upward slope extending up to $(0, 1)$. The slope is reflected across the vertical axis. The curve for half $T_o(f_x + 2f_0)$ begins at $(4f_0, 0.4)$ and rises only gently before sloping down to $(0, 0.4)$ and then continuing the downward slope into the left side of the vertical axis almost reaching the $-4f_0$ mark on the horizontal axis. The curve for half $T_o(f_x - 2f_0)$ begins at $(-4f_0, 0.4)$ and rises only gently before sloping down to $(0, 0.4)$ and then continuing the downward slope into the right side of the vertical axis almost reaching the $4f_0$ mark on the horizontal axis.

The image captured is a superposition of three images corresponding to the portions of the three spectral islands passed by the OTF. It remains to unscramble these spectral components and place them at their proper center frequencies. This can be done by introducing a phase shift in one of the illuminating beams, and capturing three images with the cosinusoidal intensity fringe shifted by 0, 120, and 240 degrees. The procedure outlined in the previous subsection can then be used to recover the three different spectral segments, and they can then be moved to their appropriate center frequencies, extending the frequency coverage and the resolution by a factor of 2. Note that in the incoherent case (unlike the coherent case), it is necessary to compensate for the frequency-dependent attenuation introduced by the OTF.

If the fluorescent object is illuminated through and imaged by the same objective lens, as in a reflective imaging geometry, the maximum extension of bandwidth in one dimension is a factor of 2. However, if the illumination optics and imaging optics are separate, and the NA of the imaging optics is smaller than the NA of the illumination optics, resolution gains larger than a factor of 2 can be obtained by imaging a grating onto the object and moving the grating in a sequence of steps producing a sequence of images that can be processed and combined to extend the bandwidth. However, like the coherent case, this method is most useful when the NA of the objective lens is

considerably less than unity and one wishes to extend the effective NA beyond the limit posed by the objective lens.

In addition, if the fluorescence process is driven into nonlinearity by a single-frequency sinusoidal fringe of illumination intensity, harmonics of the fringe frequency will be generated at equally spaced frequencies, and shifting the phase of one of the two illuminating beams will shift the phase of the fringe and its harmonics by proper increments to allow resolution improvement beyond a factor of 2 [157].

An alternative approach uses only two illumination beams separated in angle, but steps the angle between them, thereby changing the fringe frequency. A sequence of images, each obtained with a different fringe frequency, can then be combined to extend the bandwidth by a factor larger than 2, again provided overlap of the spectral islands is included to allow compensation for any unwanted constant phase shifts between islands.

Finally, the method can be extended to two dimensions by choosing a sequence of illuminating beams that create a multitude of multiplexed spectral segments spanning a two-dimensional region of the spectrum.

8.4.6 Super-Resolved Fluorescence Microscopy

Fluorescent Labelling

A powerful technique in modern microscopy is the labelling of molecules with fluorophores that emit light of a certain wavelength when stimulated by incident light at a shorter wavelength. The simplest form of fluorescence is illustrated in the molecular energy-level diagram of [Fig. 8.17](#) (the complete energy diagram is more complicated than this).

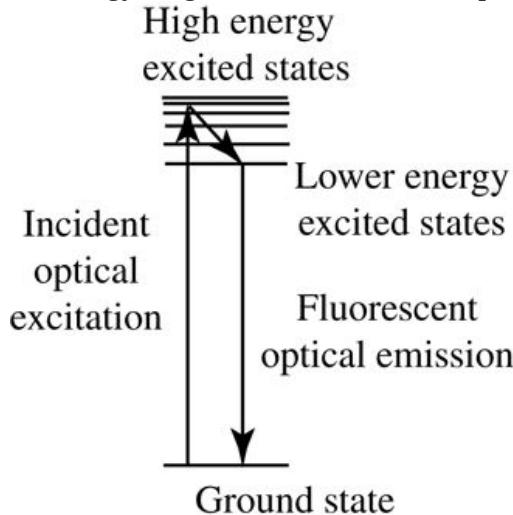


Figure 8.17

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 8.17 Energy band diagram illustrating fluorescent emission.

At the bottom of the illustration is a horizontal line representing ground state. At the top of the illustration are several closely packed parallel lines representing high energy excited states right above loosely packed parallel lines representing lower energy excited states. An upward pointing

arrow extending from ground state is the incident optical excitation. It passes halfway into the high energy excited states and then returns in a rightward downward slope to the lowest parallel line, from where it drops perpendicularly to the ground state as the fluorescent optical emission.

The absorption of an incident short-wavelength photon, which takes place on a time scale of about $10-15 \text{ } 10^{-15}$ seconds, creates an excited electron in the upper-level energy state. This transition is followed by a relaxation of the excited electrons to a lower energy state, which takes place on a time on the order of $10-12 \text{ } 10^{-12}$ seconds. The final process is emission of a longer wavelength photon, returning the molecule to the ground state, which takes place on a time on the order of $10-9 \text{ } 10^{-9}$ seconds. The emitted photon constitutes fluorescent emission.

W.E. Moerner and L. Kador were the first to apply spectroscopy to a single fluorescent molecule [257]. Stimulated by this achievement, the ability to tag interesting molecules (particularly biological molecules) with fluorophores then led to the field of fluorescent microscopy, some details of which will be explained in what follows.

Localization Precision

Suppose that it is desired to image a dilute set of molecules that have been tagged with fluorophores and emit mutually incoherent light when an optical excitation of an appropriate wavelength is applied. Each such point-source generates an intensity point-spread function of the imaging system. If the set of point source objects is sufficiently dilute, the probability of overlap of point-spread functions is small. If one is interested primarily in the locations of the molecules in the dilute array, then this set of locations can be identified with far greater resolution than the width of the point-spread function might imply. It has been shown [265] that while the so-called Abbe limit of resolution is given by

$$\Delta x = \lambda / 2n \sin \alpha,$$

$$\Delta x = \frac{\lambda}{2n \sin \alpha},$$

(8-42)

where α is the half-angle captured by the objective lens and n is the refractive index in the object space, one can nevertheless (neglecting finite pixel size and background noise) localize a point-source to an accuracy

$$\Delta x \approx \Delta x N,$$

$$\tilde{\Delta x} \approx \frac{\Delta x}{\sqrt{N}},$$

(8-43)

where N represents the total number of photons captured in the point-spread-function image of the point source. In effect, each detected photon gives a noisy estimate of the position of the point-source, and N such detected photons improve the localization, reducing the error by $1/\sqrt{N}$.

Stimulated Emission Depletion Microscopy (STED)

The fluorescence microscopy method known as STED was first developed by [Stephan Hell and colleagues \[166\] \[195\]](#) (Hell shared the 2014 Nobel Prize in Chemistry for discovering and demonstrating this technique). STED is a scanning microscopy technique, in which either the object is translated under a diffraction-limited spot of light, or the spot is scanned over a stationary object. The fundamental idea behind STED is to excite a fluorescent molecule with a proper wavelength to generate fluorescence, but at the same time to surround the diffraction-limited scanning spot by a donut-shaped spot of higher intensity and different wavelength that quenches the normal fluorescence by generating stimulated emission (of a different wavelength that can therefore be blocked by a filter) in a ring around the outer portions of the resolution element on the object. The result is that only a small region at the location of the scanning spot is actually contributing to the imaged light, and the location of the origin of the fluorescence can be identified with higher precision than the width of the point-spread function. While the standard diffraction limit of a fluorescence microscope operating at 400 nm with an NA=1.4 would be about 140 nm, STED resolutions below 60 nm have been demonstrated.

Photoactivated Localization Microscopy (PALM) and Stochastic Optical Reconstruction Microscopy (STORM)

The fluorescence microscopy method known as PALM was first developed by [Eric Betzig, Harold Hess, and colleagues \[27\] \[28\]](#). Betzig shared the 2014 Nobel Prize in Chemistry with S. Hell and W.E. Moerner. While STED microscopy is accomplished by scanning, PALM microscopy is a full-field imaging technique that is able to image dense sets of fluorescent molecules by imaging different dilute subsets of molecules over time. In PALM, photo-switchable protein fluorophores (such as photoactivatable green fluorescent protein (PA-GFP)) are used to genetically tag a protein of interest. The imaging process starts with the fluorescent molecules in an off state. The sample is then illuminated with photoactivation light of weak intensity, such that only a dilute random subset of the molecules is activated. The image of the dilute subset is captured, from which a high resolution image can be obtained by localization. After capturing the image, the sample is illuminated by a different wavelength with sufficient intensity to photobleach the activated molecules, thus permanently turning them off. The photoactivation light is again applied to turn on a new dilute set of fluorescent molecules, the image is again captured and localization is applied. The process is repeated enough times to accumulate the desired image in full. The images are combined to obtain a single high-resolution image.

A close relative of PALM is stochastic optical reconstruction microscopy (STORM), which differs from PALM primarily in the nature of the fluorescent tagging of the sample. STORM uses photo-switchable dyes to tag the protein of interest. A red laser turns all fluorophores off, then a green laser turns a dilute subset on and an image is captured. The red laser then turns all fluorophores off, and the green laser is applied again, exciting a different dilute set of fluorophores. The excited set varies randomly from image to image, and in this fashion a sequence of images of different sets of molecules is captured, to which localization is applied. The images are then combined to yield a complete super-resolution image. [Figure 8.18](#) shows an example of the resolution gain that can be obtained by these techniques, in this case by a variant of STORM. Resolutions on the order of 20 nm have been demonstrated with these techniques.

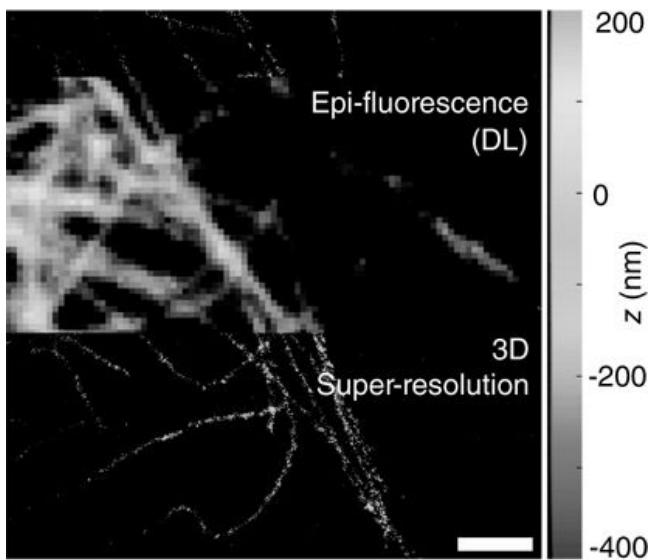


Figure 8.18

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 8.18 This figure shows a diffraction-limited image (upper left.) on a super-resolution image (lower left) of immunolabeled microtubules in a BSC-1 cell over a 14x14 micron field of view. Three-dimensional information was determined by the double-helix point spread function technique. Reproduced from Hsiao-lu D. Lee, Steffen J. Sahl, Matthew D. Lew, and W. E. Moerner, “The double-helix microscope super-resolves extended biological structures by localizing single blinking molecules in three dimensions with nanoscale precision,” *Appl. Phys. Lett.* **100**, 153701 (2012), with permission of AIP Publishing.

An illustration shows a square image along whose right edge is a scale ranging from minus 400 to + 200 in the bottom up direction. The scale is labeled z (nm). In the top right corner of the image, near the 100 mark, is the label “Epi-fluorescence (DL).” Next to the minus 200 mark is the label that reads, “3D super-resolution.” The image itself is a dark square with a few strings of bright spots emerging from the bottom right corner and rising to the top left corner where they form a tangle of bright streaks.

There are many variants of these techniques, some using different fluorophores, but space limitations prevent us from going further. For more information see [337] and [256].

8.5 Light Field Photography

The light field concept has its origins in the computer graphics community in the early 1990s [2]. The application of light field concepts to the light field or plenoptic camera was pioneered by Ren Ng in his Ph.D. thesis at Stanford University [264].

[Figure 8.19](#) shows a much simplified diagram of a light field camera. The object plane is imaged onto a lenslet array. Each lenslet then images the exit pupil of the primary lens onto a detector “super-pixel” consisting of a subarray of detector elements in a larger detector array. An individual lenslet can be thought of as gathering the light associated with one super-pixel in the image, while the detector elements associated with that super-pixel measure the strength of rays arriving from different points in the primary lens exit pupil, or equivalently with different angles of arrival. The coordinates (x_i, y_i) represent the coordinates of the i^{th} super-pixel center coordinates, while the coordinates (u, v) represent points in the exit pupil of the primary imaging lens and the coordinates (u_i, v_i) are coordinates in the image of the exit pupil formed by the i^{th} lenslet. Thus the full detector array measures the light intensities arriving in a four-dimensional coordinate system (x, y, u, v) . This set of measurements constitutes a measure of the light field arriving at the detector. As we shall see, from the measurement of the light field we can, within certain limits, reconstruct a focused image of any object plane in front of the lens, while retaining the depth of focus of that lens. Note that light fields are a geometrical optics concept, and the light field camera is an incoherent imaging system.

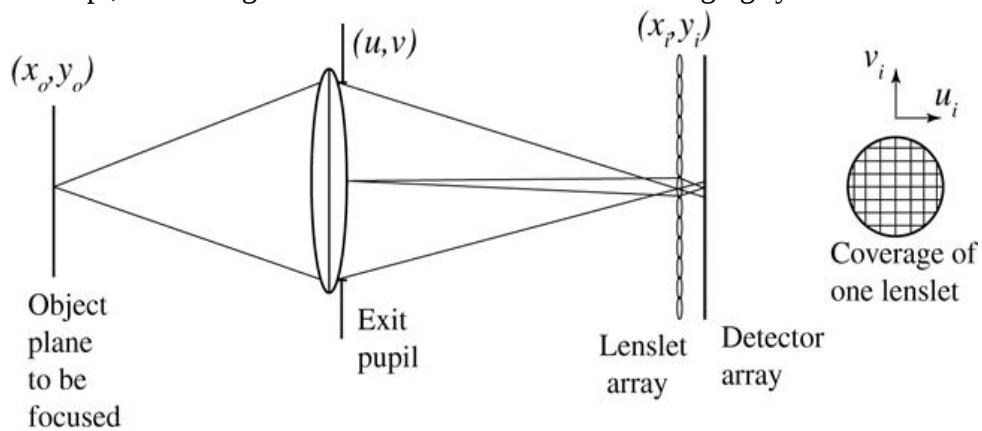


Figure 8.19

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 8.19 Geometry of the light field camera.

The illustration shows in the left extreme, at $(x \text{ subscript } o, y \text{ subscript } o)$, the object plane to be focused. From its center, rays diverge rightward to the extremes ends of a biconvex lens covering the exit pupil at (u, v) . From the pupil, the rays converge rightward to the center of the vertical lenslet array at $(x \text{ subscript } I, y \text{ subscript } i)$ and then diverge to the center of the detector array that is close behind. From the center of the exit pupil two rays are shown diverging to a very small

angle and reaching points on either sides of the center of the lenslet array and then converging behind it to the center of the detector array. An accompanying illustration shows coverage of one lenselet, which is a grid of horizontal and vertical lines set within a circle. Above the circle is an image of horizontal axis u subscript i and vertical axis v subscript i.

Visualizing light fields in four dimensions poses a challenge. It is much easier to gain understanding by considering two-dimensional light fields, one space coordinate x for pixel location and one coordinate u for angle of arrival. To understand the principles behind the light field camera, we need to understand how light fields are transformed in optical systems. Consider Fig. 8.20, which shows an original hypothetic light field and the effects on that light field due to Fourier transformation, forward propagation by distance z , and backward propagation by distance $-z$. Paraxial geometrical optics is assumed throughout. The region covered by the original light field is in essence a plot of the space/angle occupancy of the object, as captured by the imaging system. The effect of propagation is to shear the light field by an amount proportional to a product of the propagation distance and the angle of propagation. The effect of a Fourier transform is to rotate the light field by 90 degrees. The relations between x and x' are shown above the sheared light fields in the figure.

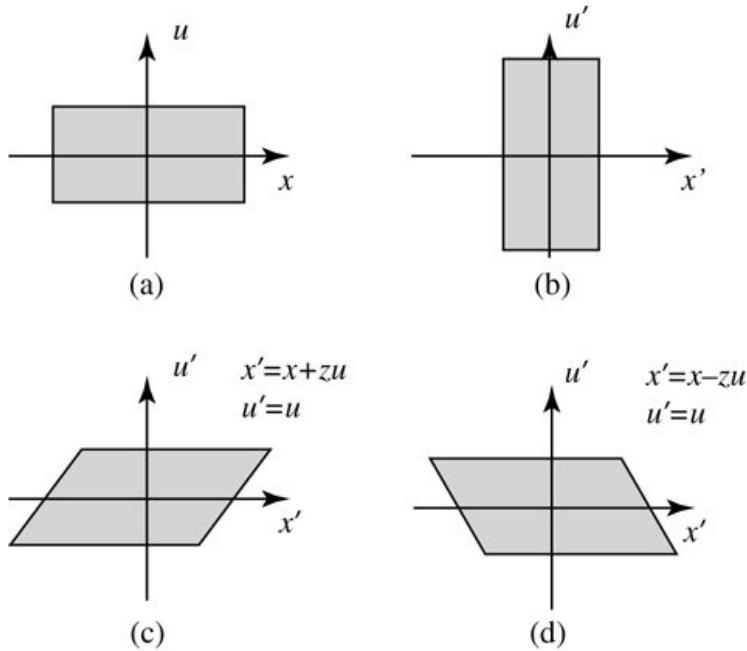


Figure 8.20
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 8.20 Light field transformations: (a) the original light field, (b) light field after a Fourier transform, (c) light field after forward propagation by distance z , (d) light field after backward propagation by distance $-z$.

Image a shows rightward horizontal axis x and upward vertical axis u . A rectangle is shown such that its longer sides are parallel to the horizontal axis and the intersection of its diagonals is at the origin. The remaining three images each show rightward horizontal axis x' and upward vertical axis u' . In image b, the rectangle is rotated such that the longer sides are parallel to the vertical axis. In image c, the longer sides are parallel to the horizontal axis and the shorter sides are upward sloping, giving it the appearance of a parallelogram. Two equations read, $x' = x + zu$ and $u' = u$. In image d, the longer sides are parallel to the horizontal axis and the shorter sides are downward sloping, giving it the appearance of a parallelogram. Two equations read, $x' = x - zu$ and $u' = u$.

x_u and u' = u . Image d is the same as image e but here the shorter sides are downwards sloping. Two equations read, $x' = x - x_u$ and $u' = u$.

Next we consider the effects of misfocus of the light field, illustrated in Fig. 8.21. On the left, the light field incident on the i^{th} super-pixel is shown for a point-source object when the lenslet array is located in the correct image plane. All angles arrive at the same sub-pixel in the super-pixel, shown by the dark box in the center of the lower left ((x_i, u_i)) light field representation. In the middle the object plane is closer to the lens than the ideal, in-focus object plane, and therefore the image appears below the detector array. The light field that occurs in the detector plane has in effect been back-propagated from the image plane to the detector plane, and hence the super-pixel detector array sees a light field as shown at the bottom center of the figure. Finally, on the right of the figure we show the opposite case, i.e. an object that is further from the lens than the ideal object plane, which results in an image closer to the lens than the detector plane. To find the light field incident on the detector in this case we must in effect forward propagate the light from the image plane to the detector plane, resulting in a light field at this detector super-pixel shown on the lower right of the picture. Note that we have ignored magnification and demagnification effects for simplicity. Clearly, in the center and right-hand cases, the detector does not record an in-focus image. However, it is possible to recover an in-focus image from the detected light field, as we next explain.⁷

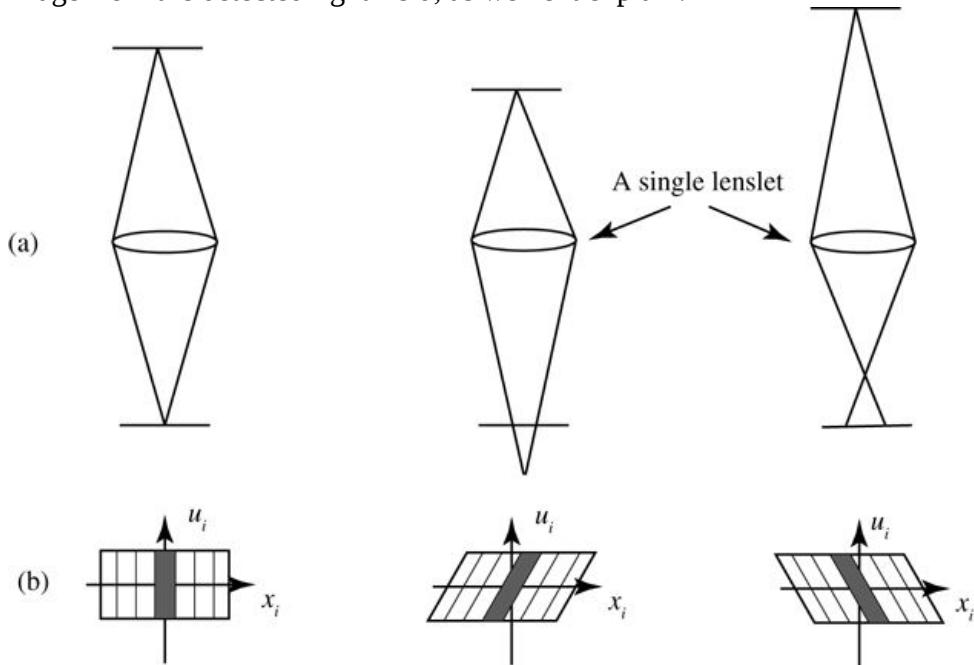


Figure 8.21

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 8.21 Effects at the i^{th} super-pixel of misfocus on the light field: (a) the imaging geometries, (b) the light fields at the detector plane.

Illustration a shows three figures. The first figure has two horizontal parallel lines equidistant from a horizontal lenslet. Rays from the center of the one parallel line diverge to the lenslet and then pass through it to converge at the center of the other parallel line. The second and the third figures are similar to the first one with some differences. In the second figure the upper parallel line is closer to the the single lenslet and the converging lines meet below the lower parallel line. In the

third figure the upper parallel line is farther from the the single lenslet and the converging lines meet above the lower parallel line and diverge to a short distance thereafter to the line.

Illustration b shows three figures. The first figures shows horizontal axis x_i and vertical axis u_i . A rectangle with its longer sides parallel to the horizontal axis is shown sectioned into 7 equal vertical rectangular parts. The part in the center is shaded and overlaps the vertical axis. The second and the third figures are similar to the first one with some differences. In the second one the rectangle slants to the right whereas in the third one it slants to the left.

To recover an image that is coincident with the detector array (the in-focus case), we simply need to add all the intensities detected at u_i -values that lie above and below each x_i coordinate and assign that sum to that x_i -coordinate in the image. That is, we project, at an angle in the u_i space when the image is not in focus, onto the x_i -axis. If this operation is performed on all super-pixels, the result is an in-focus image. Note that, according to the projection-slice theorem, projecting in the (x, u) space is equivalent to performing a two-dimensional Fourier transform of the function of (x, u) , slicing the spectrum at an angle normal to the projection direction, and performing a one-dimensional inverse Fourier transform of the spectral slice. In the actual four-dimensional case with coordinates (x, y, u, v) , the initial transform must of course be four-dimensional and the slice two-dimensional.

Problems - Chapter 8

1. 8-1. Show that for any rotating point-spread function that rotates $q \times 360^\circ$ degrees as z passes from $-\infty$ to ∞ , when the values of $\ln l_n$ are plotted versus p_n , the allowable pairs (\ln, p_n) lie on a line in the discrete plane passing through the origin and having slope $1/(q-1/2)$.
2. 8-2. It is claimed that a properly chosen set \square of Laguerre-Gaussian modes will exhibit "self-imaging" as they propagate down the z -axis. That is, if the intensity distribution of the beam is given by $I_0(x, y)$ at $z=0$, then that intensity distribution will be duplicated at discrete locations along the z -axis, without rotation and with changes only in transverse scale size.

1. Show that, for self imaging, the allowable values of indices (\ln, p_n) for the modes in the set \square must satisfy

$$|\ln| + 2p_n = M,$$

$$|l_n| + 2p_n = M,$$

(8-44)

where where M is an integer and $M > 4$.

2. Find the indices of the allowable modes in \square when $M=5$ and when $M=10$.
3. Find the z -axis locations of the self images in both cases.
3. 8-3. A telescope containing a Lyot coronagraph as depicted in Fig. 8.5 is pointed at a star with no planets. Let ρ_1 represent the radius of the primary collector, and let ρ_2 represent the radius of the image stop. The radius coordinate in the image plane is r_1 and radius coordinate in the Lyot-stop plane is r_2 . The focal length of the primary collector is f and the wavelength is λ_0 . Let ρ_2 be a radius such that it blocks 10 rings of the Airy pattern incident on the image plane, counting the central lobe as a ring. For simplicity of calculation, assume $\rho_1/\lambda_0 f = 1$.
 1. Find the required radius ρ_2 of the image-plane stop.
 2. Find an analytical expression for the light amplitude transmitted through the first image plane.
 3. Use any technique to plot the light intensity incident on the plane containing the Lyot stop.

4. 8-4. To suppress starlight in a preferred direction, as accomplished by the prolate spheroidal apodization shown in [Fig. 8.7](#), other apodization functions can be considered. For example, consider the *Nuttall apodization* defined by

$$g(x) = 0.355768 - 0.487396 \cos(\pi(x-1)) + 0.144232 \cos(2\pi(x-1)) - 0.012604 \cos(3\pi(x-1)),$$

$$\begin{aligned} g(x) &= 0.355768 - 0.487396 \cos(\pi(x-1)) \\ &\quad + 0.144232 \cos(2\pi(x-1)) - 0.012604 \cos(3\pi(x-1)), \end{aligned}$$

for $-1 \leq x \leq 1$, zero otherwise.

1. Plot the function $g(x)$ defined above and its first derivative with respect to x .
2. On a logarithmic plot, show the squared magnitude of the Fourier transform of this apodization function along the f_x axis, normalized to unity at the origin, and a similar logarithmic plot for the circular aperture of radius 1 with no apodization.
3. By what fraction has the total energy transmission been reduced by the presence of the apodization?

9 Wavefront Modulation

In many applications in optics, the ability to manipulate the spatial amplitude and phase distributions of light waves is needed. Such a capability is essential in holography, in creating systems with diffractive optical elements, and also in certain types of optical information processing. Therefore, in this chapter attention is focused on methods for spatially modulating optical wavefields, especially coherent fields. Historically the most common means of both detection and modulation has been through the use of photographic materials. However, in the 1990s, CCD detectors were sufficiently well developed to allow them to replace photographic materials in many image-capture applications. Ultimately, CCD and CMOS detectors replaced film in consumer cameras. In addition, spatial light modulators, based most commonly on either liquid crystals or micro-electromechanical devices (MEMs), became competitive with photographic film and plates in many applications in which optical wavefields are to be manipulated. As a consequence, photographic materials are much less important today than they have been historically. While the former major manufacturers of film and plates have in most cases left this business, nonetheless, both film and plates of various kinds can be obtained from smaller manufacturers in various countries around the world. Because of the historical importance of photographic materials in optics, as well as their continuing importance in many forms of holography, we present a short summary of their properties in [Section 9.1](#). In [Chapter 11](#) we consider holography in detail.

It is useful to distinguish between “fixed” masks that modulate the spatial distribution of light in a static way and “dynamic” devices for which the spatial modulation of the light can be changed rapidly in time. Photographic materials as well as lithographically defined masks are of the fixed type, while liquid-crystal devices, micro-mechanical electrically driven devices and acousto-optic cells are of the dynamic type. Very flexible and powerful optical systems can be realized if fixed masks are replaced by dynamic devices, although many applications do exist in which fixed masks are quite sufficient. We will discuss examples of both fixed masks and dynamic devices in what follows.

Synthesis of diffractive optical elements, holographic imaging, and optical information processing are application areas in which fixed masks play an important role. In [Section 9.2](#) we consider two approaches to constructing diffractive optical elements that control the complex amplitude of transmitted light in fixed but complicated ways. As their name implies, these elements control transmitted light through diffraction rather than refraction. Often a computer is employed in the design and construction of these elements, and their properties can be much more complicated than those of refractive elements.

9.1 Wavefront Modulation with Photographic Film

Photographic film and plate have historically been a basic component of imaging systems. Glass photographic plates were used in astronomy, high-energy physics, electron microscopy and medical imaging for many years. Photographic film and plates can play three very fundamental roles in optics. First, they can serve as a detector of optical radiation, a task they perform remarkably efficiently. Second, they can serve as a storage medium for images, capable of retaining information for long periods of time. Third, they can serve as a spatial modulator of transmitted or reflected light, a role of particular importance in optical information processing. All of these functions are achieved at extremely low cost.

Because of the importance photographic film has played in optics in general, as well as the importance it continues to play in holography, we devote some time here to discussing its properties. For a more comprehensive treatment of the photographic process, see, for example, [247]. Other useful references include [323] and [29].

9.1.1 The Physical Processes of Exposure, Development, and Fixing

An unexposed photographic film or plate generally consists of a very large number of tiny silver halide (often AgBr) grains suspended in a gelatin support, which in turn is attached to a firm “base” consisting of acetate or mylar¹ for films, and glass for plates. The soft emulsion also has a thin layer of a protective overcoating on its exposed surface, as illustrated in the cross section shown in Fig. 9.1. In addition, certain sensitizing agents are added to the gelatin; these agents have a strong influence on the introduction of dislocation centers within the silver halide crystals. Light incident on the emulsion initiates a complex physical process that is outlined as follows:

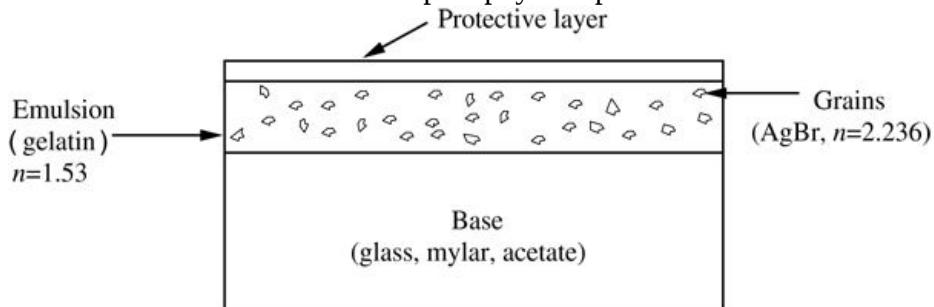


Figure 9.1

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 9.1 Structure of a photographic film or plate.

The rectangular cross section shows three layers. The top most layer, also the thinnest, is the protective layer. Below it is a layer of gelatin emulsion, $n = 1.53$, embedded in it are irregular shaped grains (Ag Br, $n = 2.236$). The third layer, taking up more than half the cross-section, is the base consisting of glass, mylar, and acetate.

1. A photon incident on a silver halide grain may or may not be absorbed by that grain. If it is absorbed, an electron-hole pair is released within the grain.
2. The resulting electron is in the conduction band, is mobile within the silver halide crystal, and eventually, with some probability, becomes trapped at a crystal dislocation.
3. The trapped electron electrostatically attracts a silver ion; such ions are mobile even before exposure by light, a consequence of thermal agitation.
4. The electron and the silver ion combine to form a single atom of metallic silver at the dislocation site. The lifetime of this combination is rather short, of the order of a few seconds.
5. If within the lifetime of this first silver atom, a second silver atom is formed by the same process at the same site, a more stable two-atom unit is formed with a lifetime of at least several days.
6. Typically at least two additional silver atoms must be added to the silver speck in order for it ultimately to be developable. The existence of a threshold, requiring several trapped electrons to activate the development process, is responsible for good stability of unexposed film on the shelf.

The speck of silver formed as above is referred to as a *development speck*, and the collection of development specks present in an exposed emulsion is called the *latent image*. The film is now ready for the development and fixing processes.

The exposed photographic transparency is immersed in a chemical bath, the developer, which acts on silver specks containing more than the threshold number² of silver atoms. For such grains, the developer causes the entire crystal to be reduced to metallic silver. The ratio of the number of silver atoms in a developed grain to the number of photons that must be absorbed to make the grain developable is typically of the order of 10^9 10^9 , a number which is often called the “gain” of the photographic process.

At this point the processed emulsion consists of two types of grains, those that have been turned to silver, and those that did not absorb enough light to form a development center. The latter crystals are still silver halide and, without further processing, will eventually turn to metallic silver themselves simply through thermal processes. Thus in order to ensure stability of the image, it is necessary to remove the undeveloped silver halide grains, a process called *fixing* the emulsion. The transparency is immersed in a second chemical bath, which removes the remaining silver halide crystals from the emulsion, leaving only the stable metallic silver.

The processes of exposure, development and fixing are illustrated in [Fig. 9.2](#).

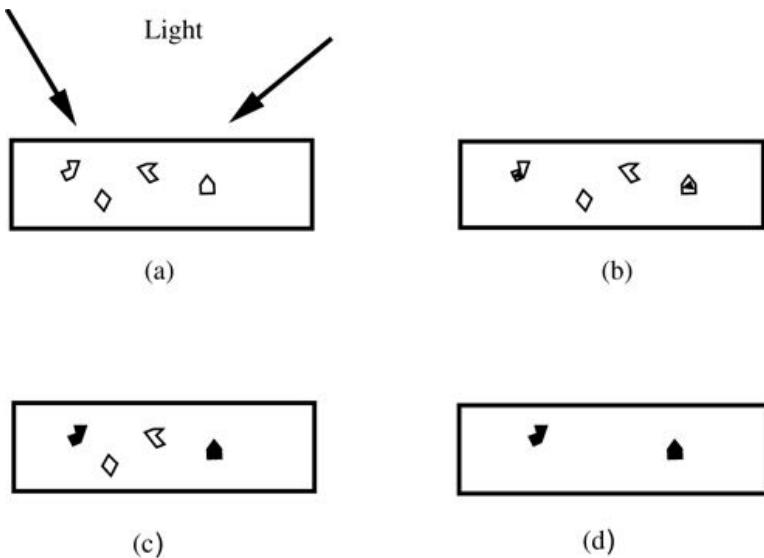


Figure 9.2
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 9.2 Pictorial representation of the photographic process. (a) Exposure, (b) latent image, (c) after development, and (d) after fixing. Only the emulsion is shown.

All four images are identical rectangles with tiny grains in each. Illustration a shows four tiny irregular shaped colorless grains, downward arrows from above the rectangle indicate exposure to light. In illustration b, two of the grains have partially turned black. In illustration c, the partially blackened grains are now fully black, while the other two grains remain unchanged. In illustration d, only the blackened grains remain while the non-blackened ones have disappeared.

9.1.2 Definition of Terms

The field of photography has developed a certain nomenclature that should be mastered if the properties of photographic emulsions are to be discussed in any detailed way. At this point we introduce the reader to some of these terms.

Exposure. The energy incident per unit area on a photographic emulsion during the exposure process is called the exposure. Represented by the symbol E , it is equal to the product of incident intensity \mathcal{J} at each point and the exposure time T ,

$$E(x,y) = \mathcal{J}(x,y)T.$$

$$E(x, y) = \mathcal{J}(x, y)T.$$

The units for exposure are mJ/cm^2 . Note that the symbol \mathcal{J} is used for intensity incident on the film during exposure; we reserve the symbol I to represent intensity incident on (or transmitted by) the transparency after development.

Intensity transmittance. The ratio of intensity transmitted by a developed transparency to the intensity incident on that transparency, averaged over a region that is large compared with a single grain but small compared with the finest structure in the original

exposure pattern, is called the intensity transmittance. Represented by the symbol τ , it is equal to

$$\tau(x, y) = \text{local average } I_{\text{transmitted at } (x, y)} / I_{\text{incident at } (x, y)}.$$

$$\tau(x, y) = \frac{\text{local average} \left\{ I_{\text{transmitted at } (x, y)} \right\}}{I_{\text{incident at } (x, y)}}.$$

Photographic density. In the year 1890, F. Hurter and V.C. Driffield published a classic paper in which they showed that the logarithm of the reciprocal of the intensity transmittance of a photographic transparency should be proportional to the silver mass per unit area of that transparency. They accordingly defined the photographic density D as

$$D = \log_{10} \tau.$$

$$D = \log_{10} \left(\frac{1}{\tau} \right).$$

The corresponding expression for intensity transmittance in terms of density is

$$\tau = 10^{-D}.$$

$$\tau = 10^{-D}.$$

Hurter-Driffield curve. The most common description of the photometric properties of a photographic emulsion is the Hurter-Driffield curve, or the H&D curve, for short. It is a plot of photographic density D versus the logarithm of the exposure E that gave rise to that density. A typical H&D curve is shown in Fig. 9.3 for the case of a photographic negative.

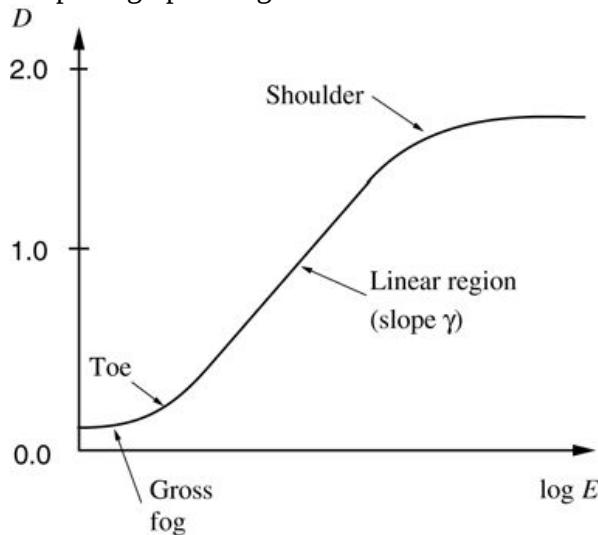


Figure 9.3

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 9.3 The Hurter-Drifford curve for a typical emulsion.

The graph plots $\log E$ along the horizontal line and D along the vertical line marked from 0 to 2. The S shaped curve begins near the origin on the vertical axis and rises gently upward for about one sixth of its length, bowed out toward the horizontal axis. This part is labeled “Gross fog,” at the end of which is a part where the curve begins to go steeper and is labeled “Toe.”

The curve then follows a steep, straight line upward path for a distance approximately half its entire length; this section is labeled “Linear region (slope gamma)” and extends up to a level that approximately matches the 1.5 mark on the vertical axis. Thereafter about a third of the curve begins to bow out toward the vertical axis and then go almost parallel to the horizontal axis. This part is labeled shoulder.

Note the various regions of the H&D curve. When the exposure is below a certain level, the density is independent of exposure and equal to a minimum value called *gross fog*. In the *toe* of the curve, density begins increasing with exposure. There follows a region of considerable extent in which the density is linearly proportional to the logarithm of exposure—this is the region most commonly used in ordinary photography. The slope of the curve in this linear region is referred to as the *gamma* of the emulsion and is represented by the symbol γ . Finally the curve saturates in a region called the *shoulder*, beyond which there is no change of density with increasing exposure. In the linear region of the H&D curve where the slope is γ , the relationship between intensity transmission and exposure is given by

$$\tau = kE^{-\gamma},$$

$$\tau = kE^{-\gamma},$$

where k is a constant while γ is positive for a negative transparency and negative for a positive transparency.

More relevant than the H&D curve in optical information processing and holography is a direct relationship between the light *complex amplitude* incident on the developed film or plate and the light amplitude transmitted by that transparency. We consider that relationship for coherent optical systems in the next section.

9.1.3 Photographic Film or Plate in Coherent Optical Systems

When film is used as an element of a *coherent* optical system, it is most appropriately regarded as providing either (1) a mapping of intensity incident during exposure into complex field transmitted after development, or (2) a mapping of complex amplitude incident during exposure into complex amplitude transmitted after development. The second viewpoint can be used, of course, only when the light that exposes the transparency is itself coherent, and must incorporate the fact that all phase information about the incident complex wavefield is lost upon detection. Only when interferometric detection is used can phase information be captured, and such detection systems will be seen to benefit from the first viewpoint, rather than the second.

Since the complex amplitude of the transmitted light is, from both viewpoints, the important quantity in a coherent system, it is necessary to describe a transparency in terms of its complex amplitude transmittance t_A [221]. It is most tempting to define t_A simply as the positive square root of the intensity transmittance $t_A = \sqrt{\tau}$. However, such a definition neglects the relative phase shifts that can occur as the light passes through the film [178]. Such phase shifts arise as a consequence of variations of the film or plate thickness, or from internal refractive-index

changes resulting from the exposure pattern. Thickness changes can originate in two distinct ways. First, there are generally random thickness variations across the base of the film, i.e. the base is not optically flat. Second, the thickness of the emulsion is often found to vary with the density of the silver in the developed transparency. This latter variation is strongly dependent on the exposure variations to which the film has been subjected. Internal refractive index changes can also be present, and indeed can be emphasized by *bleaching* of the emulsion, a subject to be discussed shortly. Thus a complete description of the amplitude transmittance of the film must be written

$$tA(x,y)=\tau(x,y)\exp[j\phi(x,y)]$$

$$t_A(x, y) = \sqrt{\tau(x, y)} \exp [j\phi(x, y)]$$

(9-1)

where $\phi(x,y)$ describes the pattern of phase shifts introduced by the transparency.

In most applications, the thickness variations are entirely undesired, for they cannot easily be controlled. It is possible to remove the effects of these variations by means of a device called a *liquid gate*. Such a device consists of two pieces of glass, each ground and polished to be optically flat on one side, between which the transparency and an index matching fluid (often oil) can be sandwiched, as illustrated in [Fig. 9.4](#). The flat surfaces of the glass are, of course, facing the outside, and the index of refraction of the fluid must be chosen as a compromise, for it is impossible to match simultaneously the different indices of the base, the emulsion, and the glass. However, with a proper choice of fluid, the optical path length through the liquid gate can be made nearly constant, allowing the amplitude transmittance of the film and gate to be written

$$tA(x,y)=\tau(x,y).$$

$$t_A(x, y) = \sqrt{\tau(x, y)}.$$

(9-2)

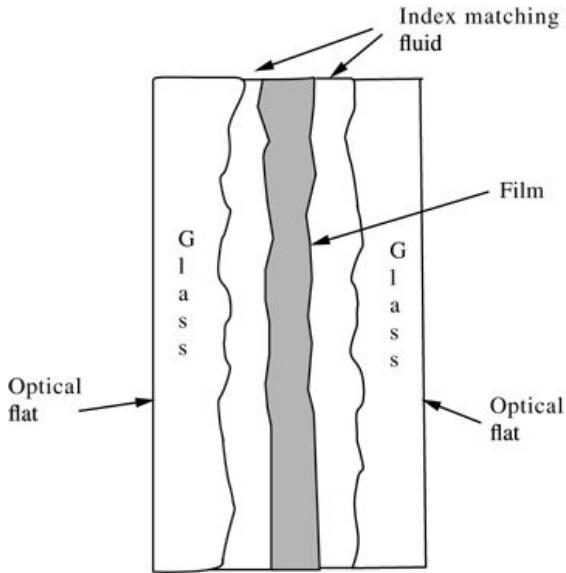


Figure 9.4

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 9.4 A liquid gate for removing film thickness variations. The thickness variations are greatly exaggerated.

The illustration shows a magnified cross section of a photographic film the roughness of whose two sides is represented by uneven lines. The film is sandwiched between uneven layers of index matching fluid followed by sheets of optical glasses. The inner surface of the optical glass is uneven while the outer surface is polished and flat.

As will be seen in many of the examples to be discussed in later sections, it is often desirable to have film act as a *square-law* mapping of complex amplitude. It is possible, to obtain square-law action over a limited dynamic range with a transparency of any gamma, be it a positive or a negative. This point is most easily seen by abandoning the traditional H&D curve description of film and making instead a direct plot of amplitude transmittance versus exposure (on a linear scale). Such a description was advocated at an early stage by Maréchal and was very successfully used by [Kozma \[210\]](#) in an analysis of the effects of photographic nonlinearities. [Figure 9.5](#) shows a plot of amplitude transmittance versus exposure (the $t_A - E$ curve) for a typical negative transparency. If the film is “biased” to an operating point that lies within the region of maximum linearity of this curve, then over a certain dynamic range the film will provide a square-law mapping of incremental changes in incident amplitude into incremental changes of amplitude transmittance. Thus if E_b represents the bias exposure and t_b the corresponding bias amplitude transmittance, we may represent the $t_A - E$ curve within its region of linearity by

$$t_A \approx t_b + \beta(E - E_b) = t_b + \beta'|\Delta U|^2$$

$$t_A \approx t_b + \beta(E - E_b) = t_b + \beta'|\Delta U|^2$$

(9-3)

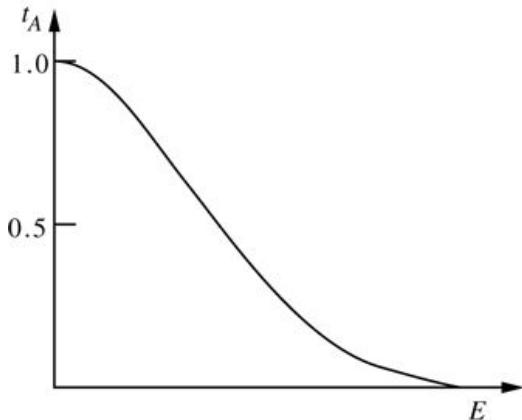


Figure 9.5

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 9.5 Typical amplitude transmittance versus exposure curve.

where β is the slope of the curve at the bias point, $\Delta U \Delta U$ represents the incremental amplitude changes, and $\beta' \beta'$ is the product of $\beta \beta$ and the exposure time. Note that $\beta \beta$ and $\beta' \beta'$ are negative numbers for a negative transparency.

In general, a high-gamma film has a steeper slope to its $t_A - E$ curve than does a low-gamma film and therefore is more efficient in transferring small changes of exposure into changes of amplitude transmittance. However, this increased efficiency is often accompanied by a smaller dynamic range of exposure over which the $t_A - E$ curve remains linear. As an additional point of interest, the bias point at which maximum dynamic range is obtained is found to lie in the toe of the H&D curve.

Before closing this section we note that when thin gratings are recorded by interference in a photographic emulsion, as is often the case in the construction of spatial filters and in recording holograms, it may often be desirable to achieve the highest possible diffraction efficiency, rather than the widest possible dynamic range. It can be shown (see [323], page 7) that, for small modulations, the maximum diffraction efficiency for a thin sinusoidal grating recorded photographically will occur for a recording made in the region where the magnitude of the slope α of the t_A versus $\log E$ curve of the emulsion is maximum. This curve is yet another description of the properties of photographic emulsions that is relevant in some applications.

9.1.4 The Modulation Transfer Function

To this point we have tacitly assumed that any variations of exposure, however fine on a spatial scale, will be transferred into corresponding variations of silver density according to the prescription implied by the H&D curve. In practice, one finds that when the spatial scale of exposure variations is too small, the changes of density induced may be far smaller than would be implied by the H&D curve. We can say in very general terms that each given type of film has a limited spatial frequency response.

The spatial frequency response of an emulsion is limited by two separate phenomena:

1. Light scattering within the emulsion during exposure.
2. Chemical diffusion during the development process.

Both of these phenomena are linear ones, although the physical quantities with respect to which they are linear are different. Light scattering is linear in the variable exposure, while chemical diffusion is linear in the variable density. It might be hoped that the linear phenomena that limit spatial frequency response could be separated from the highly nonlinear behavior inherent in the H&D curve. This in fact can be done by regarding the photographic process as a cascade of several separate mappings, as illustrated in Fig. 9.6. The first operation in this cascade is a linear, invariant filter representing the effects of light scattering and the resulting spread or blur of the exposure pattern E . The output of this filter, E' , then passes through the H&D curve, which is regarded as a *zero-spread nonlinearity*, analogous to the zero-memory nonlinearities often encountered in the analysis of communications systems. The output of the H&D curve is a density D' , which is itself subjected to linear spreading and blur by the chemical diffusion process to produce a final density D . This model is often referred to as the “Kelley model,” after D.H. Kelley who originated it. Often the model is simplified to include only a single linear filter that precedes the nonlinear H&D curve, thus ignoring the linear filter associated with diffusion.

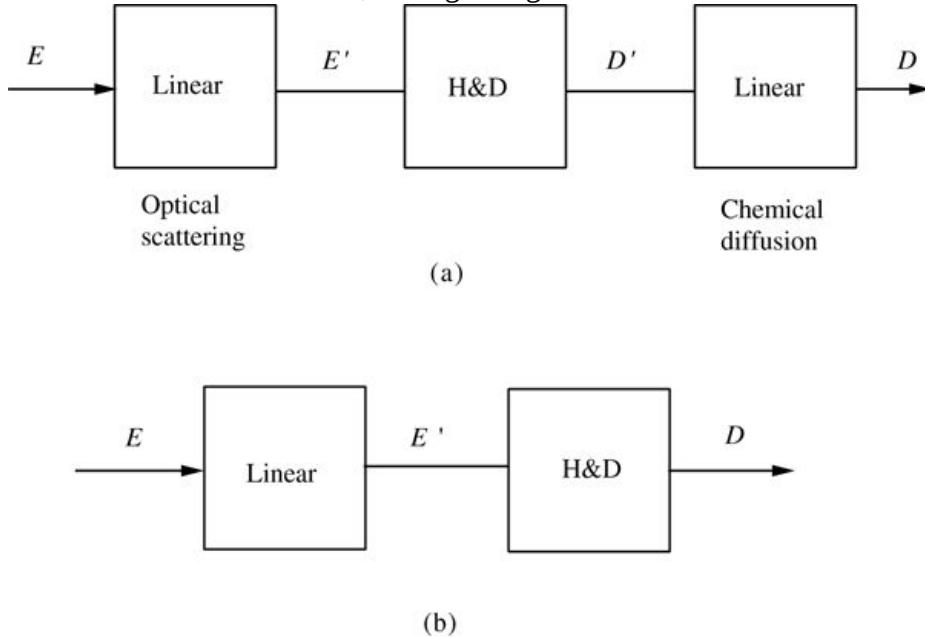


Figure 9.6
Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 9.6 The Kelley model of the photographic process. (a) Full model; (b) simplified model.

Mapping a shows a series of three identical squares connected by a single rightward arrow. The squares are Linear (Optical scattering), H and D, and Linear (Chemical diffusion). The connecting arrow segments are labeled to indicate inputs and outputs. The mapping thus reads input E, Linear, Output E dash, H and D, output D dash, Linear, output D. Mapping b shows two identical squares, namely, Linear and H and D, connected by a single rightward arrow. The connecting arrow segments are labeled to indicate inputs and outputs. The mapping thus reads input E, Linear, Output E dash, H and D, output D.

The effects of the linear filters are, of course, to limit the spatial frequency response of the emulsion. If the model is simplified to one with a single linear filter preceding the nonlinear mapping ([Fig. 9.6\(b\)](#)), then it is of some interest to find the transfer function of the filtering operation, usually referred to as the *modulation transfer function* of the photographic process. To measure the characteristics of the linear filter, a cosinusoidal exposure pattern

$$E = E_0 + E_1 \cos 2\pi f x$$

$$E = E_0 + E_1 \cos 2\pi f x$$

(9-4)

can be applied (such a pattern is easily generated by interference of two mutually coherent plane waves on the emulsion). The “modulation” associated with the exposure is defined as the ratio of the peak variation of exposure to the background exposure level, or

$$M_i = \frac{E_1}{E_0}$$

$$M_i = \frac{E_1}{E_0}$$

(9-5)

If the variations of density in the resulting transparency are measured, they can be referred back to the exposure domain through the H&D curve (assumed known) to yield an inferred or “effective” cosinusoidal exposure pattern, as indicated in [Fig. 9.7](#). The modulation M_{eff} of the effective exposure distribution will always be less than the modulation M_i of the true exposure distribution. Accordingly the modulation transfer function of the film is defined as

$$M(f) = M_{\text{eff}}(f) M_i(f)$$

$$M(f) = \frac{M_{\text{eff}}(f)}{M_i(f)}$$

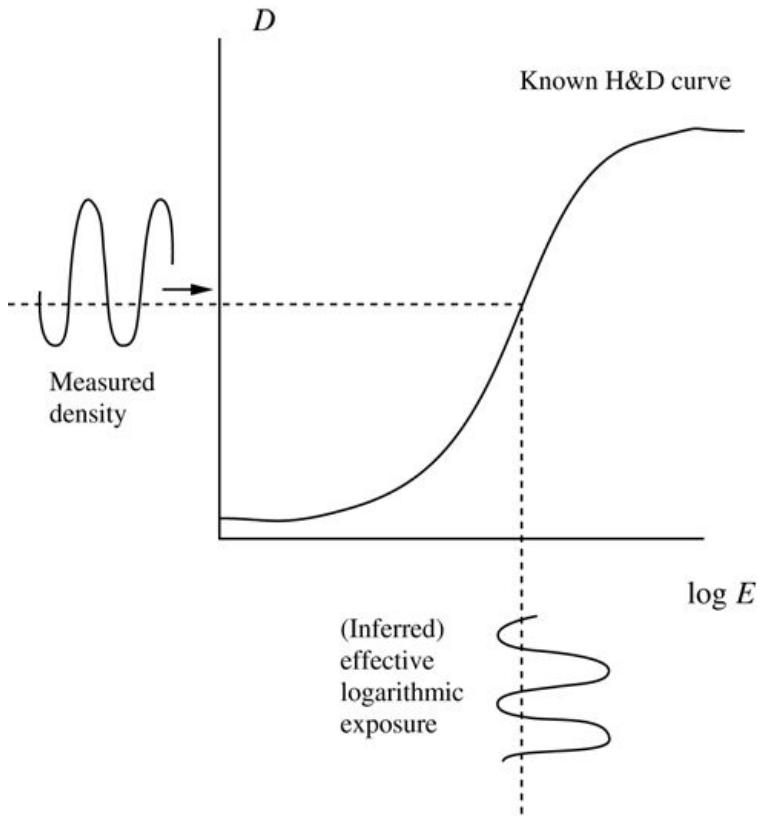


Figure 9.7

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 9.7 Measurement of the MTF by projecting back through the H&D curve.

The graph plots $\log E$ values along the horizontal axis and D values along the vertical axis. The S shaped Known H D and D curve begins near the origin on the vertical axis and rises gently upward for about a quarter of its length, bowed out toward the horizontal axis. It then becomes steep for a distance approximately half its entire length. The final quarter of the curve is a smooth, upward, and bowed out toward the vertical axis. From the center of the curve a dotted perpendicular is dropped to the horizontal axis and extended beyond it; in the extended part a wavelike curve is drawn and labeled “(Inferred) effective logarithmic exposure.” Similalry, from the center of the curve a dotted perpendicular is dropped to the vertical axis and extended beyond it; in the extended part a wavelike curve is drawn and labeled “Measure density.” The curve appears with a rightward arrow.

where the dependence on the spatial frequency f^f of the exposure has been emphasized. In most cases encountered in practice, the form of the point-spread function of the scattering process (approximately Gaussian) is circularly symmetric and there are no phase shifts associated with the transfer function. The effective exposure distribution applied to the nonlinear portion of the film mapping may therefore be written

$$E' = E_0 + M(f) E_1 \cos 2\pi f x.$$

$$E' = E_0 + M(f) E_1 \cos 2\pi f x.$$

(9-6)

Figure 9.8 illustrates the typical measured frequency dependence of the MTF of an emulsion, plotted versus radial spatial frequency ρ . The small hump rising above unity at low frequencies is caused by chemical diffusion (the final linear filtering box in our model, which was ignored in the procedure for measuring the MTF) and is referred to as arising from the *adjacency effect*.

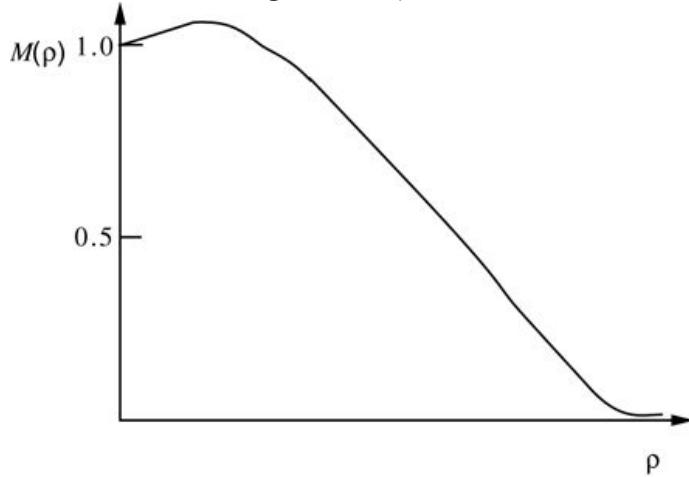


Figure 9.8

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 9.8 Typical measured MTF curve.

The graph plots ρ along the horizontal axis and $M(\rho)$ along the vertical axis marked from 0 to 1. It shows a downward sloping curve beginning at $(0, 1)$, rising only marginally before taking a downward slope reaching the right extreme of the horizontal axis.

The range of frequencies over which significant frequency response is obtained varies widely from emulsion to emulsion, depending on grain size, emulsion thickness, and other factors. High-resolution photographic plates of the kind used in holography have MTFs with significant value beyond 2000 line-pairs/mm, which implies a spatial resolution of at least 250 nm.

9.1.5 Bleaching of Photographic Emulsions

Conventional photographic emulsions modulate light primarily through absorption caused by the metallic silver present in the transparency. As a consequence, significant amounts of light are lost when an optical wave passes through such a spatial modulator. In many applications it is desired to have a more efficient modulator, one that can operate primarily through *phase modulation* rather than absorption. Such structures can be realized with photographic materials, provided they are subjected to *chemical bleaching*.

The bleaching process is one that removes metallic silver from the emulsion and leaves in its place either an emulsion thickness variation or a refractive index variation within the emulsion. The chemical processes that lead to these two different phenomena are in general different. A thickness variation results when a so-called *tanning bleach* is used, while a refractive index modulation occurs when a *nontanning bleach* is used.

Considering first the tanning bleach, the chemical agents used in this type of bleach release certain chemical byproducts as they remove the metallic silver, and these byproducts cause a cross-linking of the gelatin molecules within the emulsion in regions where the silver concentration was high. As the transparency is dried, the hardened areas shrink less than do the

unhardened areas, with the result that a *relief image* is formed, with the thickest regions of the emulsion being where the density was highest, and the thinnest regions where the density was lowest. [Figure 9.9](#) illustrates the phenomenon for the case of a square-wave density pattern. This phenomenon is found to depend strongly on the spatial frequency content of the density pattern and to act as a bandpass filter, with no relief induced at very low spatial frequencies and at very high spatial frequencies. For a $15\text{-}\mu\text{m}$ -thick emulsion, the peak thickness variations are found to occur at a spatial frequency of about 10 cycles/mm, with a maximum relief height in the $1\text{-}2\text{-}\mu\text{m}$ range. Using such a bleach it is possible, for example, to make an approximately sinusoidal relief grating, which will exhibit diffraction efficiencies typical of sinusoidal phase gratings, considerably higher than those of sinusoidal amplitude gratings.

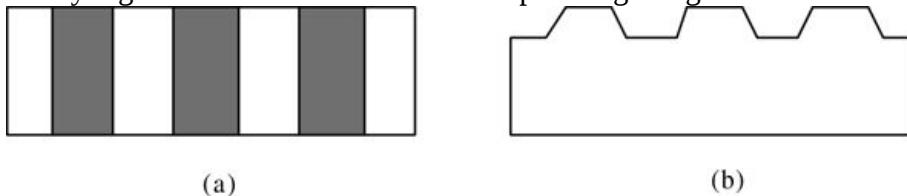


Figure 9.9

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 9.9 A relief image produced by a tanning bleach. (a) Original density image. (b) Relief image after bleaching.

Illustration a shows a rectangle with its longer sides oriented horizontally. It is sectioned into a series of seven identical rectangular parts, their longer sides oriented vertically, and the second, third, and fourth rectangles being shaded dark. Illustration b shows a rectangle with its longer sides horizontally oriented. The top horizontal side has three identical projections, each in the shape of an inverted cup. That is, each begins with an upward slope then a horizontal line followed by a downward slope as long as the initial upward slope. The horizontal side of the projection is in line with the upper side of the rectangle in image a.

Nontanning bleaches, on the other hand, produce internal refractive index changes within the emulsion, rather than relief images. For such bleaches, the metallic silver within the developed transparency is changed back by the chemical bleach to a transparent silver halide crystal, with a refractive index considerably larger than that of the surrounding gelatin. In addition, the bleach must remove the sensitizing agents found in unexposed silver halide crystals to prevent them from turning to metallic silver due to thermal effects and additional exposure to light. The resulting refractive index structures constitute a pure phase image. The spatial frequency response of this kind of bleached transparency is not a bandpass response, but rather is similar to that of the original silver image. Very high-frequency phase structures can be recorded using this method.

Phase shifts of the order of 2π radians can be induced in a wavefront passing through the bleached emulsion, although this number obviously depends on the emulsion thickness.

9.2 Wavefront Modulation with Diffractive Optical Elements

The vast majority of optical instruments in use today use *refractive* or *reflective* optical elements (e.g. lenses, mirrors, prisms, etc.) for controlling the distribution of light. In some cases it is possible to replace refractive or reflective elements with *diffractive* elements, a change that can lead to some significant benefits in certain applications. Diffractive optics can be made to perform functions that would be difficult or impossible to achieve with more conventional optics (e.g., a single diffractive optical element can have several or many different focal points simultaneously). Diffractive optical elements also generally have much less weight and occupy less volume than their refractive or reflective counterparts. They may also be less expensive to manufacture and in some cases may have superior optical performance (e.g. a wider field of view). Examples of applications of such components include optical heads for compact disks, beam shaping for lasers, grating beamsplitters, and reference elements in interferometric testing.

Along with these several advantages comes one significant difficulty with diffractive optical components: because they are based on diffraction, they are highly dispersive (i.e. wavelength sensitive). For this reason they are best applied in problems for which the light is highly monochromatic. Such is the case for most coherent optical systems. However, diffractive optics can be used together with either refractive optics or additional diffractive elements in such a way that their dispersive properties partially cancel (cf. [331], [261], [108]), allowing their use in systems for which the light is not highly monochromatic.

For additional background on diffractive optics, the reader may wish to consult review articles [338], [107], and Vol. 2, [Chapter 8](#) of [17].

The tremendous advances in the ability to lithographically define and image exquisitely fine patterns onto various materials, as driven by the needs of the integrated circuit industry, has opened opportunities for the construction of “fixed,” non-silver-halide optical elements. Various types of diffractive optical elements can be made. The simplest types of structures are gratings, which may have a fixed period or a period that varies over space. Square wave phase gratings, which have higher diffraction efficiency than square wave amplitude gratings, are common and require only a single lithographic exposure. A diffractive optical lens can be made by such a process provided the period of the grating is properly changed over space. Often more complicated structures are desired, such as a sawtooth phase grating which can have nearly 100% diffraction efficiency if made properly. Such a structure can be made by a process requiring a single lithographic exposure or by a process that requires multiple lithographic exposures. In what follows, we outline these two approaches to making general diffractive optical elements without the use of silver halide materials. Our focus will be on pure phase diffractive elements because of their superior diffraction efficiency.

9.2.1 Single Step Lithography

A process that requires only a single lithography step can be used to make structures that have a wide variety of different phase profiles. The process begins with a glass plate with a thin layer of chromium, on top of which a positive photoresist layer has been spun. By means of electron beam writing or laser writing, a binary pattern is written into the photoresist; the photoresist is then developed, leaving exposed portions of the chromium mask. The exposed regions of chromium

can now be etched away using an appropriate solvent. In this fashion a binary master of chromium on glass is created.

A thin layer of photoresist is now spun onto another substrate, again usually glass. The master is illuminated optically, and the photoresist is exposed through the master. For a positive photoresist, the plate is chemically processed (developed) to remove photoresist where it has been exposed, in proportion to the amount of exposure. While photoresist is often regarded as a binary recording material, i.e. it produces only a binary relief pattern, nonetheless, many photoresists have a region of exposure where the amount of photoresist removed during processing varies with exposure. [Figure 9.10](#) shows a hypothetical plot of the fraction of photoresist removed versus the exposure dose given to the photoresist. Note that unlike the case of silver-halide materials, for which the symbol D refers to silver density, in lithography the symbol D is used to represent exposure dose. The symbol D represents the dose where the two dotted lines meet, and corresponds to the dose where no photoresist is removed after development. The symbol D_0 represents the dose that removes 100% of the photoresist after development. A region in which the remaining fraction of photoresist is linearly proportional to the logarithm of dose is evident. The contrast D_{100}/D_0 of the resist is defined to be the slope of the linear part of the curve. The contrast achieved depends on the particular photoresist used, the details of the development process, and the wavelength of the exposure. For more details on photoresist and its patterning, see [50] and [271], for example.

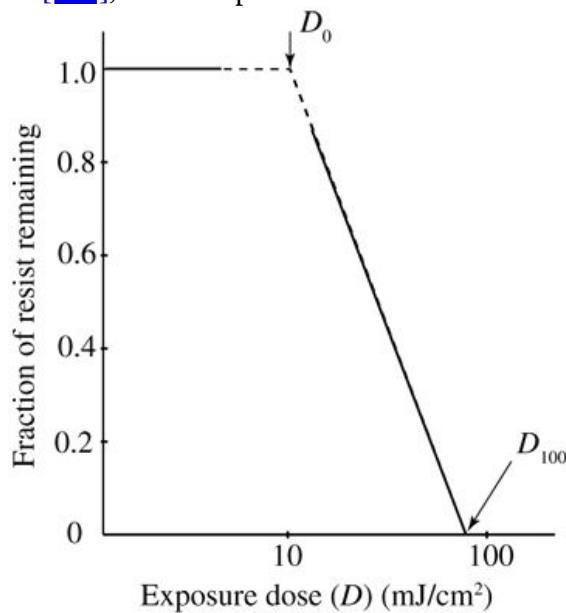


Figure 9.10

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 9.10 Fraction of the photoresist remaining versus the exposure dose D (positive resist assumed).

The graph plots exposure dose (D) (milli joules per centimeter squares) along the horizontal axis marked from 0 to 100 and fraction of resist remaining along the vertical axis marked from 0 to 1. The graph includes a horizontal solid line extending from $(0, 1)$ to around $(6.5, 1)$, a point labeled D subscript 0, where a dotted line begins and extends up to $(10, 1)$, where it slants downward and reaches in a straight line slope a point marked D subscript 100 located around the $(90, 0)$ mark.

Another solid line begins around (11, 0.85) and stretches up to (90, 0) where the slanting dotted line meets the horizontal axis. The dotted line almost completely overlaps the solid line.

The binary master can be thought of as consisting of a discrete set of cells, each of a size roughly corresponding to the resolution of the lithography imaging system to be used. Given that the electron beam or laser writing system used to pattern the master has finer resolution than the lithographic imaging system that follows, within each cell on the master it is possible to write one or more binary subcells, a different set for each analog value of exposure the lithography system should deliver to the photoresist. An example is shown in [Fig. 9.11](#). This process is analogous to the half-toning process used in the printing industry. In this way, a set of analog (rather than binary) exposures is delivered to the photoresist, and a set of analog relief heights is produced. Because a distinct set of binary representations of the analog values desired in the relief pattern is used in the master cells, the relief pattern is generally quantized to a finite set of values. The subsection that follows discusses the approximation of a sawtooth phase grating to a set of quantized values.

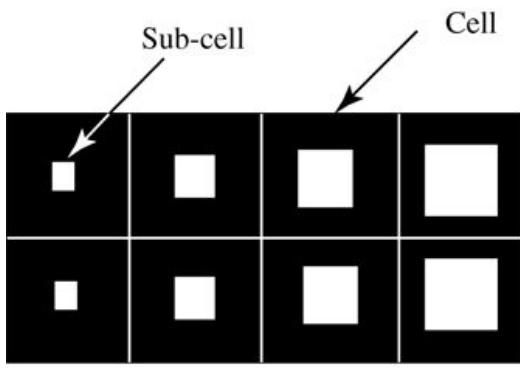


Figure 9.11

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 9.11 Example of two rows of a four-level mask pattern.

The refractive index of photoresists is in the range of 1.5, so rather thin layers are required if the element is to be used at a visible wavelength. A 2π radian phase shift in transmission at 550 nm wavelength requires a thickness of only 367 nm. Photoresists of relatively high viscosity are required to achieve such a thin layer.

9.2.2 Multistep Lithography

The term *binary optics* has come to have different meanings to different people, but there are certain threads that are common and can serve to define the field. First and foremost is the fact that binary optical elements are most easily manufactured using VLSI fabrication techniques, namely photolithography and micromachining (the analog exposure values discussed above require more care than binary exposure in photoresist.) Second, the phase shift imparted by a binary optical elements depends solely on the surface relief profile of the optical element and the wavelength at which the element is to be used. The elements are usually thin structures, with relief patterns on the order of sub-micron to several microns in depth, and as such they can be inexpensively replicated using well-established methods of embossing. Surprisingly, the relief patterns utilized are often not binary at all, and therefore in a certain sense these elements are misnamed. However, such elements are usually defined through a series of binary exposure steps, and this fact has provided the rationale for retention of the name.

Approximation by a Stepped Thickness Function

Binary optical elements have stepped approximations to ideal continuous phase distributions. We briefly discuss the approximation process here, and then turn to the most common methods of fabrication.

We suppose that a certain thickness function $\Delta(x, y)$ is desired for the element (as usual, x and y are the transverse coordinates on the face of the element). Presumably this function has been derived from a design process, which may have been simple or may have been quite complex itself. As an example of a simple case, the element may be a grating of constant spatial frequency, the purpose of which is to deflect the incident light through a certain angle with the highest possible optical efficiency. An example of a more complex case might be a focusing element which generates an aspheric wavefront such that certain aberrations are reduced or eliminated. We shall assume that the desired thickness function $\Delta(x, y)$ is known and that the problem at hand is how to fabricate a thin relief element that closely approximates this desired thickness function.

An approximation to the desired thickness function is made by quantizing that function to a set of 2^N discrete levels (usually equally spaced). [Figure 9.12](#) shows an ideal phase grating profile with a perfect sawtooth period, and a quantized version of that grating with 2^N levels. The continuous blazed grating has the property that, if the peak-to-peak phase variation it introduces is exactly 2^N radians (or an exact multiple of 2π radians), 100% of the incident light will be diffracted into a single first diffraction order (cf. [Prob. 4-15](#)). The binary optic approximation to the grating shown in [Fig. 9.12](#) is a quantized version with 4 discrete levels. More generally 2π quantization levels can be realized through a series of N exposure and micromachining operations, as described below. The peak-to-peak thickness change of the

quantized element is $2^N - 1$ times the peak-to-peak thickness of the unquantized element.³

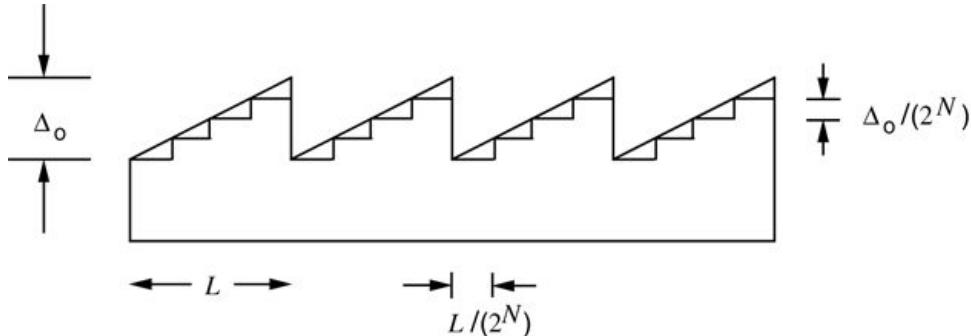


Figure 9.12

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 9.12 Ideal sawtooth thickness profile for a blazed grating, and binary optic approximation to that profile ($N=2$)
The illustration shows a figure that is a rectangle whose longer, top horizontal side is a series of four identical projections, each shaped like a right angled triangle whose base is L units long and whose height is Δ_0 units long. Thus we have, beginning at the top left corner, an

upward sloping line that drops a perpendicular that is delta subscript o units long, at the end of which an upward slope identical to the first begins and then similarly drops a perpendicular. This pattern is repeated to give the four projections. Below and along the sloping side is a step like formation of 4 horizontal lines, each measuring L/ (2 to the power N), and four vertical lines, each measuring delta o/(2 to the power N).

The diffraction efficiency of the step approximation to the sawtooth grating can be obtained by expanding its periodic amplitude transmittance in a Fourier series. A straightforward but tedious calculation shows that the diffraction efficiency of the q^{th} diffraction order can be expressed by [87]

$$\eta_q = \text{sinc}2q2Ns \text{sinc}2q-\phi_o 2\pi \text{sinc}2q-\phi_o 2\pi 2N,$$

$$\eta_q = \text{sinc}^2\left(\frac{q}{2^N}\right) \frac{\text{sinc}^2\left(q - \frac{\phi_o}{2\pi}\right)}{\text{sinc}^2\left(\frac{q - \frac{\phi_o}{2\pi}}{2^N}\right)},$$

(9-7)

where ϕ_o is the peak-to-peak phase difference of the continuous sawtooth grating, and is related to the peak-to-peak thickness variation (again, of the continuous grating) through

$$\phi_o = 2\pi\Delta_o(n_2 - n_1)\lambda_o,$$

$$\phi_o = 2\pi \frac{\Delta_o(n_2 - n_1)}{\lambda_o},$$

(9-8)

n_2 being the refractive index of the substrate and n_1 that of the surround, and λ_o being the vacuum wavelength of the light.

Of special interest is the case of a quantized approximation to the blazed grating with a peak-to-peak phase difference of $\phi_o = 2\pi$. Substitution in (9-7) yields

$$\eta_q = \text{sinc}2q2Ns \text{sinc}2q-1 \text{sinc}2q-12N.$$

$$\eta_q = \text{sinc}^2\left(\frac{q}{2^N}\right) \frac{\text{sinc}^2(q - 1)}{\text{sinc}^2\left(\frac{q - 1}{2^N}\right)}.$$

(9-9)

Consider for the moment only the last factor, consisting of the ratio of two sinc functions. The numerator is zero for all integer q except $q=1$, when it is unity. The denominator is also unity for $q=1$ and is nonzero except when

$$q-1=p2N,$$

$$q - 1 = p 2^N,$$

where p^P is any integer other than zero. For values of q^Q for which the numerator and denominator vanish simultaneously, l'Hôpital's rule can be used to show that the ratio of the two factors is unity. Thus the factor in question will be zero except when

$$q=p2^N+1,$$

$$q = p 2^N + 1,$$

in which case it is unity. The diffraction efficiency therefore is given by

$$\eta(p2^N+1) = \text{sinc}^2\left(p + \frac{1}{2^N}\right).$$

(9-10)

As the number, $2N 2^N$, of phase levels used increases, the angular separation between nonzero diffraction orders increases as well, since it is proportional to $2N 2^N$. The primary order of interest is the $+1 + 1$ order ($p=0$), for which the diffraction efficiency is

$$\eta_1 = \text{sinc}^2\left(\frac{1}{2^N}\right).$$

(9-11)

[Figure 9.13](#) shows the diffraction efficiencies of various nonzero orders as a function of the number of levels. It can be seen that, as $N \rightarrow \infty$, all diffraction orders except the $+1 + 1$ order vanish, and the diffraction efficiency of that nonvanishing order approaches 100%, identical with the case of a continuous blazed grating with the same peak-to-peak phase shift. Thus the properties of a stepped approximation to the continuous blazed grating do indeed approach those of the continuous grating as the number of steps increases.

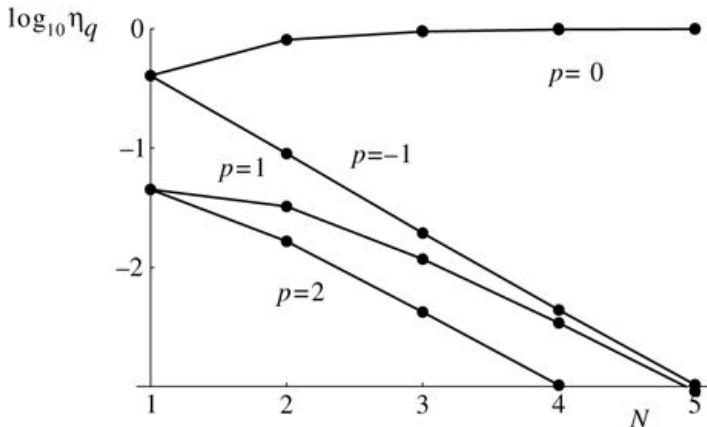


Figure 9.13

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 9.13 Diffraction efficiencies of various orders of a stepped approximation to a sawtooth grating. The parameter p determines the particular diffraction order, with the order number given by $p2^{N+1}$, and the number of discrete levels is 2^N .

The graph plots N values along the horizontal axis marked from 1 to 5 and values of $\log_{10} \eta_q$ along the vertical axis marked from minus 3 to 0, where minus 3 coincides with 1 on the horizontal axis. There are four curves, one for each of the p values minus 1, 0, 1, and 2. Curves for p values minus 1, 0, and 1 have five equidistant dots each, including one at each extreme. The curve for $p = 2$ has 4 such dots. The $p = -1$ curve begins near the 0.4 mark on the vertical axis and is a downward slope all the way till the +5 mark on the horizontal axis. The $p = 0$ curve also begins near the minus 0.4 mark on the vertical axis and slopes gently upward up to the next dot and thereafter runs almost parallel to the horizontal axis, that is, with a slight upward slope. The $p = 1$ curve begins near the minus 1.4 mark on the vertical axis and is a gentle downward slope till the next dot, thereafter the curve is steeper in an almost straight line path to the +5 mark on the horizontal axis. The $p = 2$ curve also begins near the minus 1.4 mark on the vertical axis but ends at the +4 mark on the horizontal axis, getting a little steeper along the way.

The Fabrication Process

Figure 9.14 illustrates the process by which a four-level binary optic approximation to a sawtooth thickness function is generated. The process consists of a number of discrete steps, each of which consists of photoresist application, exposure through one of several binary masks, photoresist removal, and etching of the substrate. Masks are often made by electron-beam writing. For a binary optic element with 2^N levels, N separate masks are required. Part (a) of the figure shows a substrate overcoated with photoresist, which is exposed through the first binary mask,

having transparent cells of width equal to $1/2^N$ of the period of the desired final structure. After exposure, the photoresist is developed. For a positive photoresist, the development process removes the exposed areas and leaves the unexposed areas, while for a negative photoresist the opposite is true. We will assume a positive photoresist here. Following the photoresist development process, micromachining is applied to remove material from the uncovered portions of the substrate, as illustrated in part (b) of the figure. The two most common

micromachining methods are reactive ion etching and ion milling. This first micromachining step removes substrate material to a depth of $1/2N^{1/2^N}$ th of the desired peak-to-peak depth of the grating. Now photoresist is spun onto the substrate a second time and is exposed through a second mask which has openings of width equal to $1/2N-1^{1/2^{N-1}}$ th of the desired final period, as shown in part (c) of the figure. Micromachining again removes the exposed portions of the substrate, this time with an etch depth $1/2N-1^{1/2^{N-1}}$ th of the final desired maximum depth, as illustrated in part (d) of the figure. For the case of a four-level element, the fabrication process now terminates. If there are $2N^{2^N}$ levels desired, $N^{N/2}$ different masks, exposures, development, and etching processes are required. The last etch process must be to a depth that is $1/2^{1/2}$ of the total desired peak-to-peak depth. A variety of different materials can be used for the substrate of such elements including silicon and glass. It is also possible to make reflective optical devices by overcoating the etched profile with a thin layer of metal. With the use of electron beam writing, it is possible to control the accuracy of the masks to about one-tenth of a μm . When the profile is more complex than purely binary, alignment of several masks is required, and the accuracy is reduced.

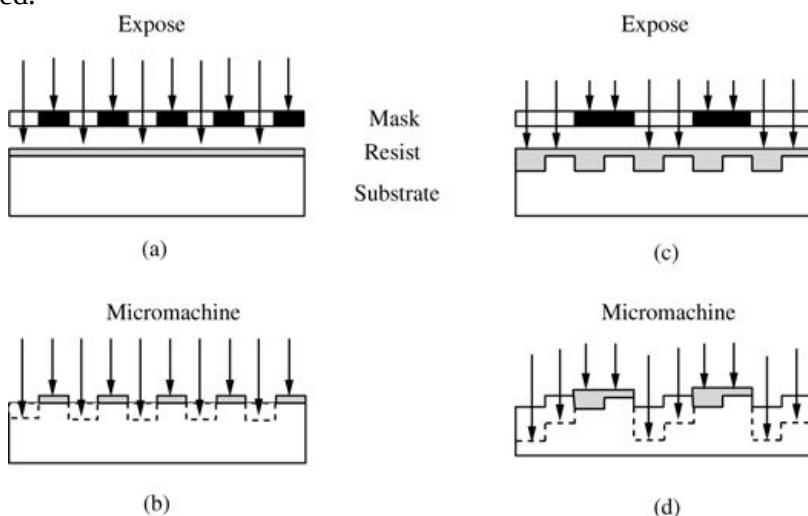


Figure 9.14

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 9.14 Steps in the fabrication of a four-level binary optic element.

Illustration a, labeled “Expose,” shows a rectangular substrate lined at the top with a thin layer of resist. Above it is the mask represented as a series of ten alternately filled and blank rectangles. Vertical downward pointing arrows reach up to the filled rectangles and do not penetrate it. Arrows directed at the blank rectangles are shown crossing the mask.

Illustration b, labeled “micro machine,” shows downward pointing parallel rays such that alternate arrows penetrate the substrate while the rest stop at the resist that is now discontinuous. The penetration level is marked by a dotted horizontal line at whose extremes perpendiculars are dropped from the resist sections that are in their original place.

Illustration c, labeled “Expose,” is the same as illustration a, but with two differences. The surface of the substrate is such that it has rectangular protrusions. The resist fills the gaps between these

projections. The mask is a series of only 5 but longer rectangles.

Illustration d, labeled “micro machine,” shows downward pointing parallel rays such that the leftmost arrow has penetrated the substrate the most, the second arrow a little less and the third and fourth arrows do not penetrate at all. The resist layer for the third arrow is much thicker than that for the fourth arrow. This pattern repeats to the right up to the tenth arrow. The extent of the penetration of the rays is marked by dotted horizontal lines that are connected to form a step like formation.

Diffraction efficiencies of 80% to 90% are quite common for these types of elements.

9.2.3 Other Types of Diffractive Optics

Attention has been focused above on two types of diffractive optics, which are fabricated by the techniques widely used in the semiconductor industry. Many other approaches to fabricating diffractive optical elements exist. Some methods use similar substrates to those mentioned above, but use different methods of micromachining, for example diamond turning or laser ablation. Some differ through their use of photographic film, rather than etchable substrates, as the means for creating the element. Computer-generated holographic optical elements are an example that will be discussed in more detail in [Chapter 11](#). Some depend on more conventional methods for recording holograms, for example, interference of two beams of light to expose photoresist.

For an overview of the field, including examples of many different approaches, the reader is referred to the proceedings of a series of meetings held on this general subject [\[64\]](#), [\[65\]](#), [\[66\]](#), [\[67\]](#), [\[68\]](#).

9.2.4 A Word of Caution

The capability of semiconductor fabrication techniques to make structures of ever smaller physical size has led already to the construction of diffractive optical elements with individual feature sizes that are comparable with and even smaller than the size of a wavelength of the light with which the element will be used. Such small structures lie in the domain where the use of a scalar theory to predict the properties of these optical elements is known to yield results with significant inaccuracies. It is therefore important to use some caution when approaching the analysis of the properties of diffractive optical elements. If the minimum scale size in the optical element is smaller than a few optical wavelengths, then a more rigorous approach to diffraction calculations will probably be needed, depending on the accuracy desired from the computation. For a discussion of such issues, see, for example, [\[285\]](#) and [\[240\]](#).

9.3 Liquid Crystal Spatial Light Modulators

Wavefront modulating elements made with photographic film or by lithographic means have one distinct disadvantage in many applications. Namely, they are fixed modulators of wavefronts rather than dynamic ones. However, if information is being rapidly gathered, perhaps by some electronic means, one would prefer a more direct interface between the electronic information and an optical data processing system. For this reason, a large number of devices capable of converting data in electronic form (or sometimes in incoherent optical form) into spatially modulated coherent optical signals have been explored. Such a device is called a *spatial light modulator*, a term that is abbreviated by SLM.

There is a broad categorization of SLMs into two classes that can be made at the start: (1) electrically written SLMs and (2) optically written SLMs. In the former case, electrical signals representing the information to be input to the system (perhaps in raster format) directly drives a device in such a way as to control its spatial distribution of absorption or phase shift. If the information is to be written in optical form, it may be input to the SLM in the form of an optical image at the start, rather than in electrical form. In this case the function of the SLM may be, for example, to convert an incoherent image into a coherent image for subsequent processing by a coherent optical system. Often a given SLM technology may have two different forms, one suitable for electrical addressing and one suitable for optical addressing.

Optically addressed SLMs have several key properties besides their fast temporal response that are very useful for optical processing systems. First, they can convert incoherent images into coherent images, as alluded to above. Second, they can provide image amplification: a weak incoherent image input to an optically addressed SLM can be read out with an intense coherent wave. Third, they can provide wavelength conversion: e.g. an incoherent image in the infrared could be used to control the amplitude transmittance of a device in the visible.

SLMs are used not only to input data to be processed, but also to create spatial filters that can be modified in real time. In such a case the SLM is placed in the back focal plane of a Fourier transforming lens, where it modifies the transmitted amplitude of the fields in accord with a desired complex spatial filter.

A great many different SLM technologies have been explored. Books have been written on this subject (see, for example, [\[102\]](#)). For a review article covering the properties of more types of SLMs than will be discussed here, the reader may wish to consult [\[263\]](#) and its associated references. In addition, a series of meeting proceedings on the subject provides valuable information [\[99\]](#), [\[100\]](#), [\[101\]](#). Here we limit ourselves to presenting the barest outlines of the principles of operation of what are currently regarded as the most important SLM technologies. These include (1) liquid crystal SLMs, (2) deformable mirror SLMs, and (3) acousto-optic cells.

9.3.1 Properties of Liquid Crystals

The use of liquid crystals in displays is commonplace. Examples include television displays and screens for laptop computers. In such applications voltages applied to pixelated electrodes cause a change in the intensity of the light transmitted by or reflected from the display. Similar principles can be used to construct a spatial light modulator for input to an optical system.

Background on the optics of liquid crystals can be found in [305], Section 6.5. For an additional reference that covers liquid crystal displays in detail, the reader can consult [189]. See also [102], [Chapters 1](#) and [2](#).

Mechanical Properties of Liquid Crystals

Liquid crystal materials are interesting from a physical point of view because they share some of the properties of both solids and liquids. The molecules composing such materials can be visualized as ellipsoids, with a single long axis about which there is circular symmetry in any transverse plane. These ellipsoidal molecules can stack next to one another in various ways, with different geometrical configurations defining different general types of liquid crystals. Adjacent molecules are not rigidly bound to one another, and can rotate or slide with respect to one another under the application of mechanical or electrical forces, thus exhibiting some of the properties of a liquid. However, there are constraints on the geometrical organization of collections of molecules, and these constraints introduce some properties normally associated with solids.

There are three different general classes (or *phases*) of liquid crystals that are of general interest in optics: (1) nematic, (2) smectic, and (3) cholesteric. The classes are differentiated by the different molecular orders or organizational constraints, as illustrated in [Fig. 9.15](#). For *nematic* liquid crystals (NLC), the molecules throughout the entire volume of the material favor a parallel orientation, with randomly located centers within that volume. For *smectic* liquid crystals, the molecules again favor parallel alignment, but their centers lie in parallel layers, with randomness of location only within a layer. Finally, a *cholesteric* liquid crystal is a distorted form of a smectic liquid crystal in which, from layer to layer, the alignment of molecules undergoes helical rotation about an axis. Spatial light modulators are based primarily on nematic liquid crystals and on a special class of smectic liquid crystals (the so-called smectic-C* phase) called *ferroelectric* liquid crystals (FLC), so our discussions will focus on these types primarily.

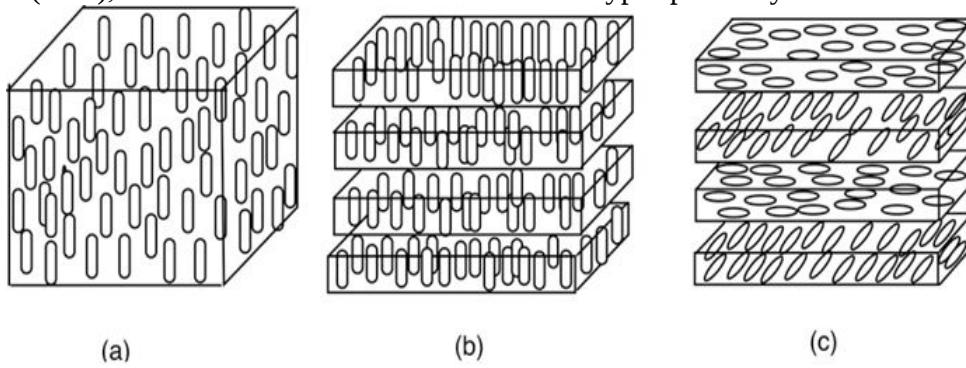


Figure 9.15
Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 9.15 Molecular arrangements for different types of liquid crystals. (a) Nematic liquid crystal, (b) smectic liquid crystal, and (c) cholesteric liquid crystal. The layers in (b) and (c) have been separated for clarity. Only a small column of molecules is shown.

Image a shows a cuboid block filled with identical oblong shapes floating in it vertically and each parallel to the rest. Image b shows the cuboid now divided into four horizontally arranged slabs of equal size with a slight gap between two adjacent slabs. Each slab is crowded with the oblong shapes in a similar vertically parallel arrangement. Image c is same as image b but here the oblong shapes in the top slab are horizontal while those in the next slab are upward sloping. In the next slab they are again horizontal, with the bottom slab showing upward sloping shapes.

It is possible to impose boundary conditions on the alignment of nematic liquid crystal molecules contained between two glass plates by polishing soft alignment layers coated on those plates with strokes in the desired alignment direction. The small scratches associated with the polishing operation establish a preferred direction of alignment for the molecules that are in contact with the plate, with their long direction parallel with the scratches. If the two alignment layers are polished in different directions (for example, in orthogonal directions), then the tendency of the molecules to remain aligned with one another (characteristic of the nematic liquid crystal phase) and the alignment of the molecules with the direction of polish at the glass plates combine to create a *twisted* nematic liquid crystal, as illustrated in [Fig. 9.16](#). Thus as we move between the two plates, the directions of the long axes of the various molecules remain parallel to one another in planes parallel to the glass plates, but gradually rotate between those planes to match the boundary conditions at the alignment layers.

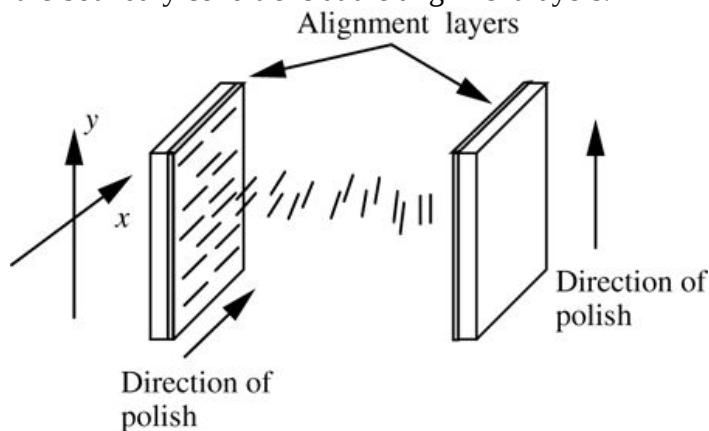


Figure 9.16

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 9.16 Molecular arrangements in a twisted nematic liquid crystal. The lines between the alignment layers indicate the direction of molecular alignment at various depths within the cell.

The illustration shows in the left extreme a coordinate plane with upward pointing vertical axis y and horizontal axis x suggesting depth running from near end to the far end. Two vertical glass plates with their alignment layers facing each other are next to the axes, their height parallel to the y axis and width parallel to the x axis. Horizontal line segments on the face of the left alignment layer, also parallel to the x axis, represent polishing; its direction is suggested by an arrow in the direction of the x axis. There are such horizontal line segments also between the alignment layers. Next to the right alignment layer, an upward arrow parallel to the y axis suggests direction of polish.

The structure of ferroelectric liquid crystals is more complex. Since they are of the smectic type, their molecules are arranged in layers. Within a given layer, the molecules are aligned in the same direction. For smectic-C* materials, the angle of the molecules within a single layer is

constrained to lie at a specific declination angle θ_t with respect to the layer normal, and thus there is a cone of possible orientations for any given layer. [Figure 9.17](#) illustrates the structure of the surface stabilized FLC for large cell thickness. The directions of orientation between layers form a helical spiral. The angular directions of the two layers at the interfaces with the glass plates can be stabilized by aligned polishing⁴ [69]. In practice, the cells are made sufficiently thin (typically only a very few microns of thickness) to eliminate the possibility that different layers will be in different allowed states.

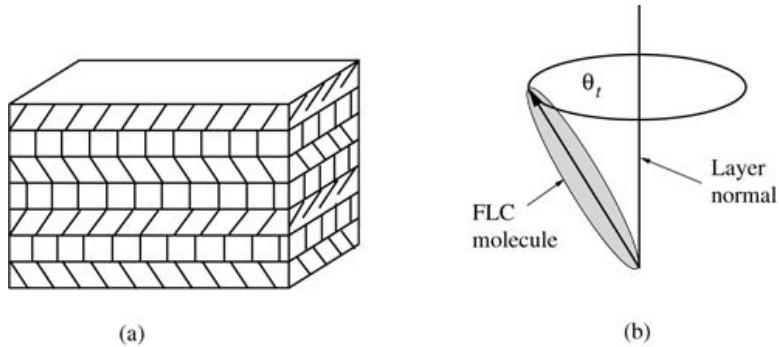


Figure 9.17

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 9.17 Ferroelectric liquid crystal (a) smectic-C* layered structure, and (b) allowed molecular orientations.

The illustration a is a cuboid block of seven layers. The orientation of the sections representing molecules forming each layer is different from that of those in the neighboring layer. In the top down order, the orientations, indicated by parallel strokes of lines, are as follows: upward sloping, perpendicular, downward sloping, perpendicular, upward sloping, perpendicular, downward sloping. Illustration b shows a vertical line labeled “Layer normal” passing through the center of a horizontally oriented circle. An upward pointing slanting arrow set along the length of an oval labeled “F L C molecule” connects the lower extreme of the layer normal and the circumference of the circle. The angle between the F L C molecule and the layer normal is theta subscript t.

Electrical Properties of Liquid Crystals

Both displays and SLMs exploit the ability to change the transmittance of a liquid crystal by means of applied electric fields. Usually those fields are applied between the glass plates that contain the liquid crystal material using transparent conductive layers (indium tin oxide films) coated on the inside of the glass plates. In order to achieve alignment of the liquid crystal at the interface, the conductive layer is covered with a thin alignment layer (often polyimide) which is subjected to polishing, as shown in Fig. 9.18.

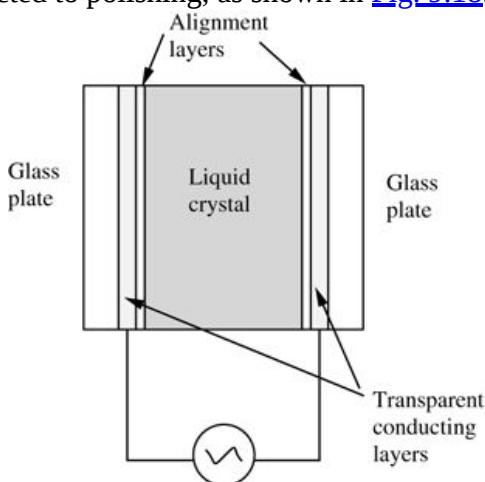


Figure 9.18

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 9.18 Structure of an electrically controlled liquid crystal cell.

The illustration shows a rectangle, representing liquid crystal, sandwiched between two thin alignment layers, which in turn are sandwiched between two slightly thicker transparent conducting layers, which in turn are held between two glass plates. Two wires from the sides of an AC source connect to the transparent conducting layers.

The application of an electric field across such a device can induce an electric dipole in each liquid crystal molecule, and can interact with any permanent electric dipoles that may be present. If, as is usually the case, the dielectric constant of a molecule is larger in the direction of the long axis of the molecule than normal to that axis, the induced dipoles have charge at opposite ends of the long direction of the molecule. Under the influence of the applied fields, the torques exerted on these dipoles can cause the liquid crystal molecules to change their natural spatial orientation.

For nematic liquid crystals, which do not have the extra constraints of smectic and cholesteric materials, a sufficiently large applied voltage will cause the molecules that are not in close proximity to the alignment layers to rotate freely and to align their long axes with the applied field. Thus the arrangement of the molecules within the twisted nematic liquid crystal cell shown previously in [Fig. 9.16](#) will change under sufficient applied field to the arrangement shown in [Fig. 9.19](#), in which the vast majority of the molecules have their long axis aligned with the field, i.e. pointing in a direction normal to the glass plates. As we shall discuss shortly, the change in the orientation of the molecules changes the optical properties of the cell as well. To avoid permanent chemical changes to the NLC material, cells of this type are driven by AC voltages, typically with frequencies in the range of 1 kHz to 10 kHz and with voltages of the order of 5 volts. Note that because the dipole moment of a nematic liquid crystal is an *induced* moment rather than a permanent moment, the direction of the moment reverses when the applied field reverses in polarity. Thus the direction of the torque exerted by the field on the molecules is independent of the polarity of the applied voltage, and they align in the same direction with respect to the applied field, regardless of polarity.

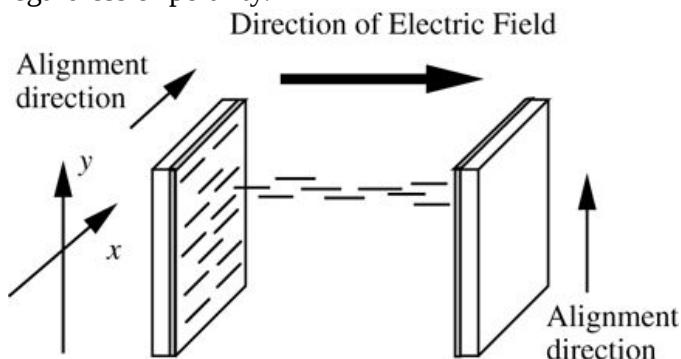


Figure 9.19

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 9.19 Twisted nematic liquid crystal with a voltage applied. Only a small column of molecules is shown.

The illustration shows in the left extreme a coordinate plane with upward pointing vertical axis *y* and horizontal axis *x* suggesting depth running from the near end to the far end. There are two vertical glass plates, their heights parallel to the *y* axis and width parallel to the *x* axis. A rightward pointing arrow from the left layer to the right layer indicates the direction of electric field. Horizontal line segments are shown moving in this direction. On the side of the left alignment layer that is facing the other alignment layer, there are horizontal line segments whose alignment direction matches the *x* axis. An arrow pointing from the near end to the far end indicates the

alignment direction in the left layer. An upward pointing vertical arrow suggests the alignment direction in the right layer.

In the case of the ferroelectric liquid crystal cell, the molecules can be shown to have a permanent electric dipole (with an orientation normal to the long dimension of the molecules), which enhances their interaction with the applied fields and leads to only two allowable orientation states, one for each possible direction of the applied field. [Figure 9.20](#) shows the molecules oriented at angle $+θ_t$ to the surface normal for one direction of the applied field and $-θ_t$ to the surface normal for the other direction of applied field. Because of the permanent dipole moment of the FLC molecules, the current state is retained by the material even after the applied field is removed. The FLC cell is thus *bistable* and has memory. It is because of the permanent dipole moment that the direction of the applied field matters. Unlike the case of nematic liquid crystals, DC fields of opposite polarity must be applied to the ferroelectric liquid crystal in order to switch between states.

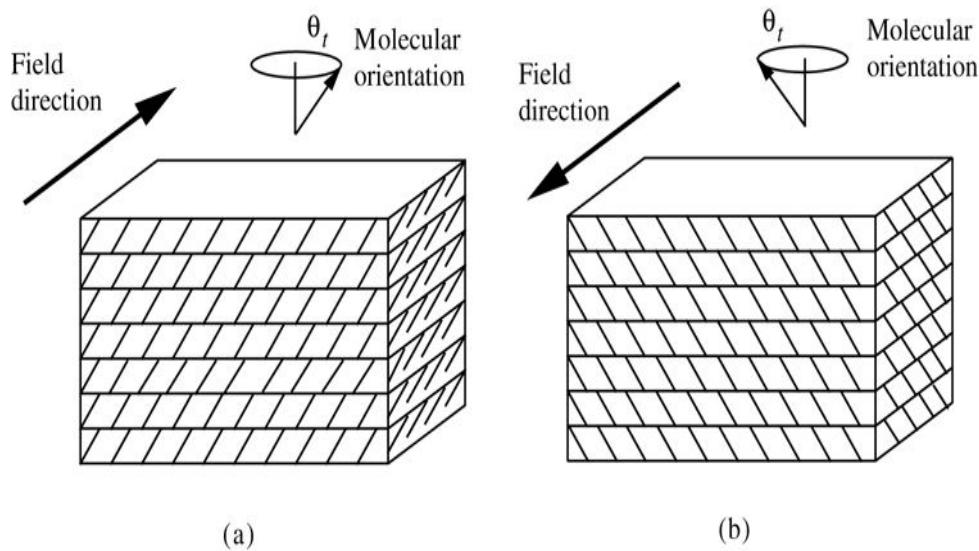


Figure 9.20

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 9.20 Ferroelectric liquid crystal molecules align in one of two allowed directions, depending on the direction of the field.

The angles of orientation in the two states are separated by $2\theta_t$.

Illustration a shows a cuboid that is a stack of seven identical slabs with all its constituent molecules being upward sloping. The cuboid's longer side is oriented horizontally. An arrow moving in the direction of its width, that is, from the near end to the far end, marks the field direction. The molecular orientation is shown in an accompanying diagram showing a perpendicular line passing through the center of a circle in a horizontal plane. A rightward upward slanting arrow connects the lower end of the perpendicular to the circumference of the circle and making an angle with the perpendicular measuring theta subscript t.

Illustration b is the same as illustration a, but with some differences. Here the constituent molecules are downward sloping. The field direction is the opposite, that is, from the far end to the near end. And the molecular orientation shown in the accompanying diagram is leftward upward slanting.

Liquid crystals have high resistivity and therefore act primarily as an electrical dielectric material. The electrical response of a liquid crystal cell is predominantly that of a simple RC circuit, where the resistance arises from the finite resistivity of the transparent electrodes and the capacitance is that of a parallel plate capacitor (the NLC cell is typically 5 to 10 μm thick). For sufficiently small cells, or sufficiently small pixels on a large array, the electrical time constant is small by comparison with the time constant associated with the mechanical rotation of the molecules. Typical time constants for NLC materials are approximately 100 μs for the molecules to align with an applied field, and 20 ms for the molecules to relax back to their original state. The permanent dipole moment of the FLC materials makes them considerably faster; cell thicknesses are typically in the 1- to 2- μm range, applied voltages are typically in the 5- to 10-volt range, and switching times of the order of 50 μs . In some cases even submicrosecond response times are observed [70].

Optical Properties of Nematic and Ferroelectric Liquid Crystals

A quantitative understanding of the behavior of SLMs based on liquid crystals, as well as many other types of SLMs that operate by means of polarization effects, requires the use of a mathematical formalism known as the *Jones calculus*. This formalism is outlined in [Appendix C](#), to which the reader is referred. The state of polarization of a monochromatic wave with X U_X and Y U_Y components of polarization expressed in terms of complex phasors U_X and U_Y is represented by a *polarization vector* \vec{U} with components U_X and U_Y ,

$$U \rightarrow = U_X U_Y.$$

$$\vec{U} = \begin{bmatrix} U_X \\ U_Y \end{bmatrix}.$$

(9-12)

The passage of light through a linear polarization-sensitive device is described by a 2×2 2×2 *Jones matrix*, such that the new polarization vector $U \rightarrow'$ is related to the old polarization vector $U \rightarrow$ through the matrix equation

$$U \rightarrow' = L U \rightarrow = l_{11} l_{12} l_{21} l_{22} U \rightarrow.$$

$$\vec{U}' = L \vec{U} = \begin{bmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \end{bmatrix} \vec{U}.$$

(9-13)

If we can characterize a given device by specifying its Jones matrix, we will then be able to understand completely the effect of that device on the state of polarization of an incident wave.

The elongated structure of liquid crystal molecules causes such materials to be *anisotropic* in their optical behavior, and in particular to exhibit *birefringence*, or to have different refractive indices for light polarized in different directions. The most common materials have larger refractive indices for optical polarization parallel to the long axis of the crystal (the *extraordinary*

refractive index, n_e), and smaller uniform refractive index for all polarization directions normal to the long axis (the *ordinary* refractive index, n_o). One of the highly useful properties of these materials is the very large difference between the extraordinary and ordinary refractive indices they exhibit, often in the range of 0.2 or more.

It can be shown (see [305], pp. 232–234) that, for a twisted nematic liquid crystal with no voltage applied, having a helical twist of α radians per meter in the right-hand sense along the direction of wave propagation and introducing a relative retardation β radians per meter between the extraordinary and ordinary polarization components, a wave polarized initially in the direction of the long molecular axis at the entrance surface of the cell will undergo polarization rotation as the light propagates through the cell, with the direction of polarization closely tracking the direction of the long crystal axis, provided only that $\beta \gg \alpha$. The Jones matrix describing such a transformation can be shown to be the product of a coordinate rotation matrix $L_{\text{rotate}}(-\alpha d)$ and a wave retarder $L_{\text{retard}}(\beta d)$,

$$L = L_{\text{rotate}}(-\alpha d)L_{\text{retard}}(\beta d),$$

$$\mathbf{L} = \mathbf{L}_{\text{rotate}}(-\alpha d) \mathbf{L}_{\text{retard}}(\beta d),$$

(9-14)

where the coordinate rotation matrix is given by

$$L_{\text{rotate}}(-\alpha d) = \cos \alpha d - \sin \alpha d \sin \alpha d \cos \alpha d,$$

$$\mathbf{L}_{\text{rotate}}(-\alpha d) = \begin{bmatrix} \cos \alpha d & -\sin \alpha d \\ \sin \alpha d & \cos \alpha d \end{bmatrix},$$

(9-15)

and the retardation matrix is (neglecting constant phase multipliers)

$$L_{\text{retard}}(\beta d) = 100e^{-j\beta d},$$

$$\mathbf{L}_{\text{retard}}(\beta d) = \begin{bmatrix} 1 & 0 \\ 0 & e^{-j\beta d} \end{bmatrix},$$

(9-16)

where β is given by

$$\beta = 2\pi(n_e - n_o)\lambda_0.$$

$$\beta = \frac{2\pi(n_e - n_o)}{\lambda_0}.$$

(9-17)

Here λ_0 is the vacuum wavelength of light and d is the cell thickness. With the help of this Jones matrix, the effects of the twisted nematic cell with no voltage applied can be found for any

initial state of polarization.

When voltage is applied to an NLC cell along the direction of wave propagation, the molecules rotate so that the long axis coincides with that direction, and no polarization rotation occurs. Under this condition both α and β go to zero, and the cell has no effect on the incident polarization state. Thus an NLC can be used as a *changeable* polarization rotator, with rotation experienced in the unexcited state (no voltage applied) by an amount determined by the orientation of the alignment layers on the two glass plates as well as the thickness of the cell, and no rotation experienced in the excited state (voltage applied).

To consider the case of an FLC cell, a bit of further background is needed. When a liquid crystal cell of thickness d has all of its molecules tilted such that the long dimension of the molecule lies in the (x, y) plane, but tilted at angle $+θ_t$ to the y (vertical) axis, the effects of the cell on incident light can be represented by a Jones matrix that is the sequence of a coordinate rotation with angle $θ_t$ which aligns the direction of the y axis with the long axis of the molecules, a retardation matrix representing the phase shift experienced by polarization components oriented parallel to the long and short axes of the liquid crystal molecule, followed by a second rotation matrix with angle $-θ_t$ which returns the y axis to its original orientation at angle $-θ_t$ to the long axis of the molecule. Taking account of the proper ordering of the matrix product,

$$L = L_{\text{rotate}}(-θ_t) L_{\text{retard}}(βd) L_{\text{rotate}}(θ_t) = \cos θ_t - \sin θ_t \sin θ_t \cos θ_t 100e^{-jβd} \cos θ_t \sin θ_t - \sin θ_t \cos θ_t,$$

$$\begin{aligned} L &= L_{\text{rotate}}(-θ_t) L_{\text{retard}}(βd) L_{\text{rotate}}(θ_t) \\ &= \begin{bmatrix} \cos θ_t & -\sin θ_t \\ \sin θ_t & \cos θ_t \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & e^{-jβd} \end{bmatrix} \begin{bmatrix} \cos θ_t & \sin θ_t \\ -\sin θ_t & \cos θ_t \end{bmatrix}, \end{aligned}$$

(9-18)

where $β$ is again given by (9-17).

The Jones matrix for such an FLC cell has two possible forms, one for each direction of the applied field. When the applied field switches the direction of alignment so that the long molecular axis is at angle $+θ_t$ to the y axis, then from (9-18) the Jones matrix is of the form

$$L_+ = L_{\text{rotate}}(-θ_t) L_{\text{retard}}(βd) L_{\text{rotate}}(θ_t),$$

$$L_+ = L_{\text{rotate}}(-θ_t) L_{\text{retard}}(βd) L_{\text{rotate}}(θ_t),$$

(9-19)

whereas for the field in the opposite direction we have

$$L_- = L_{\text{rotate}}(θ_t) L_{\text{retard}}(βd) L_{\text{rotate}}(-θ_t).$$

$$L_- = L_{\text{rotate}}(θ_t) L_{\text{retard}}(βd) L_{\text{rotate}}(-θ_t).$$

(9-20)

A case of special interest is that of a cell thickness d such that the retardation satisfies $\beta d = \pi$ (i.e. the cell is a half-wave plate). The reader is asked to verify (see [Prob. 9-2](#)) that the two Jones matrices above can be reduced to the forms

$$L_+ = \cos 2\theta_t \sin 2\theta_t \sin 2\theta_t - \cos 2\theta_t L_- = \cos 2\theta_t - \sin 2\theta_t - \sin 2\theta_t - \cos 2\theta_t.$$

$$\begin{aligned} L_+ &= \begin{bmatrix} \cos 2\theta_t & \sin 2\theta_t \\ \sin 2\theta_t & -\cos 2\theta_t \end{bmatrix} \\ L_- &= \begin{bmatrix} \cos 2\theta_t & -\sin 2\theta_t \\ -\sin 2\theta_t & -\cos 2\theta_t \end{bmatrix}. \end{aligned}$$

(9-21)

Furthermore, when the input to the FLC cell is a linearly polarized wave with polarization vector inclined at angle $+t\theta_t$ to the y axis, the output polarization vectors in the two respective cases are found to be

$$U' \rightarrow + = \sin \theta_t - \cos \theta_t U' \rightarrow - = -\sin 3\theta_t - \cos 3\theta_t.$$

$$\begin{aligned} \vec{U}'_+ &= \begin{bmatrix} \sin \theta_t \\ -\cos \theta_t \end{bmatrix} \\ \vec{U}'_- &= \begin{bmatrix} -\sin 3\theta_t \\ -\cos 3\theta_t \end{bmatrix}. \end{aligned}$$

(9-22)

Finally we note that if the tilt angle of the liquid crystal is 22.5° , the two vectors above are orthogonal, aside from a sign change indicating a 180° phase shift. Thus for this particular tilt angle, a wave with linear polarization coincident with the long molecular axis in one state of the device is rotated by 90° when the device is switched to the opposite state. Such a device is therefore a 90° rotator for this particular direction of input polarization.

Liquid crystal cells are often used to construct intensity modulators, and indeed such modulation is important for several different types of SLMs. Consider first the case of nematic liquid crystals. If the NLC cell has a polarizer on its front surface and a polarization analyzer on its rear surface, it can modulate the intensity of the light it transmits. For example, in the case of a 90° twist illustrated previously in [Fig. 9.16](#), with a polarizer oriented parallel to the front-surface alignment and an analyzer oriented parallel to the rear-surface alignment, light will pass through the exit analyzer when no voltage is applied to the cell (a consequence of rotation), but will be blocked due to the absence of rotation when the full extinction voltage is applied to the cell. If less than the full extinction voltage is applied, then over a certain range of voltage, partial intensity transmission will occur, with a limited dynamic range of analog operation. Similarly, an FLC can act as a 90° polarization rotator (as explained above) and therefore can act as a binary intensity modulator.

It is also possible to make a *reflection* modulator using a liquid crystal cell, as illustrated in [Fig. 9.21](#). For NLC materials, an untwisted cell is simplest. Consider a cell with the long

molecular axis (the “slow” axis) aligned parallel to the y^y axis throughout the cell. Let the thickness of the cell be chosen to ensure a $90^\circ 90^\circ$ relative retardation of polarization components oriented along and orthogonal to the slow axis after one pass through the cell (i.e. the cell is a quarter-wave plate). The output glass plate on the cell is replaced by a mirror, and a polarizer oriented at $45^\circ 45^\circ$ to the x^x axis is inserted at the front of the cell.

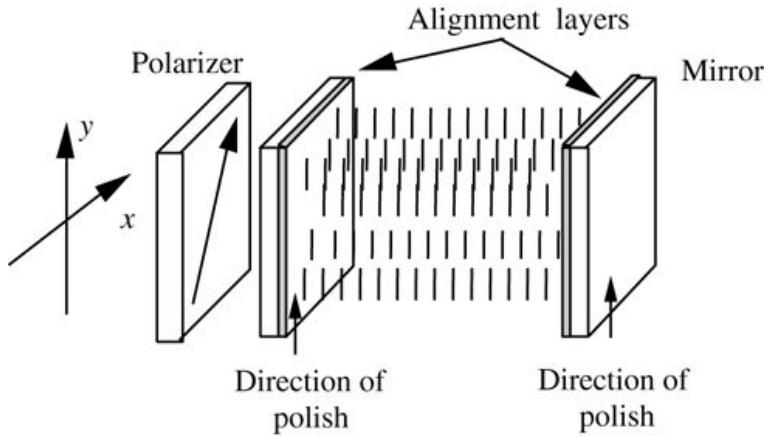


Figure 9.21

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 9.21 Intensity modulation with a reflective NLC cell.

The illustration shows in the left extreme a coordinate plane with upward pointing vertical axis y and horizontal axis x suggesting depth running from near end to the far end. To its right is a vertically placed slab representing a polarizer. An upward pointing arrow runs diagonally from the lower near corner to the top far corner. To its right are two alignment layers oriented parallel to the polarizer, all three in a row. There are rows of short vertical line segments shown between the two layers. An upward arrow on each of the layers indicates the direction of polish. The layer on the right is labeled “Mirror.”

The operation of this cell can be understood intuitively as follows. The light incident on the cell is linearly polarized at $+45^\circ +45^\circ$ to the x^x axis due to the presence of the polarizer. When no voltage is applied across the cell, there is no molecular rotation. After the first passage through the cell, the incident linear polarization has been converted to circular polarization. Reflection from the mirror reverses the sense of circular polarization, and a second pass back through the quarter-wave plate results in a linear polarization that is orthogonal to the original polarization. Thus the reflected light is blocked by the polarizer.

On the other hand, in the presence of a sufficiently large applied voltage, the long axes of the molecules all rotate to alignment with the direction of the applied field, which coincides with the direction of propagation of the wave, eliminating the birefringence of the cell. The direction of linear polarization is therefore maintained after passage through the cell, is unchanged after reflection from the mirror, and is unchanged after the second passage through the cell. The reflected light is therefore transmitted by the polarizer.

Application of a voltage that is less than that required to fully rotate the molecules will result in partial transmission of the reflected light.

In a similar fashion, it is possible to show that an FLC cell with tilt angle $22.5^\circ 22.5^\circ$ will act as a binary reflection intensity modulator if the input polarizer is aligned along one of the long

molecular orientation axes and the cell thickness is chosen to realize a quarter-wave plate.

This completes the background on liquid crystal cells, and we now turn attention to specific spatial light modulators based on these materials.

9.3.2 Spatial Light Modulators Based on Liquid Crystals

Of the SLM technologies that have been explored over a period of many years, the liquid crystal devices have survived the longest and remain the most important devices in practice. There are many variants of these devices, some using nematic liquid crystals and others using ferroelectric liquid crystals. We present a brief overview of the most important types.

Electrically Driven Liquid Crystal Spatial Light Modulators

The use of liquid crystals in television displays is widespread, and the technology of such displays has advanced rapidly in recent years. While displays of this type are not made for use in coherent light, nonetheless they can be adapted for use in a coherent optical system [148].

TV displays of this type are, of course, electrically addressed, and for high-definition 1080p, display 1920×1080 pixels. Liquid crystal microdisplays are available for use, for example, in heads-up displays. They have more modest resolution, typically VGA resolution of 640×480 , although they are available in resolutions up to QXGA, 2048×1536 at higher cost. These displays are made from nematic liquid crystals, usually with either a 90° or a 270° twist. For wavefront modulation devices, the displays with a small footprint are preferred. To use them in a coherent optical system, it is first necessary to remove polarizers attached to the display, and to remove any attached diffusing screen. Displays of this kind have not been manufactured with attention to their optical flatness, since the TV display application does not require it. Their most important attribute is their low cost, as compared with many other SLM technologies. The form of such an SLM is as shown in Fig. 9.18.

An electrically driven liquid crystal SLM can be constructed to be either a pixelated amplitude modulator or a pixelated phase modulator. For phase modulation, the liquid crystal layer does not require a twist, so the alignment layers on each side of the liquid crystal layer can be parallel. In the absence of an applied voltage, light polarized in the direction of the alignment layers encounters the extraordinary index of the liquid crystal molecules, which are co-aligned with the direction of polarization. In the presence of an applied field, the molecules rotate, with their long axes now parallel to the direction of propagation, thus presenting the ordinary index of refraction to the light propagating through the cell. The cell thickness and the difference between the two indices of refraction determine the amount of phase modulation that occurs when the states are switched.

For amplitude modulation, the pixelated cell should have a twist. Suppose the twist is 90° , and the light incident on the cell is polarized parallel to the long axis of the liquid crystal molecules at the input to the device. In the absence of an applied field, the liquid crystal molecules rotate by 90° from one side of the cell to the other, and the linear polarization of the light propagating through the cell likewise rotates by the same amount. A polarization analyzer at the output of the device, oriented orthogonally to the direction of the incident polarization, will pass the light that has propagated through the cell, and the pixel is in a “on” state. If on the other hand sufficient voltage is applied across the liquid crystal layer in the pixel, the molecules rotate to align themselves with the applied field, and no polarization rotation occurs. The output polarization analyzer then blocks the light propagating through the cell, and the pixel is in an “off” state.

Other ways of accomplishing the same behaviors are possible. In particular, as alluded to earlier, reflection devices can be constructed to accomplish either of the same behaviors.

Optically Driven Liquid Crystal Spatial Light Modulators

As an example of an optically driven liquid crystal SLM, we choose the SLM developed by Hughes Research Laboratories in 1970s. Unlike the devices discussed in the previous sections, which had their states changed by direct application of an electric field, this device is written optically, rather than electrically. However, optical writing results in the establishment of certain internal electric fields, and therefore the functioning of this device can be understood based on the previous background. A complete description of this rather complex device can be found in [149]. Our description will be somewhat simplified.

A diagram of the structure of the device is shown in Fig. 9.22. The device can be written with incoherent or coherent light of any state of polarization, and it is read with polarized coherent light. A polarizer and analyzer are external to the device, as will be discussed. To understand the operation of the device, we begin with the “write” side shown on the right of Fig. 9.22.

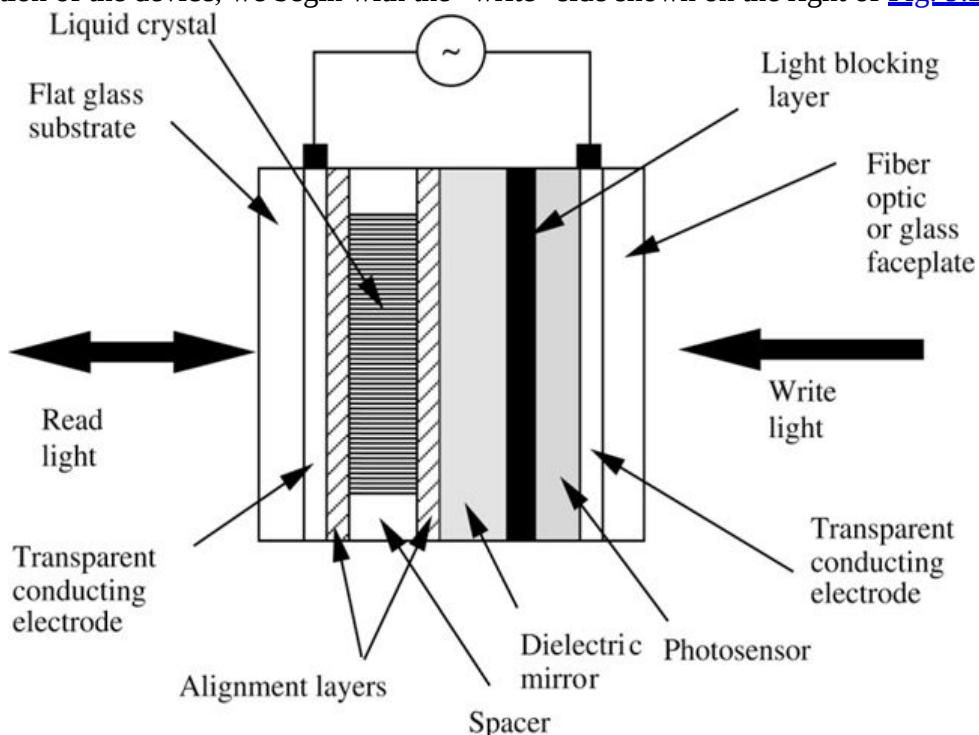


Figure 9.22

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 9.22 Hughes liquid crystal SLM.

The illustration shows a rectangular representation of the device divided into narrow rectangles representing its constituents, all extending from the top to the bottom. In the left half, a layer of liquid crystal flanked at its top and bottom ends by spacers, is sandwiched between two alignment layers. To its right, is a dielectric mirror followed by a light blocking layer followed by a photo sensor. All these layers are together sandwiched between two transparent conducting electrodes connected to an AC source. On the right extreme we have a layer of fiber optic or glass faceplate. And on the left extreme is a flat glass substrate. On the right side, a leftward pointing horizontal

arrow is labeled “Write light.” On the left side, a double directional horizontal arrow is labeled “Read light.”

Let an optical image be cast on the right-hand entrance of the device, which can consist of a glass plate or, for better preservation of resolution, a fiber-optic faceplate. The light passes through a transparent conducting electrode and is detected by a photoconductor, which in the most common version of the device is cadmium sulfide (CdS). The photoconductor should have the highest possible resistivity in the absence of write light, and the lowest possible resistivity in the presence of strong write light. Thus light absorbed by the photoconductor increases its local electrical conductivity in proportion to the incident optical intensity. To the left of the photoconductor is a light-blocking layer composed of cadmium telluride (CdTe), which optically isolates the write side of the device from the read side. An audio frequency AC voltage, with an rms voltage in the 5-to 10-volt range, is applied across the electrodes of the device.

On the read side of the device, an optically flat glass faceplate is followed to the right by a transparent conducting electrode, to the right of which is a thin NLC cell with alignment layers on both sides. The alignment layers are oriented at 45° to one another, so that with no applied voltage the liquid crystal molecules undergo a 45° twist. Following the liquid crystal is a dielectric mirror which reflects incident read light back through the device a second time. The dielectric mirror also prevents DC currents from flowing through the device, which extends its lifetime.

From the electrical point of view, it is the rms AC voltage applied across the liquid crystal layer that determines the optical state of the read side of the device. A simplified electrical model [14] for the device is shown in Fig. 9.23. In the off state (no write light applied), the two resistances are sufficiently large that they can be neglected, and the values of the capacitances of the photosensor and the dielectric stack must be sufficiently small (i.e. their impedances at the drive frequency must be sufficiently high) compared with the capacitance of the liquid crystal layer that the rms voltage across the liquid crystal layer is too small to cause the molecules to depart from their original twisted state. In the on state, ideally there is no voltage drop across the photosensor, and the fraction of the applied rms voltage appearing across the liquid crystal must be large enough to cause significant rotation of the molecules. The capacitances involved can be controlled in the design of the device, through appropriate choice of layer thicknesses, to satisfy these requirements.⁵

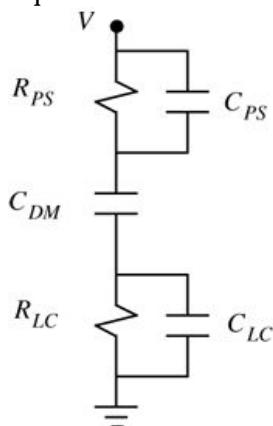


Figure 9.23

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 9.23 Electrical model for the optically written SLM. RPS R_{PS} and CPS C_{PS} are the resistance and capacitance of the photosensor, CDM C_{DM} is the capacitance of the dielectric mirror, and RLC R_{LC} and CLC C_{LC} are the resistance and capacitance of the liquid crystal layer.

The diagram shows a vertically extending circuit that begins at the top with a dot labeled V. The parts in the top down direction are as follows. resistance R subscript P S, capacitance C subscript D M, resistance R subscript L C, and ground. Connected parallel to the two the P S and L C resistances are the capacitances C subscript P S and C subscript L C. Each resistance is symbolized by a zigzag of four lines and each capacitance is symbolized by a pair of equal parallel lines. The ground is symbolized by a set of three parallel lines of unequal length with their centers aligned; the circuit line joins the center of the longest line.

Figure 9.24 illustrates the write and read operations. The liquid crystal layer is operated in a so-called “hybrid-field-effect” mode, which is explained as follows. The polarization of the incident read light is chosen to be in a direction parallel to the long axis of the aligned liquid-crystal molecules at the left-hand alignment layer. Thus as light passes through the liquid crystal layer, the direction of polarization follows the twisted direction of the liquid crystal molecules, arriving at the dielectric mirror with a 45° 45° polarization rotation. After reflection, the light propagates back through the liquid crystal a second time, with the direction of polarization again following the alignment of the molecules, thus returning to its original state. A polarization analyzer oriented at 90° 90° to the direction of incident polarization then blocks the reflected light, yielding a uniformly dark output image when there is no write light. If write light is applied to the device, a spatially varying AC electric field is established across the liquid crystal layer, and the long axis of the liquid crystal molecules begins to tilt away from the plane of the electrode. If the electric field were strong enough to fully rotate the molecules, then the birefringence of the material would vanish, the device would not change the direction of polarization, and again the reflected light would be completely blocked by the output analyzer. However, the fields are not sufficient to fully rotate the molecules, and hence they only partially tip away from the transverse plane, with an amount of tip that is proportional to the strength of the field (and therefore the strength of the write image). The partially tipped molecules retain some birefringent effect, and therefore the linearly polarized input light is transformed into elliptically polarized light, with a degree of ellipticity that depends on the strength of the applied field. The elliptically polarized field has a component that is parallel to the direction of the output analyzer, and therefore some of the reflected light passes that analyzer.

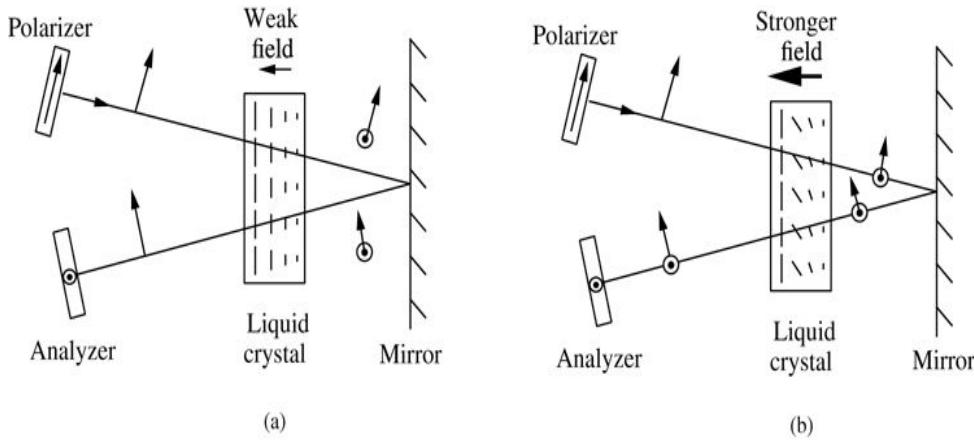


Figure 9.24

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 9.24 Readout of the Hughes liquid crystal SLM with (a) no write light present, and (b) write light present.

Illustration a shows a liquid crystal represented by a rectangle whose longer sides are oriented vertically. Inside the rectangle are columns of vertical dashes, longest in the leftmost column and shortest in the rightmost column. A leftward pointing horizontal arrow at the top of the rectangle is labeled “Weak field.” In the top left corner is a polarizer represented by a small upward sloping rectangle with an upward rightward pointing arrow inside it, along its length. In the bottom left corner is an analyzer represented by another similar but downward sloping rectangle with a circle at its center. A distinct dot marks the center of the circle. A downward sloping line from the polarizer passes through the liquid crystal and meets behind it at the center of a mirror located parallel to the crystal. The ray is reflected back through the liquid crystal to the analyzer in a sloping line. To the left of the liquid crystal, on the line from the polarizer, is an upward rightward pointing arrow, its inclination matching the arrow inside the polarizer. A similarly oriented arrow is located between the crystal and the mirror and above the line, and originating from the center of tiny symbolic circle.

Similarly, to the left of the liquid crystal, on the line from the analyzer, is an upward leftward pointing arrow. A similarly oriented arrow is located between the crystal and the mirror and below the line, and originating from the center of tiny symbolic circle. Illustration b is same as illustration a, but with a few differences. The arrow atop the liquid crystal is labeled “Stronger field.” The columns of dashes inside the crystal, except the leftmost, are downward sloping. The two arrows between the crystal and the mirror are located on the lines.

Contrast ratios of the order of 100:1 can be achieved with this device, and its resolution is several tens of line pairs per mm. The write time is of the order of 10 msec and the erase time about 15 msec. Due to the optically flat faceplate on the read side, the wavefront exiting the device is of good optical quality and the device is therefore suitable for use within a coherent optical system. The nonmonotonic dependence of reflectance on applied voltage (both no voltage and a very high voltage result in the analyzer blocking all or most of the light) allows the device to be operated in several different linear and nonlinear modes, depending on that voltage.

Ferroelectric Liquid Crystal Spatial Light Modulators

Ferroelectric liquid crystals provide the basis for several different approaches to the construction of spatial light modulators. SLMs based on these materials are inherently binary in nature, but gray scales can be created with the help of half-tone techniques. Both optically addressed and

electrically addressed FLC SLMs have been reported. An excellent reference can be found in [\[102\]](#), [Chapter 6](#).

Unlike NLC based devices, the FLC device must operate by reversal of the direction of the electric field across the liquid crystal layer. A different photoconductor, hydrogenated amorphous silicon, which has a faster response time than CdS, has been used. These devices are driven with audio-frequency square waves. The layer thicknesses (and therefore the capacitances in [Fig. 9.23](#)) are chosen so that the voltages appearing across the liquid crystal layer always remain sufficiently negative or sufficiently positive (depending on whether write light is or is not present) to drive the FLC material into its appropriate state. The tilt angles of the FLC molecules are again 45° apart and the FLC layer thickness is chosen for quarter-wave retardation, appropriate for a reflective modulator operating by polarization rotation.

Unlike the optically addressed SLMs, electrically addressed FLC SLMs are discrete pixelated devices, i.e. they display sampled images rather than continuous images. The FLC SLM is a pixelated version of the FLC intensity modulator described in a previous section.

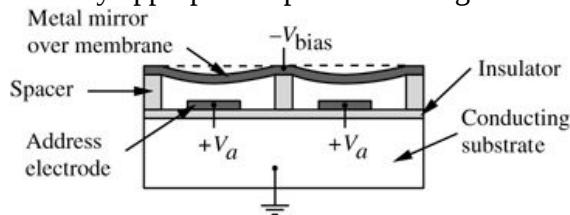
Liquid Crystal on Silicon (LCOS)

The term “liquid crystal on silicon” is used for a display technology having the virtue that it combines CMOS semiconductor technology with liquid crystal technology. This technology is pixelated, has high contrast ratio, and is typically small in size (typical LCOS panels have a dimension of the order of 3/4 inch on a side). The fundamental idea behind LCOS is to place a liquid crystal layer on top of a small metallic mirror in each pixel, with an additional transparent electrode on top of the liquid crystal. The voltage between the metallic mirror and the transparent electrode is driven by the CMOS circuitry under the mirror for each pixel. The operation of the device is similar to that of the reflective liquid crystal amplitude modulator described previously, except that polarized incoherent light (often from an LED) is used for the readout in display applications. This technology has found application particularly in so-called “pico-projector” applications, as used in head-mounted displays and other near eye applications. It can also be used as a pixelated amplitude modulator for coherent light. For references on LCOS technology, see [\[183\]](#) and [\[245\]](#).

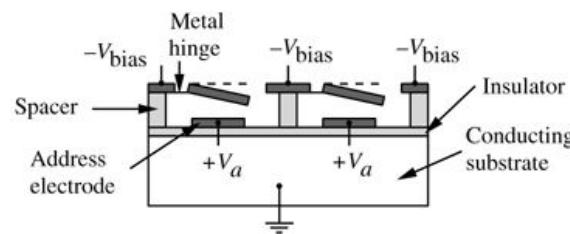
9.4 Deformable Mirror Spatial Light Modulators

A variety of devices have been reported that use electrostatically induced mechanical deformation to modulate a reflected light wave. Such devices are usually referred to as “deformable mirror devices,” or DMDs. The most advanced SLMs of this type have been developed by Texas Instruments, Inc. Early devices utilized continuous membranes which deformed under the fields exerted by pixelated driving electrodes. These SLMs gradually evolved into deformable mirror devices in which discrete cantilevered mirrors were individually addressed via voltages set on floating MOS (metal oxide semiconductor) sources, the entire device being integrated on silicon. The most recent versions have used mirror elements with two points of support, which twist under the application of an applied field. An excellent discussion of all of these approaches is found in [174].

Figure 9.25 shows the structures for a membrane device and for a cantilever beam device. For the membrane device, a metalized polymer membrane is stretched over a spacer grid, the spacers forming an air gap between the membrane and the underlying address electrodes. A negative bias voltage is applied to the metalized membrane. When a positive voltage is applied to the address electrode under the membrane, it deflects downward under the influence of the electrostatic forces. When the address voltage is removed, the membrane moves upward to its original position. In this way a phase modulation is introduced, but that phase modulation can be converted to an intensity modulation by appropriate optics following the mirror.



(a)



(b)

Figure 9.25

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 9.25 Deformable mirror pixel structures for (a) a membrane SLM and (b) a cantilever beam SLM.

Illustration a shows a conducting substrate represented by a rectangle with its longer side oriented horizontally. It is grounded at the center. Above the substrate is a thin layer of insulator, on top of

which are three posts labeled “Spacers,” one at each of the two extremes, left and right, and one in the middle. On either side of the middle spacer post is an address electrode connected to the substrate through the insulator; the connection is labeled “ $+ V$ subscript a.” A metal mirror over membrane runs across the tops of the spacer posts, sagging marginally between the spacer posts. At the middle of the membrane is an upward connection labeled “negative V subscript bias.” A dotted horizontal line connects the tops of the spacer posts. Illustration b is the same as illustration a, but with differences. All spacer posts are connected to negative V subscript bias voltage. A downward sloping metalized beam is attached to the left and middle spacer posts by a thin metal hinge such that the beam lies over the electrodes.

For the cantilever beam device, the structure is quite different. The metalized beam, which is biased to a negative voltage, is attached to a spacer post through a thin metal hinge. When the underlying address electrode is activated with a positive voltage, the cantilever rotates downward, although not far enough to touch the address electrode. An incident optical beam is thus deflected by the tilted pixel, and will not be collected by an optical system that follows. In this way an intensity modulation is induced at each pixel.

The most advanced DMD structures are based on a geometry related to that of the cantilever beam, but instead use a torsion beam which is connected at two points rather than through a single metal hinge. [Figure 9.26](#) shows a top view of the metalized pixel. As shown in part (a) of the figure, the torsion rod connects the mirror to supports at the ends of one diagonal. Again the mirror is metallized and connected to a negative bias voltage.

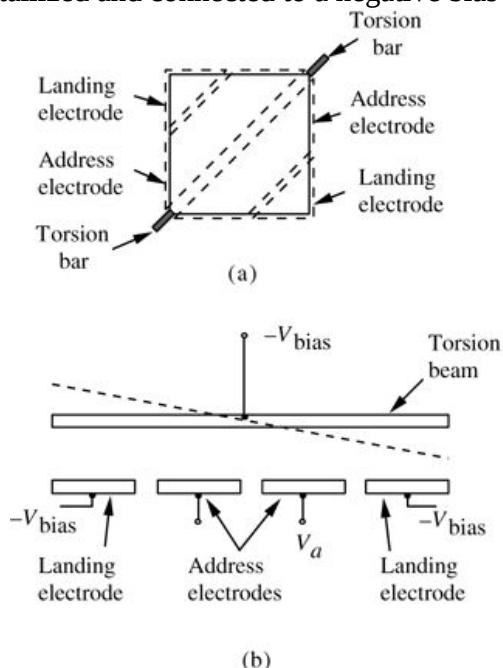


Figure 9.26
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 9.26 Torsion beam DMD: (a) top view and (b) side view.

Illustration a shows a square surrounded by a dotted line. A pair of dotted parallel lines runs from the bottom left corner to the top right corner. In line with these diagonal lines are projections from the two corners; these are torsion bars. Along somewhere between the diagonal and the top left corner is a pair of dotted parallel lines parallel to the diagonal. The exterior of the square's left

side above these lines is the landing electrode and the part below is address electrode. Similarly, along somewhere between the diagonal and the bottom right corner is a pair of dotted parallel lines parallel to the diagonal. The exterior of the square's right side below these lines is the landing electrode and the part above the lines is address electrode. Illustration b shows a row of four identical rectangles with their longer sides oriented horizontally. The first and the fourth are landing electrodes with negative V subscript bias voltage and the second and the third are address electrodes with V subscript a voltage. Located above and parallel to the electrodes is a torsion beam extending over all four electrodes. It is connected to a negative V subscript bias voltage. A downward sloping dotted line passes through the center of the torsion bar.

As shown in part (b) of the figure, two address electrodes exist for each such pixel, one on either side of the rotation axis. When one address electrode is activated with a positive voltage, the mirror twists in one direction, and when the other electrode is activated, the mirror twists in the opposite direction. Under each mirror element are two landing electrodes, held at the bias voltage, so that when the mirror tip twists so far as to hit the underlying landing electrode, there is no electrical discharge. The light incident on each pixel is deflected in either of two directions by the mirror when it is activated, and is not deflected when it is not activated. The device can be operated in either an analog mode, in which twist is a continuous function of applied address voltage, or in a digital mode, in which the device has either two stable states or three stable states, depending on the bias voltage applied [174].

A major advantage of this type of SLM technology is that it is silicon-based and compatible with the use of CMOS (complementary metal oxide semiconductor) drivers on the same substrate used for the SLM pixels. Both line-addressed DMDs and frame-addressed DMDs have been reported in sizes 128×128 128×128 and above. Devices of this type with as many as 1152×2048 1152×2048 pixels have been reported for use as high-definition TV (HDTV) displays. A second advantage is the ability of the device to operate at any optical wavelength where good mirrors can be fabricated in integrated form.

Measurements of the electrical and optical properties of this type of DMD have been reported in the literature [336]. Maximum deflection angles approaching 10° 10° are measured with applied voltages of about 16 volts. Deflection times of about $28 \mu\text{sec}$ $28 \mu\text{sec}$ were measured for an individual pixel, but this number depends on pixel size and can be shorter for smaller pixels. The resonant frequency of a pixel was found to be of the order of 10 kHz.

9.5 Acousto-Optic Spatial Light Modulators

The SLMs considered in the above sections are capable of modulating a two-dimensional wavefront, either in a continuous fashion or with a discrete two-dimensional array of elements. We turn now to an SLM technology that is most commonly one-dimensional, but which has been developed over a period of many years into a highly mature technology. This approach to wavefront modulation uses the interaction of a column of traveling acoustic waves with an incident coherent optical beam to modulate the properties of the transmitted optical wavefront. For alternative references that treat acousto-optic interactions and their applications in coherent optical systems, see, for example, [205], [26], and [356].

[Figure 9.27](#) illustrates two versions of acousto-optic SLMs, each operating in a different physical regime. In both cases, the acousto-optic cell consists of a transparent medium (e.g. a liquid or a transparent crystal) into which acoustic waves can be launched by a piezoelectric transducer. The transducer is driven by an RF voltage source and launches a compressional wave (or, in some cases, a shear wave) into the acoustic medium. The acoustic wave propagates in the medium through small local displacements of molecules (strain). Associated with these strains are small changes of the local refractive index, a phenomenon known as the acousto-optic effect or the photo-elastic effect. The driving voltage has an RF spectrum that is centered at some center frequency f_c with a bandwidth B about that center frequency.

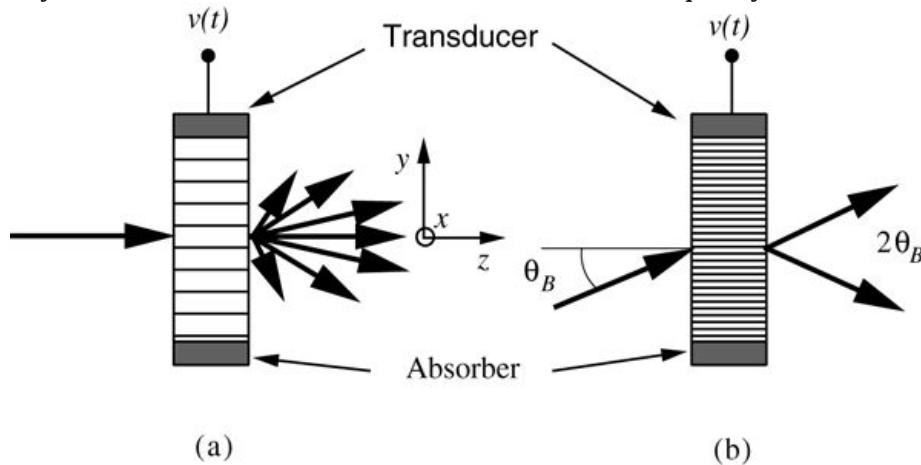


Figure 9.27
Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 9.27 Acousto-optic cells operating in the (a) Raman-Nath regime and the (b) Bragg regime.

Illustration a shows a cell represented by a rectangle with its longer side oriented vertically; parallel lines running horizontally across it section the rectangle into several parts. The topmost part, connected to voltage $V(t)$, is the transducer and the bottommost part is the absorber. A single rightward arrow points at the left vertical side of the cell; on the other side of the cell at the same level several arrows are shown to emerge. An accompanying coordinate plane has horizontal axis z , vertical axis y , and third axis x .

Illustration b is same as illustration a but with differences. The single rightward arrow pointing at the left vertical side of the cell is upward pointing, making an angle measuring theta subscript B with the perpendicular at the point where the arrow reaches the surface. On the right side of the cell is an upward pointing arrow parallel to the arrow on the left. Where the left arrow enters and where the right arrow exits are at the same horizontal level. On the right side, another arrow emerges at the same point; it is downward pointing such that the angle between the two arrows on the right is 2 theta subscript B.

A CW Drive Voltage

For a perfectly sinusoidal drive voltage of frequency f_c (i.e. a CW voltage), the transducer launches a sinusoidal traveling acoustic wave in the cell, which moves with the acoustic velocity V characteristic of the medium. This traveling wave induces a moving sinusoidal phase grating with period $\Lambda = V/f_c$ and interacts with the incident optical wavefront to produce various diffraction orders (cf. [Section 4.4.4](#)). However, there are two different regimes within which the acousto-optic interaction exhibits different properties, the *Raman-Nath* regime and the *Bragg* regime.

In the Raman-Nath regime, which is typically encountered for center frequencies in the range of several tens of MHz in cells that use liquid as the acoustic medium, the moving grating acts as a *thin* phase sinusoidal grating exactly as described in the example of [Section 4.4.4](#), with the one exception that, as a consequence of the grating motion through the cell, the various diffraction orders emerge from the cell with different optical frequencies. If the cell is illuminated normal to the direction of acoustic wave propagation, as shown in [Fig. 9.27\(a\)](#), the zero-order component remains centered at frequency v_o of the incident light, but higher-order components suffer frequency translations, which can be interpreted as Doppler shifts due to the motion of the grating.

Since the period of the grating is Λ , the q th diffraction order leaves the cell with angle θ_q with respect to the incident wave, where

$$\sin\theta_q = q\lambda\Lambda,$$

$$\sin\theta_q = q\frac{\lambda}{\Lambda},$$

(9-23)

λ being the optical wavelength within the acousto-optic medium. The optical frequency of the q th diffraction order can be determined from the Doppler-shift relation

$$v_q = v_o - V_c \sin\theta_q \approx v_o - q f_c.$$

$$\nu_q = \nu_o \left(1 - \frac{V}{c}\right) \sin\theta_q \approx \nu_o - q f_c.$$

(9-24)

Thus the optical frequency of the q th diffraction order is translated by q times the RF frequency, where q can be a positive or a negative number. In accord with the convention introduced in [Appendix D](#), the integer q is positive if the diffracted order is deflected in the

counter-clockwise direction, and negative if it is deflected in the clockwise direction. Thus in Fig. 9.27(a), the orders deflected downwards have negative values of q^q while those deflected upwards have positive values of q^q . The negative orders have optical frequencies that have been upshifted by $|q|f_c$. As for any thin sinusoidal phase grating, in the Raman-Nath regime the intensities associated with the various diffraction orders are proportional to the squares of the Bessel functions of the first kind, $J_q^2(\Delta\phi)$, where $\Delta\phi$ is the peak-to-peak phase modulation, as shown in Fig. 4.14.

For RF frequencies in the hundreds of MHz to the GHz range, and in acoustic media consisting of crystals, the thickness of the acousto-optic column compared with the acoustic wavelength introduces a preferential weighting for certain diffraction orders, and suppresses others. This effect is known as the *Bragg effect* and will be discussed at greater length in Chapter 11. For the moment it suffices to point out that in this regime the dominant diffraction orders are the zero order and a single first order (in the case shown in Fig. 9.27(b), a -1 -1 order). Diffraction into a first diffraction order occurs only when the angle of the incident beam, with respect to plane of the acoustic wavefronts, has the particular value θ_B satisfying

$$\sin\theta_B = \pm \frac{\lambda}{2\Lambda}$$

$$\sin\theta_B = \pm \frac{\lambda}{2\Lambda}$$

(9-25)

(cf. Fig. 9.27(b)), where again λ is the optical wavelength within the acoustic medium. An angle satisfying the above relation is known as a *Bragg angle*. Equivalently, if \vec{k}_i is the wave vector of the incident optical wave ($|\vec{k}_i| = 2\pi/\lambda$) and \vec{K} is the wave vector of the acoustic wave ($|\vec{K}| = 2\pi/\Lambda$), then

$$\sin\theta_B = \pm |\vec{K}| / 2|\vec{k}_i|.$$

$$\sin\theta_B = \pm \frac{|\vec{K}|}{2|\vec{k}_i|}.$$

(9-26)

The optical frequency of the first-order diffracted component is $v_o + f_c$ for the geometry shown in Fig. 9.27(b). The strength of the first-order component can be far greater than in the Raman-Nath regime, as discussed in more detail in Chapter 11.

An aid for visualizing the relations between the optical and acoustical wave vectors is a *wave vector diagram*, as shown in Fig. 9.28. For strong Bragg diffraction, the wave vector diagram must close as shown, a property that can be viewed as a statement of conservation of momentum.

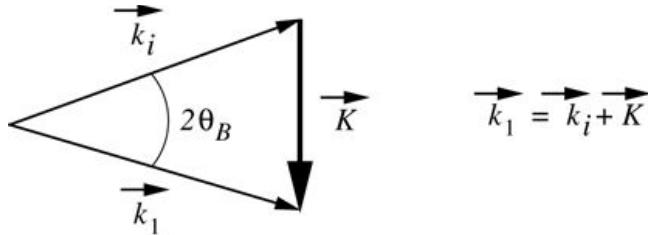


Figure 9.28

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 9.28 Wave vector diagram for Bragg interaction. \vec{k}_i is the incident optical wave vector, \vec{k}_1 is the optical wave vector of the component diffracted into the first diffraction order, and $K \rightarrow \vec{K}$ is the acoustical wave vector.

The illustration shows two vector rays of equal length originating at the same point, one upward sloping and rightward pointing labeled k_i and the other downward sloping and rightward pointing labeled k_1 . The angle between the two is $2\theta_B$. The vertical distance between the right extremes of the arrows is marked by a downward pointing arrow labeled vector K . An accompanying equation reads vector $k_1 = k_i + K$.

The boundary between the Raman-Nath regime and the Bragg regime is not a sharp one but is often described in terms of the so-called Q factor given by

$$Q = 2\pi\lambda_0 d n \Lambda^2$$

$$Q = \frac{2\pi\lambda_0 d}{n \Lambda^2}$$

(9-27)

where d is the thickness of the acoustic column in the z direction, n is the refractive index of the acousto-optic cell, and λ_0 is the vacuum wavelength of the light. If $Q < 2\pi$, operation is in the Raman-Nath regime, while if $Q > 2\pi$, operation is in the Bragg regime.

A Modulated Drive Voltage

Until now the voltage driving the acousto-optic cell has been assumed to be a perfect CW signal. We now generalize by allowing the voltage to be an amplitude and phase-modulated CW signal, of the form

$$v(t) = A(t) \sin(2\pi f_c t - \psi(t)),$$

$$v(t) = A(t) \sin[2\pi f_c t - \psi(t)],$$

(9-28)

where $A(t)$ and $\psi(t)$ are the amplitude and phase modulations, respectively. The refractive index disturbance generated by this applied voltage then propagates through the cell

with velocity V . With reference to [Fig. 9.27](#), if y is a coordinate running opposite to the direction of travel of the acoustic wave and is centered in the middle of the cell (as indicated in [Fig. 9.27](#)), and x is normal to the page of that figure, then at any instant of time t the distribution of refractive index perturbation in the cell can be written

$$\Delta n(y; t) = \sigma v y V + t - \tau_o,$$

$$\Delta n(y; t) = \sigma v \left(\frac{y}{V} + t - \tau_o \right),$$

(9-29)

where σ is a proportionality constant, $\tau_o = L/2V$ is the time delay required for acoustic propagation over half the length L of the cell, and we neglect the x dependence because it plays no role here or in what follows.

In the Raman-Nath regime, the optical wavefront is simply phase modulated by the moving refractive index grating, yielding a complex amplitude of the transmitted signal given by

$$U(y; t) = U_0 \exp[j2\pi d \lambda o A y V + t - \tau_o] \times \sin[2\pi f_c V + t - \tau_o - \psi(V + t - \tau_o)] \text{rect}\left(\frac{y}{L}\right),$$

$$U(y; t) = U_0 \exp\left\{j\frac{2\pi d \lambda o}{\lambda_o} A \left(\frac{y}{V} + t - \tau_o\right)\right. \\ \left. \times \sin[2\pi f_c \left(\frac{y}{V} + t - \tau_o\right) - \psi \left(\frac{y}{V} + t - \tau_o\right)]\right\} \text{rect}\left(\frac{y}{L}\right),$$

(9-30)

where U_0 is the complex amplitude of the incident monochromatic optical wave. Now the expansion

$$\exp[j\phi \sin\beta] = \sum_{q=-\infty}^{\infty} J_q(\phi) \exp(jq\beta)$$

$$\exp[j\phi \sin\beta] = \sum_{q=-\infty}^{\infty} J_q(\phi) \exp(jq\beta)$$

(9-31)

can be applied to the expression for $U(y; t)$. In addition, the peak phase modulation suffered by the optical wave as it passes through the cell is usually quite small, with the result that the approximation

$$J_{\pm 1}(\phi) \approx \pm \phi / 2$$

$$J_{\pm 1}(\phi) \approx \pm \phi / 2$$

holds for the first diffraction orders, which are the orders of main interest to us. As a result the complex amplitudes transmitted into the two first orders, represented by $U_{\pm 1}$, are given approximately by

$$U_{\pm 1} \approx \pm \pi d \lambda o U_0 A y V + t - \tau_o \times e^{\mp j\psi(y/V + t - \tau_o)} e^{\pm j2\pi y/\Lambda} e^{\pm j2\pi f_c(t - \tau_o)} \text{rect}(y/L),$$

$$U_{\pm 1} \approx \pm \frac{\pi \sigma d}{\lambda_o} U_o A \left(\frac{y}{V} + t - \tau_o \right) \\ \times e^{\mp j\psi(y/V + t - \tau_o)} e^{\pm j2\pi y/L} e^{\pm j2\pi f_c(t - \tau_o)} \text{rect} \frac{y}{L},$$

(9-32)

where the top sign corresponds to what we will call the “+1 + 1” diffraction order (diffracted upwards in [Fig. 9.27](#)) and the bottom sign corresponds to the “-1 - 1” order (diffracted downwards).

From [\(9-32\)](#) we see that the +1 + 1 diffracted order consists of a wavefront that is proportional to a moving version of the complex representation $A(y/V)e^{-j\psi(y/V)}$

$A(y/V)e^{-j\psi(y/V)}$ of applied voltage, while the -1 - 1 diffracted order contains the complex conjugate of this representation. The spatial argument of the moving field is scaled by the acoustic velocity V . A simple spatial filtering operation (see [Chapter 10](#)) can eliminate the unwanted diffraction orders and pass only the desired order. Thus the acousto-optic cell has acted as a one-dimensional spatial light modulator, transforming an electrical voltage modulation applied to the cell into an optical wavefront exiting the cell.

The discussion above has been framed in terms of Raman-Nath diffraction, but similar expressions for the diffraction orders are found in the case of Bragg diffraction, the primary difference lying in the strengths of the various orders. As mentioned earlier, the diffraction efficiency into one first order is generally considerably larger in the Bragg regime than in the Raman-Nath regime, and other orders are generally strongly suppressed by the diffraction process itself. Thus an acousto-optic cell operating in the Bragg regime again acts as a one-dimensional spatial light modulator, translating the applied voltage modulation into a spatial wavefront, albeit more efficiently than in the case of Raman-Nath diffraction.

9.6 Other Methods of Wavefront Modulation

A number of different materials for wavefront modulation are of special interest in holography. These include dichromated gelatin films, photopolymer films, photorefractive materials, and optical damage in glass. These methods will be discussed in the chapter on holography.

Problems - Chapter 9

1. 9-1. The interference between two plane waves of the form

$$U_1(x,y) = A \exp(j2\pi\beta_1 y) U_2(x,y) = B \exp(j2\pi\beta_2 y)$$

$$\begin{aligned} U_1(x, y) &= A \exp(j2\pi\beta_1 y) \\ U_2(x, y) &= B \exp(j2\pi\beta_2 y) \end{aligned}$$

is recorded on a photographic film. The film has an MTF of known form $M(f)$, and it is processed to produce a *positive* transparency with a gamma of -2^{-2} . This transparency (dimensions $L \times L$) is then placed in front of a positive lens with focal length f , is illuminated by a normally incident plane wave, and the distribution of intensity across the back focal plane is measured. The wavelength of the light is λ . Assuming that the entire range of exposure experiences the same photographic gamma, plot the distribution of light intensity in the rear focal plane, labeling in particular the relative strengths and locations of the various frequency components present.

2. 9-2. Show that, for a retardation of $\beta d = \pi$, the Jones matrices of (9-19) and (9-20) reduce to those of (9-21).
3. 9-3. A ferroelectric liquid crystal cell has a tilt angle of 22.5° . The input of the cell has a polarizer oriented parallel to the long molecular axis when the cell is in one of its two states, and the rear of the cell is a mirror. Using Jones matrices, show that if the retardation of the cell is one-quarter of a wave, the FLC cell can be used as a binary intensity modulator.
4. 9-4. An ideal grating has a profile that is illustrated by the triangular curve in Fig. P9.4. This ideal profile is approximated by a four-level quantized grating profile also shown in the figure. The peak-to-peak phase difference introduced by the continuous grating is exactly 2π radians.
1. Find the diffraction efficiencies of the $\pm 4, \pm 3, \pm 2, \pm 1, \pm 4, \pm 3, \pm 2, \pm 1$, and 0 orders of the continuous grating.
 2. Find the diffraction efficiencies of the same orders for the quantized grating.

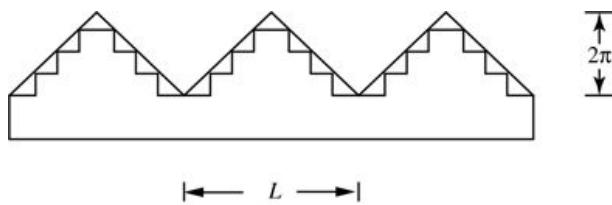


Figure P9.4

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure P9.4 Profiles of ideal and quantized gratings.

The illustration is a rectangle with its longer side oriented horizontally with three identical right angled isosceles triangles arranged in a row atop the rectangle such that their hypotenuses, each measuring L units, are along the top side of the rectangle. The height of the triangles as measured from the vertex to the hypotenuse is 2π .

10 Analog Optical Information Processing

The broad utility of linear systems concepts in the analysis of imaging systems is evident from the preceding chapters. However, if these concepts were useful *only* for analysis purposes, they would occupy a less important position in modern optics than they in fact enjoy today. Their true importance comes into full perspective only when the exciting possibilities of system *synthesis* are considered.

There exist many examples of the benefits reaped by the application of linear systems concepts to the synthesis of optical systems. One class of such benefits has arisen from the application of frequency-domain reasoning to the improvement of various types of imaging instruments. Examples of this type of problem are discussed in their historical perspective in [Section 10.1](#).

There are other applications that do not fall in the realm of imaging as such, but rather are more properly considered in the general domain of *information processing*. Such applications rest on the ability to perform general linear transformations of input data. In some cases, a vast amount of data may, by its sheer quantity, overpower the effectiveness of the human observer. A linear transformation can then play a crucial role in the *reduction* of large quantities of data, yielding indications of the particular portions of the data that warrant the attention of the observer. An example of this type of application is found in the discussion of *character recognition* ([Section 10.5](#)).

In the 1960s and 1970s, the power of computers did not quite match the power needed to manipulate very large arrays of two-dimensional data. The FFT algorithm, not made popular until the late 1960s, had a large influence on changing the capabilities available for digital processing of images of reasonable size. More important, however, has been the amazing progress of computers themselves, in terms of CPU performance (Moore's law), memory size available, and software algorithms. Processing problems that could only be handled by analog optical information processing in the 1960s and 1970s are now readily handled by even desktop computers. For these reasons, the importance of analog optical information processing has diminished significantly over the years.

Nonetheless, there are still some benefits to discussing the major ideas in optical information processing. First, the subject is intellectually interesting. The very fact that a coherent optical system can perform a two-dimensional Fourier transform in the time it takes light to propagate from a front focal plane of a lens to a back focal plane seems remarkable. Extending this capability to two-dimensional linear filtering likewise provokes thought about possible applications, particularly if the data to be processed is in optical form at the start. A second reason for discussing these ideas is that they can provoke ideas in other fields that may be useful. A good example of such a case is the arrayed waveguide grating discussed in [Section 12.5](#), in which a properly designed star coupler can play the role of a positive lens and Fourier transform one-dimensional data. Finally it should be mentioned that encountering the ideas, for example, of matched filtering and image restoration in the optical analog domain lends insight that is helpful even when the same operations will ultimately be performed digitally.

Many books devoted exclusively to the subject of optical information processing already exist (e.g. see [\[284\]](#), [\[344\]](#), [\[216\]](#), [\[53\]](#), [\[26\]](#), [\[175\]](#), and [\[356\]](#)). We shall limit our goals here to a

presentation of the most important analog optical information processing architectures and applications. We explicitly exclude from consideration the subject of “digital” or “numerical” optical computing. The reader interested in the digital domain can consult, for example, [[109](#)], [[367](#)], [[262](#)], [[243](#)], [[190](#)], or [[180](#)].

10.1 Historical Background

The history of Fourier synthesis techniques can be said to have begun with the first intentional manipulations of the spectrum of an image. Experiments of this type were first reported by Abbe in 1873 [1] and later by Porter in 1906 [286]. In both cases the express purposes of the experiments were verification of Abbe's theory of image formation in the microscope and an investigation of its implications. Because of the beauty and simplicity of these experiments, we discuss them briefly here.

10.1.1 The Abbe-Porter Experiments

The experiments performed by Abbe and Porter provide a powerful demonstration of the detailed mechanism by which coherent images are formed, and indeed the most basic principles of Fourier analysis itself. The general nature of these experiments is illustrated in Fig. 10.1. An object consisting of a fine wire mesh is illuminated by collimated, coherent light. In the back focal plane of the imaging lens appears the Fourier spectrum of the periodic mesh, and finally in the image plane the various Fourier components are recombined to form a replica of the mesh. By placing various obstructions (e.g. an iris, a slit, or a small stop) in the focal plane, it is possible to directly manipulate the spectrum of the image in a variety of ways.

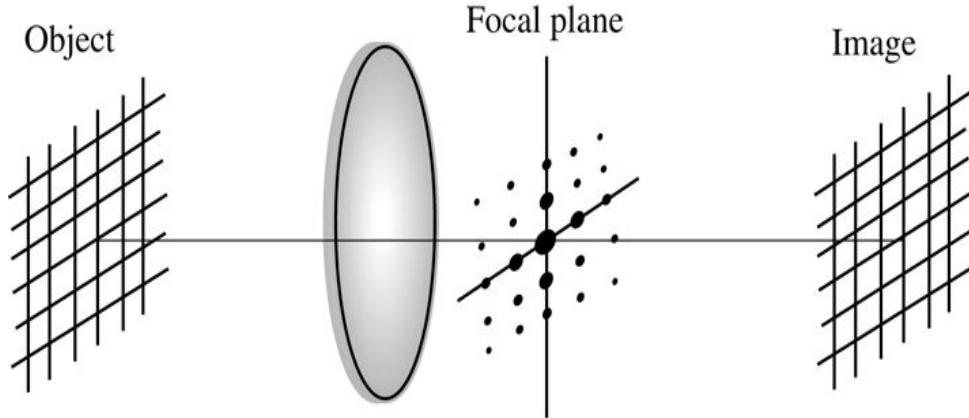


Figure 10.1

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 10.1 The Abbe-Porter experiment.

The illustration shows an object and an image, represented as sloping grids on the left and right extremes of the horizontal axis with a plane at the center. To the right of the plane is a focal plane with a vertical and a horizontal axis. Dots are arranged around the intersection of the axes in the shape of a sloping parallelogram. The dots are bigger at the center and smaller near the end.

Figure 10.2(a) shows a photograph of the spectrum of the mesh; Fig. 10.2(b) is the full image of the original mesh. The periodic nature of the object generates in the focal plane a series of isolated spectral components, each spread somewhat by the finite extent of the circular aperture within which the mesh is confined. Bright spots along the horizontal axis in the focal plane arise from complex-exponential components of the object that are directed horizontally (cf. Fig. 2.1);

bright spots along the vertical axis correspond to vertically directed complex-exponential components. Off-axis spots correspond to components directed at corresponding angles in the object plane.

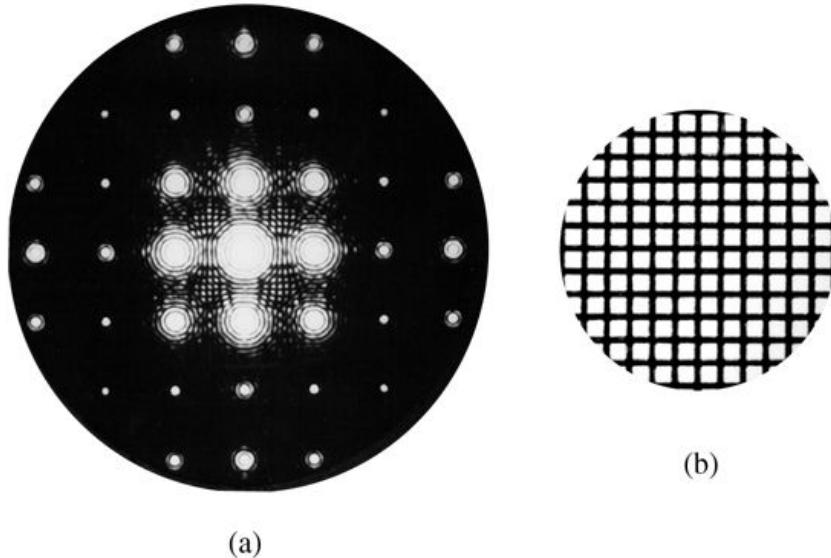


Figure 10.2

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 10.2 Photograph of (a) the spectrum of mesh and (b) the original mesh.

Illustration a shows a circular patch of darkness with 9 bright circular spots arranged in a 3 by 3 grid arrangement at the center. The dots in the middle row are bigger compared to the upper and lower rows. Outside the 3 by 3 grid, the pattern of dots continues but are of much lesser intensity. Illustration b shows a circle. Inside the circle, equidistant vertical and horizontal lines make a square mesh. The circle does not have an outline

The power of *spatial filtering* techniques is well illustrated by inserting a narrow slit in the focal plane to pass only a single row of spectral components. [Figure 10.3\(a\)](#) shows the transmitted spectrum when a horizontal slit is used. The corresponding image, seen in [Fig. 10.3\(b\)](#), contains only the *vertical* structure of the mesh; it is precisely the horizontally directed complex-exponential components that contribute to the structure in the image that is uniform vertically. The suppression of the horizontal structure is quite complete.

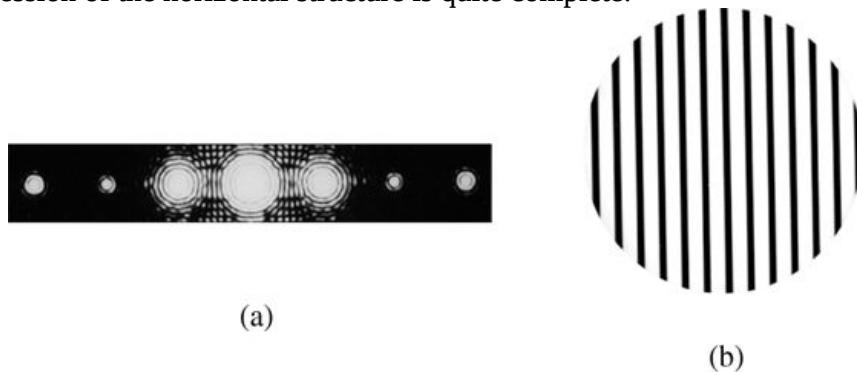


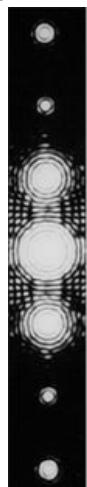
Figure 10.3

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 10.3 Mesh filtered with a horizontal slit in the focal plane. (a) Spectrum, (b) image.

Illustration a shows a lengthy rectangular patch of darkness. Three bright spots are shown at the center where the center spot is bigger in the size when compared to the spots on either side. Two dots of decreased intensity are shown on either side of the three bright dots. Illustration b shows a circle inside which vertical lines are shown.

When the slit is rotated by 90° to pass only the spectral column of [Fig. 10.4\(a\)](#), the image in part (b) of the figure is seen to contain only horizontal structure. Other interesting effects can also be readily observed. For example, if an iris is placed in the focal plane and stopped down to pass only the on-axis Fourier component, then with a gradual expansion of the iris the Fourier synthesis of the mesh can be watched step by step. In addition, if the iris is removed and a small central stop is placed on the optical axis in the focal plane to block only the central order or “zero-frequency” component, then a contrast reversal can be seen in the image of the mesh (see [Prob. 10-1](#)).



(a)

(b)

Figure 10.4

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 10.4 Mesh filtered with a vertical slit in the focal plane. (a) Spectrum, (b) image.

Illustration a shows a vertical rectangular patch of darkness with a bright spot at the center. On either side of the bright spot, a spot is shown in a slightly smaller size but with the same brightness intensity. Another two small dots of decreased intensity are shown on either side of the three bright dots. Illustration b shows a circle with thick horizontal lines inside.

10.1.2 The Zernike Phase-Contrast Microscope

Many objects of interest in microscopy are largely transparent, thus absorbing little or no light (e.g. an unstained bacterium). When light passes through such an object, the predominant effect is the generation of a spatially varying phase shift; this effect is not directly observable with a conventional microscope since detectors respond only to light intensity. A number of techniques for viewing such objects have been known for many years; these include interferometric techniques, the *central dark ground method* in which a small stop is used on the optical axis in the focal plane to block only the zero-frequency spectral component (see [Prob. 10-2](#)), and the

schlieren method in which all spectral components to one side of the zero-frequency component are excluded (see [Prob. 10-3](#)). All these techniques suffer from a similar defect—the observed intensity variations are *not* linearly related to the phase shift and therefore cannot be taken as directly indicative of the thickness variations of the object.

In 1935, [Frits Zernike \[381\]](#) proposed a new *phase contrast* technique which rests on spatial filtering principles and has the advantage that the observed intensity *is* (under certain conditions to be discussed) linearly related to the phase shift introduced by the object.¹ For this invention, Zernike won the 1953 Nobel Prize in Physics. In what follows we treat his idea in some detail.

Suppose that a transparent object, with amplitude transmittance

$$\begin{aligned} tA(\xi, \eta) &= \exp[j\phi(\xi, \eta)] \\ t_A(\xi, \eta) &= \exp[j\phi(\xi, \eta)] \end{aligned} \tag{10-1}$$

is coherently illuminated in an image-forming system. For mathematical simplicity we assume a magnification of unity and neglect the finite extent of the entrance and exit pupils of the system. In addition, a necessary condition to achieve linearity between phase shift and intensity is that the variable part of the object-induced phase shift, $\Delta\phi$, be small compared with 2π radians, in which case the crudest approximation to amplitude transmittance might be

$$\begin{aligned} tA(\xi, \eta) &= e^{j\phi_o} e^{j\Delta\phi} \approx e^{j\phi_o} [1 + j\Delta\phi(\xi, \eta)]. \\ t_A(\xi, \eta) &= e^{j\phi_o} e^{j\Delta\phi} \approx e^{j\phi_o} [1 + j\Delta\phi(\xi, \eta)]. \end{aligned} \tag{10-2}$$

In this equation we have neglected terms in $(\Delta\phi)^2$ and higher powers, assuming them to be zero in our approximation, and the quantity ϕ_o represents the average phase shift through the object, so $\Delta\phi(\xi, \eta)$ by definition has no zero-frequency spectral component. Note that the first term on the right of [Eq. \(10-2\)](#) represents a strong wave component that passes through the sample suffering a uniform phase shift ϕ_o , while the second term generates weaker diffracted light that is deflected away from the optical axis.

The image produced by a conventional microscope could be written, in our approximation, as

$$I_i \approx 1 + j\Delta\phi^2 \approx 1$$

$$I_i \approx |1 + j\Delta\phi|^2 \approx 1$$

where, to remain consistent with our approximation, the term $(\Delta\phi)^2$ has been replaced by zero. Zernike realized that the diffracted light arising from the phase structure is not observable in the image plane because it is in *phase quadrature* with the strong background, and that if this phase-quadrature relation could be modified, the two terms might interfere more directly to produce observable variations of image intensity. Recognizing that the background is brought to focus on the optical axis in the focal plane while the diffracted light, arising from higher spatial

frequencies, is spread away from the optical axis, he proposed that a phase-changing plate be inserted in the focal plane to modify the phase relation between the focused and diffracted light.

The phase-changing plate can consist of a glass substrate on which a small transparent dielectric dot has been deposited.² The dot is centered on the optical axis in the focal plane and has a thickness and index of refraction such that it retards the phase of the focused light by either $\pi/2$ radians or $3\pi/2$ radians relative to the phase retardation of the diffracted light. In the former case the intensity in the image plane becomes

$$I_i = | \exp [j(\pi/2) + j\Delta\phi/2] + j\Delta\phi/2 |^2 = |j(1 + \Delta\phi)|^2 \approx 1 + 2\Delta\phi \quad (10-3)$$

while in the latter case we have

$$I_i = | \exp [j(3\pi/2) + j\Delta\phi/2] + j\Delta\phi/2 |^2 = |-j(1 - \Delta\phi)|^2 \approx 1 - 2\Delta\phi. \quad (10-4)$$

Thus the image intensity has become linearly related to the variations of phase shift $\Delta\phi$. The case of (10-3) is referred to as *positive phase contrast* while the case of Eq.(10-4) is referred to as *negative phase contrast*. It is also possible to improve the contrast of the phase-induced variations of intensity in the image by making the phase-shifting dot partially absorbing (see Prob. 10-4).

The phase-contrast method is one technique for converting a spatial phase modulation into a spatial intensity modulation. The reader with a background in communications may be interested to note that one year after Zernike's invention a remarkably similar technique was proposed by [E.H. Armstrong](#)[10] for converting amplitude-modulated electrical signals into phase-modulated signals. As we have seen in [Chapter 7](#) and will continue to see in this chapter, the disciplines of optics and electrical engineering were to develop even closer ties in the years to follow.

10.1.3 Improvement of Photographs: Maréchal

In the early 1950s, workers at the Institut d'Optique, Université de Paris, became actively engaged in the use of coherent optical filtering techniques to improve the quality of photographs. Most notable was the work of A. Maréchal, whose success with these techniques was to provide a strong motivation for future expansion of interest in the optical information processing field.

Maréchal regarded undesired defects in photographs as arising from corresponding defects in the optical transfer function of the incoherent imaging system that produced them. He further reasoned that if the photographic transparencies were placed in a coherent optical system, then by insertion of appropriate attenuating and phase-shifting plates in the focal plane, a *compensating filter* could be synthesized to at least partially remove the undesired defects. While the optical transfer function of the initial imaging system might be poor, the product of that transfer function with the (amplitude) transfer function of the compensating system would hopefully yield an overall frequency response that was more satisfactory.

A variety of types of improvements to photographs were successfully demonstrated by Maréchal and his co-workers. For example, it was shown that small details in the image could be strongly emphasized if the low-frequency components of the object spectrum were simply attenuated. Considerable success was also demonstrated in the removal of image blur. In the latter case, the original imaging system was badly defocused, producing an impulse response which (in the geometrical-optics approximation) consisted of a uniform circle of light. The corresponding optical transfer function was therefore of the form

$$\mathcal{H}(\rho) \approx 2J_1(\pi a\rho)\pi a\rho,$$

$$\mathcal{H}(\rho) \approx 2\frac{J_1(\pi a\rho)}{\pi a\rho},$$

where a is a constant and $\rho = \sqrt{f_x^2 + f_y^2}$. The compensating filter was synthesized by placing both an absorbing plate and a phase shifting plate in the focal plane of the coherent filtering system, as shown in Fig. 10.5(a). The absorbing plate attenuated the large low-frequency peak of \mathcal{H} , while the phase-shifting plate shifted the phase of the first negative lobe of \mathcal{H} by 180° . The original and compensated transfer functions are illustrated in Fig. 10.5(b).

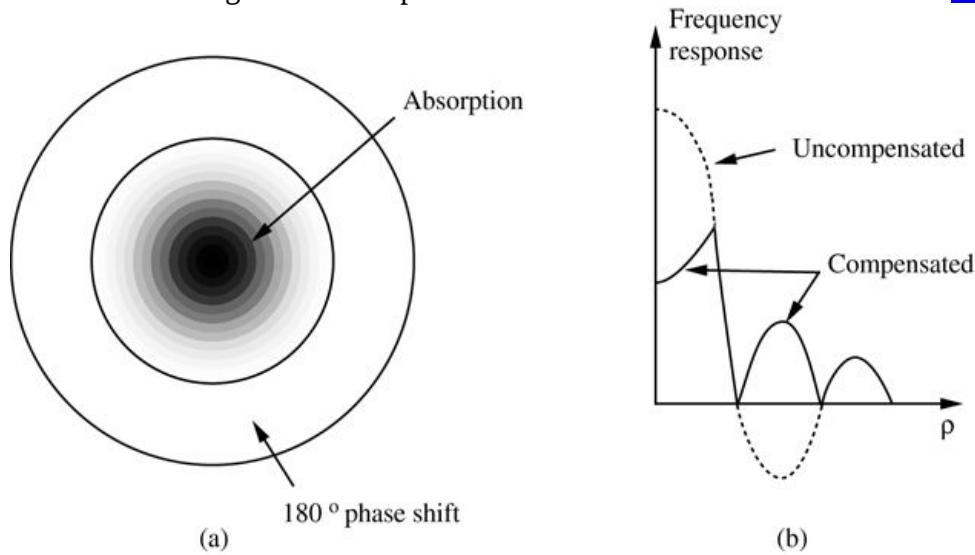


Figure 10.5

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 10.5 Compensation for image blur. (a) Focal-plane filter; (b) transfer functions.

Image a shows a dark spot at the center which is labeled absorption. Surrounding the dark spot, shaded portions of lighter shades are shown. Two concentric circles are shown surrounding the shaded portions. The region between the first and the second ring is labeled 180 degree shift. Image b is a graph in which the vertical axis represents frequency response and the horizontal axis represents rho. A dotted line starts at a point near the upper end of the vertical axis and then reduces until it reaches the midpoint of the vertical axis after which a solid line begins and reduces gradually and meets the horizontal axis. A line begins from a point below the midpoint of the vertical axis and touches the solid curve which begins at the midpoint of the vertical axis. Then the line again increases gradually and reaches a peak value and decreases and touches the horizontal

axis and again increases and reaches the peak value which is less than the second peak value after which it decreases and touches the horizontal line. A dotted graph is plotted from the point at which the solid line touches the horizontal axis for the first time and gradually increases in the opposite direction and reaches the peak value and then decreases and touches the horizontal line at the same point where the solid line touches the horizontal axis for the second time. The dotted line is labeled uncompensated line whereas the solid line is labeled compensated line.

As an additional example, it was shown that the periodic structure associated with the halftone process used in printing photographs, for example, in newspapers, could be suppressed by a simple spatial filter. The halftone process is, in many respects, similar to the periodic sampling procedures discussed in [Section 2.4](#). The spectrum of a picture printed in this fashion has a periodic structure much like that illustrated in [Fig. 2.6](#). By inserting an iris in the focal plane of the filtering system, it is possible to pass only the harmonic zone centered on zero frequency, thereby removing the periodic structure of the picture while passing all of the desired image data.

Notice a common requirement in all the applications mentioned above: a picture or photograph taken in incoherent light is filtered in a system that uses coherent light. To ensure that linear systems are used, and therefore that transfer function concepts remain valid, it is necessary that *the amplitude introduced into the coherent system be proportional to the intensity of the image we wish to filter*.

10.1.4 Application of Coherent Optics to More General Data Processing

While the early 1950s were characterized by a growing realization on the part of physicists that certain aspects of electrical engineering were of particular relevance to optics, the late 1950s and early 1960s saw a gradual realization on the part of electrical engineers that spatial filtering systems might be usefully employed in their more general data-processing problems. The potentials of coherent filtering were particularly evident in the field of radar signal processing and were exploited at an early stage by L.J. Cutrona and his associates at the University of Michigan Radar Laboratory. The publication of the paper “Optical data processing and filtering systems” [\[81\]](#) by the Michigan group in 1960 stimulated much interest in these techniques among electrical engineers and physicists alike. One of the most successful early applications of coherent filtering in the radar realm has been to the processing of data collected by synthetic aperture radar systems [\[82\]](#), a problem that is now solved exclusively by digital processing. A survey of the literature from the mid-1960s shows application of coherent processing techniques in such widely diverse fields as, for example, Fourier spectroscopy [\[334\]](#) and seismic-wave analysis [\[179\]](#).

10.2 Coherent Optical Information Processing Systems

When coherent illumination is used, filtering operations can be synthesized by direct manipulation of the complex amplitude appearing in the back focal plane of a Fourier transforming lens. Examples of this type of processing have already been seen in the discussion of the phase-contrast microscope (Zernike) and the filtering of photographs (Maréchal). In this section we outline the system architectures used for coherent optical information processing, and point out some of the difficulties encountered in attempting to synthesize general complex filters.

10.2.1 Coherent System Architectures

Coherent systems, being linear in complex amplitude, are capable of realizing operations of the form

$$I(x,y)=K \iint_{-\infty}^{\infty} g(\xi,\eta) h(x-\xi,y-\eta) d\xi d\eta$$

$$I(x,y) = K \left| \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(\xi,\eta) h(x-\xi,y-\eta) d\xi d\eta \right|^2.$$

(10-5)

There are many different system configurations that can be used to realize this operation, three of which are shown in [Fig. 10.6](#).

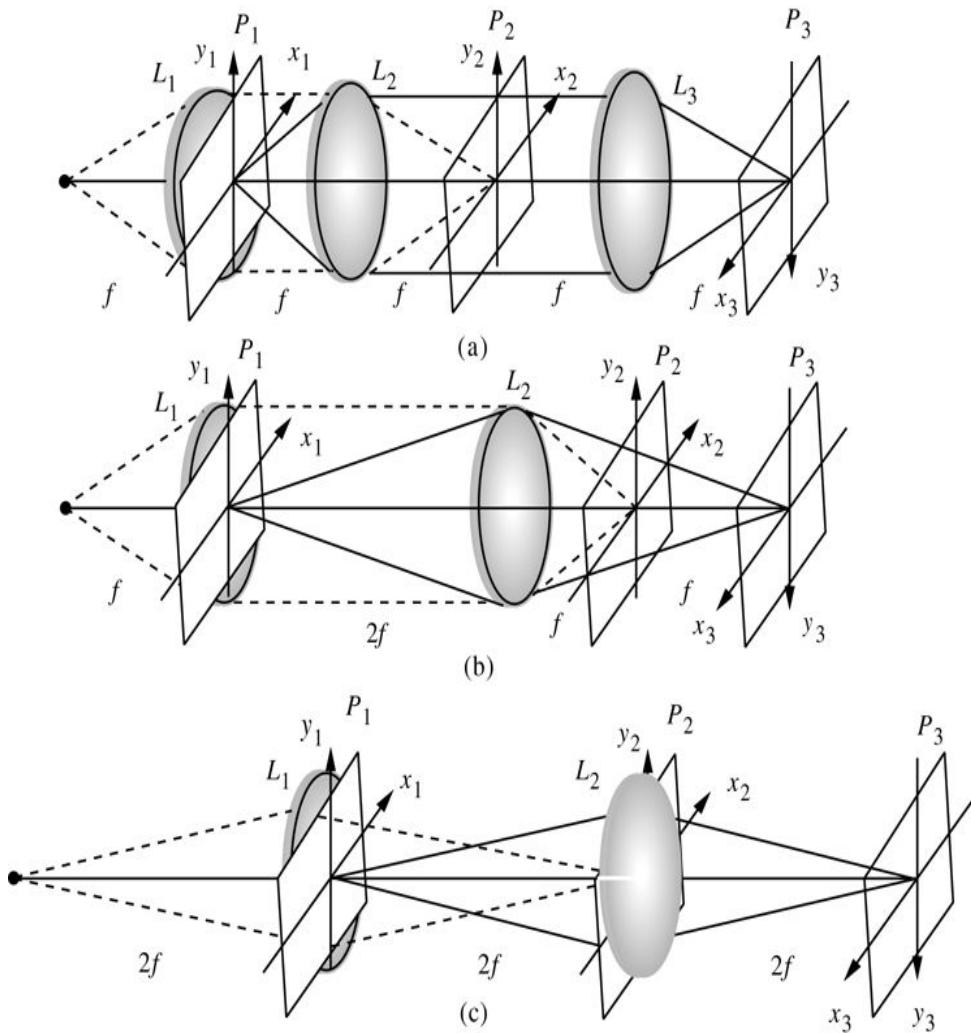


Figure 10.6

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 10.6 Architectures for coherent optical information processing.

Illustration a shows three planes \$P_1\$, \$P_2\$, and \$P_3\$ and three lens \$L_1\$, \$L_2\$, and \$L_3\$, in that order from left to right along a common horizontal axis. Lens \$L_1\$ and plane \$P_1\$ are adjoining. Lens \$L_2\$ is between planes \$P_1\$ and \$P_2\$ and Lens \$L_3\$ is between planes \$P_2\$ and \$P_3\$. The vertical axis of Plane \$P_1\$ is upward and labeled \$y_1\$. The vertical axis of Plane \$P_2\$ is upward and labeled \$y_2\$. The vertical axis of Plane \$P_3\$ is downward and labeled \$y_3\$. The third axis of plane \$P_1\$ is \$X_1\$ and points in the upper right direction. The third axis of plane \$P_2\$ is \$X_2\$ and points in the upper right direction. The third axis of plane \$P_3\$ is \$X_3\$ and points in the lower left direction. Two dotted lines connect the starting point of the horizontal axis with the upper and lower ends of lens \$L_1\$. The distance between the starting point of the horizontal axis and lens \$L_1\$ is \$f\$. Dotted lines connect the upper and lower ends of lens \$L_1\$ with the upper and lower ends of lens \$L_2\$, respectively. The distance between \$L_1\$ and \$L_2\$ is \$f\$. Two solid lines connect the center of Plane \$P_1\$ with the upper and lower ends of lens \$L_2\$, respectively. Two dotted lines connect the upper and lower ends of lens \$L_2\$ with the center of Plane \$P_2\$. Two solid lines connect the upper and lower ends of lens \$L_2\$ with the upper and lower ends of lens \$L_3\$, respectively. Two solid lines connect the upper and lower ends of lens \$L_3\$ with the center

of Plane P3. The distance between L2 and P2 is f, the distance between P2 and L3 is marked f and the distance between L3 and P3 is also marked f.

The second section shows three planes P1, P2, and P3 and two lens L1 and L2 with a common horizontal axis. Plane P1 and lens L1 are adjoining and lens L2 is placed between planes P1 and P2. The vertical axis of Plane P1 is upward and labeled y1. The vertical axis of Plane P2 is upward and labeled y2. The vertical axis of Plane P3 is downward and labeled y3. The third axis of plane P1 is labeled X1 and points in the upper right direction. The third axis of plane P2 is labeled X2 and points in the upper right direction. The third axis of plane P3 is labeled X3 and points in the lower left direction. Two dotted lines connect the starting point of the horizontal axis point with the upper end and the lower end of lens L1, respectively. Two dotted lines connect the upper end and the lower end of Lens L1 with the upper and lower end of lens L2, respectively. Two dotted lines emerge from the upper and lower ends of lens L2 and end at the center of Plane P2. Two lines emerge from the center of the Plane P1 and end at the upper and lower ends of lens L2, respectively. Two lines from the upper and lower ends of lens L2 pass through Plane P2 and end at the center of the Plane P3. The distance between the starting point of the horizontal line and lens L1 is f. The distance between L1 and L2 is 2f. The distance between L2 and P2 is f and the distance between P2 and P3 is f.

The third section shows three planes P1, P2, and P3 and two lens L1 and L2 with a common horizontal axis. Plane P1 and lens L1 are adjoining and lens L2 and Plane P2 are adjoining. The vertical axis of Plane P1 is upward and labeled y1. The vertical axis of plane P2 is upward and labeled y2. The vertical axis of Plane P3 is downward and labeled y3. The third axis of plane P1 is X1 and points in the upper right direction. The third axis of plane P2 is labeled X2 and points in the upper right direction. The third axis of plane P3 is labeled X3 and points in the lower left direction. The distance between the starting point of the horizontal axis and Plane P1 is 2f. The distance between planes P1 and P2 is 2f. The distance between Planes P2 and P3 is 2f. A dotted line from the starting point of the horizontal axis points to the top center of Plane P1, and a dotted line starts from the same point and ends at the center of lens L1. A dotted line from the starting point of the horizontal axis points toward the bottom center of Plane P1, and a dotted line starts from the same point and ends at the center of lens L1. A line from the center of plane P1 points toward the top center of plane P2 and a line starts from the same point and ends at the center of plane P3. A line from the point of plane P1 points toward the bottom center of plane P2, and a line starts from the same point and ends at the center of plane P3.

The system shown in part (a) of the figure is conceptually the most straightforward and is often referred to as a “4f” filtering architecture, due to the fact that there are four separate distances of length f separating the input plane from the output plane. Light from the point source S is collimated by lens L_1 . In order to minimize the length of the system, the input transparency, having amplitude transmittance $g(x_1, y_1)$, is placed against the collimating lens in plane P_1 . One focal length beyond the input is a Fourier transforming lens L_2 , in the rear focal plane (P_2) of which is placed a transparency to control the amplitude transmittance through that plane. An amplitude $k_1 G(x_2/\lambda f, y_2/\lambda f)$ is incident on this plane, where G is the Fourier transform of g and k_1 is a constant. A filter is inserted in plane P_2 to manipulate the spectrum of g . If H represents the desired transfer function, then the amplitude transmittance of the frequency-plane filter should be

$$t_A(x_2, y_2) = k_2 H x_2 \lambda f, y_2 \lambda f.$$

$$t_A(x_2, y_2) = k_2 H \left(\frac{x_2}{\lambda f}, \frac{y_2}{\lambda f} \right).$$

(10-6)

The field behind the filter is thus GH. After one additional focal length, lens L₃ is placed, the purpose of which is to again Fourier transform the modified spectrum of the input, producing a final output in its rear focal plane, P₃. Note that the output appears inverted in plane P₃ due to the imaging operation, or equivalently due to the fact that a sequence of two Fourier transforms has been used, rather than one transform followed by its inverse. This awkwardness can be remedied by reversing the final coordinate system (x₃, y₃), as shown in the figure, in which case the output in plane P₃ is as described by (10-5). For simplicity, the focal lengths of all three lenses have been assumed to be f, and the total length of the system is seen to be 5f. This architecture has the disadvantage that vignetting can occur during the first Fourier transform operation.

The system shown in part (b) of the figure has the same length as the previous system, but uses one fewer lens. Again lens L₁ collimates the light from the point source S, and again the input transparency is placed against L₁ to minimize the length of the system. Placed at distance 2f from the input, lens L₂ now performs both the Fourier transforming and the imaging operations, with the spectrum of the input appearing in the rear focal plane P₂ (where the Fourier filter transparency is placed) and the filtered image appearing one additional focal length beyond the focal plane, in plane P₃. Since the object and image distances are both 2f, the magnification of the system is unity. Note that in this geometry, the spectrum of the input has associated with it a quadratic-phase factor of the form exp-jk2f(x₂₂+y₂₂)

$\exp \left[-j \frac{k}{2f} (x_2^2 + y_2^2) \right]$, since the input is not in the front focal plane of the lens. The length of this system remains 5f, as before.

There are two practical disadvantages of this second geometry. First, as compared with system (a), the input is now twice the distance from lens L₂, and therefore the vignetting will be even worse than that encountered with system (a). A second disadvantage arises from the approximations that led to (6-35) in the analysis of the coherent imaging properties of a thin lens. In that formulation, we found it necessary to assume that the amplitude of the image at any particular point consisted of contributions from only a small region surrounding the geometrical object point. If the filtering operation represented by the transfer function H is of high space-bandwidth product, then the impulse response h will extend over a sizable area, and the output of this system must be regarded as a filtered version of the function g(x₁, y₁)expjk4f(x₁₂+y₁₂)

$$g(x_1, y_1) \exp \left[j \frac{k}{4f} (x_1^2 + y_1^2) \right]$$

rather than simply of g(x₁, y₁)^{g(x₁, y₁)}. This problem is not

encountered with system (a), which casts an image of *plane* P_1 onto a *plane* P_3 , rather than of a sphere onto a sphere. This difficulty can be corrected by adding an additional positive lens with focal length $2f$ in contact with the object, thus canceling the troubling quadratic-phase factor. This additional lens also results in movement of the frequency plane from f behind lens L_2 to coincidence with that lens, but the location of image plane P_3 is not affected.

As a final example which has merit (but by no means the only other system geometry possible), consider the system shown in part (c) of the figure. Again only two lenses are used.

Lens L_1 now serves as both a lens for collecting the light from the point source S and as a Fourier transforming lens. The input is placed in plane P_1 in contact with lens L_1 . This lens images the source onto the frequency plane P_2 , where the filter transparency is placed. The magnification of this imaging operation as shown is unity. The second lens, L_2 , is also placed in this plane and images the input onto the output plane P_3 with unity magnification. Note that this system has no vignetting problems, and the quadratic-phase factor across the input plane (mentioned above) is canceled by the converging illumination. The disadvantage is that the system is now of length $6f$ rather than $5f$.

Finally we mention that it is also possible to arrange a coherent system to process a stacked array of one-dimensional inputs rather than a single two-dimensional input. An example of these so-called *anamorphic* processors³ is shown in Fig. 10.7. The collimating lens L_1 is followed by the input data in plane P_1 . The input data consists of an array of one-dimensional transmittance functions each running horizontally. A cylindrical lens L_2 follows, placed one focal length f from P_1 and having power only in the vertical dimension. At distance $2f$ beyond L_2 is placed a spherical lens L_3 which again has focal length f . The “frequency plane” now appears at P_2 , where an array of one-dimensional spectra is found. The lens combination L_2, L_3 has performed a double Fourier transformation in the y direction, thus imaging in the vertical direction. Since L_2 exerts no power in the x direction, the spherical lens L_3 Fourier transforms in the horizontal dimension, up to a phase factor exp-

$$jkfx22 \exp\left(-j\frac{k}{f}x_2^2\right)$$

across P_2 . This phase factor can be removed by placing a negative cylindrical lens of focal length $f/2$ immediately in front of P_2 , thus canceling the phase curvature. If the input array is the set of transmittance functions $g_k(x_1), k=1,2,\dots,K$ $g_k(x_1), k = 1, 2, \dots, K$, then across P_2 we find displayed the corresponding set of transforms $G_k(x_2), k=1,2,\dots,K$ $G_k(x_2), k = 1, 2, \dots, K$. with the vertical order inverted by the imaging operation.

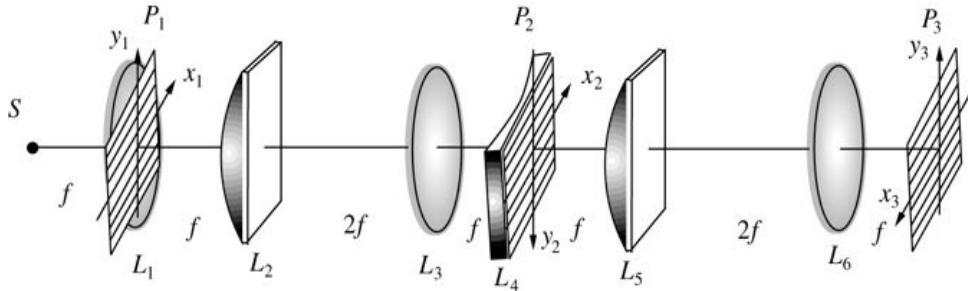


Figure 10.7

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 10.7 Example of an anamorphic processor.

The illustration shows planes P_1 , P_2 , and P_3 with a common horizontal axis whose left extreme is point S . The planes are represented by three sloping parallelograms with horizontal lines inside. The vertical axis of P_1 is upward and labeled y_1 . The vertical axis of P_2 is downward and labeled y_2 . The vertical axis of P_3 is upward and labeled y_3 . The third axis of plane P_1 is X_1 and points in the upper right direction. The third axis of plane P_2 is X_2 and points in the upper right direction. The third axis of plane P_3 is X_3 and points in the lower left direction. The distance between S and lens L_1 is f . Lens L_1 and plane P_1 are adjoining. The distance between L_1 and L_2 is f . Lens L_3 is placed at a distance of $2f$ from L_2 . The distance between L_3 and L_4 is f . Plane P_2 is shown as a sloping parallelogram adjoining lens L_4 . Lens L_5 is placed at a distance from L_4 . The distance between L_5 and L_6 is $2f$ and the distance between L_6 and plane P_3 is f .

A linear array of one-dimensional filters may now be introduced in plane P_2 . The lens pair L_5, L_6 again images in the y direction and Fourier transforms in the x direction, thus retaining the array structure but returning the original functions to the “space domain.” The phase factor associated with the final Fourier transform is generally of no concern.

10.2.2 Constraints on Filter Realization

While coherent systems are in general more flexible and have greater data-handling capacity than most incoherent systems, nonetheless there are limitations to the types of operations that can be realized with simple frequency-plane filters of the kind used earlier by Maréchal. More sophisticated techniques for realizing frequency-plane masks, based on interferometric recording, are free from some of these limitations, as will be discussed in the section to follow.

Before 1963, the conventional means for realizing a given transfer function had been the insertion of independent amplitude and phase masks in the frequency plane. The amplitude transmittance was controlled by a photographic plate, presumably immersed in a liquid gate. The phase transmittance was controlled by insertion of a transparent plate with an appropriately varying thickness. Such plates could be ruled on a substrate, much as diffraction gratings are ruled, or deposited on a flat plate using thin-film coating techniques. All such methods are rather cumbersome, and could be successfully employed only when the desired pattern of phase control was rather simple, e.g. binary and of simple geometric structure.

[Figure 10.8](#) shows the regions of the complex plane that can be reached by the transfer functions of coherent optical systems under different constraints on the frequency-plane transparency. As shown in (a), when only an absorbing transparency is used, the reachable region is limited to the positive real axis between 0 and 1. If binary phase control is added to this

absorbing transparency, then the reachable region is extended to the region $-1 \leq 1$ to 1 on the real axis, as shown in (b). If a pure phase filter is used, with arbitrary achievable values of phase, then the values of the transfer function would be restricted to the edge of the unit circle, as shown in (c). Finally part (d) of the figure shows the region of the complex plane that one would generally desire to reach if there were no constraints, namely the entire interior and edge of the unit circle.

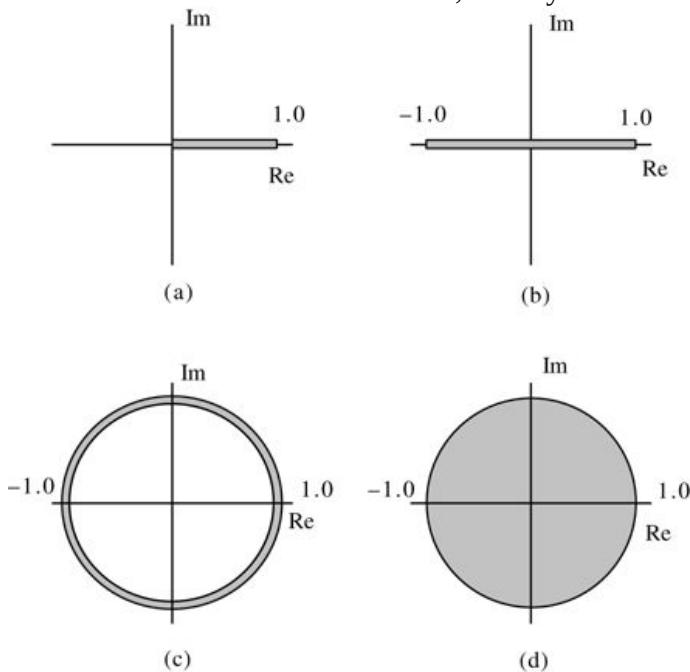


Figure 10.8

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 10.8 Reachable regions of the frequency plane for (a) a purely absorbing filter, (b) an absorbing filter and binary phase control, (c) a pure phase filter, and (d) a filter that achieves arbitrary distributions of absorption and phase control.

Graph a has the horizontal axis labeled Re and the vertical axis labeled Im. A shaded rectangular strip extends from the origin to the point 1.0 on the horizontal axis.

Graph b has the horizontal axis labeled Re and the vertical axis labeled Im. A shaded rectangular strip extends on the horizontal axis from +0.1 to minus 0.1 on the negative axis.

Graph c has the horizontal axis labeled Re and the vertical axis labeled Im. Two concentric circles are shown with 0 as its center and they intersect the horizontal axis at two points between the range 1.0 on the positive axis and -1.0 on the negative axis and the portion between the two circles is shaded.

Graph d has the horizontal axis labeled Re and the vertical axis labeled Im. A shaded circle is shown with 0 as its center and it intersects the horizontal axis at two points between the range 1.0 on the positive axis and -1.0 on the negative axis.

It should be noted that, for even a very simple impulse response (such as one in the shape of the character "P," for example), the corresponding transfer function was (1) difficult to calculate (prior to the development of the fast Fourier transform algorithm for digital computation of spectra) and (2) far too complicated to be synthesized by these rather simple techniques.

In summary, the most severe limitation to the traditional coherent processor (prior to the invention of the methods to be discussed in the next section) arose from the difficulty of simultaneously controlling the amplitude and phase transmittances in any but very simple patterns.

Thus coherent optical filters were limited to those that had very simple transfer functions. It was not until 1963, with the invention of the interferometrically recorded filter, that this serious limitation was largely overcome, extending the domain of complex filters that could be realized to those with simple impulse responses.

10.3 The VanderLugt Filter

In 1963, A.B. VanderLugt of the University of Michigan's Radar Laboratory proposed and demonstrated a new technique for synthesizing frequency-plane masks for coherent optical processors [354], [355].⁴ The frequency-plane masks generated by this technique have the remarkable property that they can effectively control both the amplitude and phase of a transfer function, in spite of the fact that they consist only of patterns of *absorption*. By means of this technique, it is possible to largely overcome the two limitations to coherent processing systems mentioned above.

10.3.1 Synthesis of the Frequency-Plane Mask

The frequency-plane mask of the VanderLugt filter is synthesized with the help of an interferometric (or holographic—see [Chapter 11](#)) system, such as that shown in [Fig. 10.9](#). The lens L_1 collimates the light from the point source S . A portion of this light strikes the mask P_1 , which has an amplitude transmittance that is proportional to the desired *impulse response* h . The lens L_2 Fourier transforms the amplitude distribution h , yielding an amplitude

distribution $\frac{1}{\lambda f} H\left(\frac{x_2}{\lambda f}, \frac{y_2}{\lambda f}\right)$ incident on the recording medium,⁵ usually film. In addition, a second portion of the collimated light passes above the mask P_1 , strikes a prism P , and is finally incident on the recording plane at angle θ , as shown.

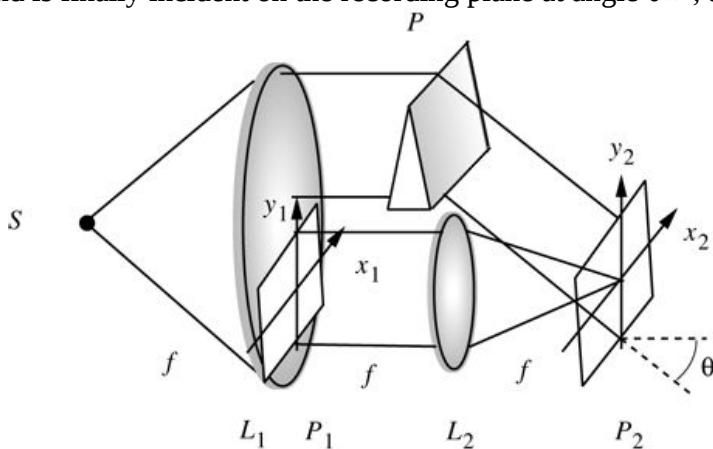


Figure 10.9

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 10.9 Recording the frequency-plane mask for a VanderLugt filter.

The illustration shows a point S. Two lines emerge from S and end at the upper and lower end of Lens L1. Plane P1 is shown as a sloping parallelogram. Planes P1 and Lens L1 are adjoining. The vertical axis of P1 is y1. The horizontal axis of P2 is x1. The distance between S and L1 is f. To the right of Lens L1 is a prism. A line emerges from the center of Lens L1 and ends at the base of

the prism. A line from the upper end of the lens touches the midpoint of the prism. Lens L2 is placed at a distance f from plane P1. A line starts from the top center of P1 and ends at the upper portion of L2 and a line from the bottom center of P2 ends at the bottom portion of L2. Plane P2 is placed at a distance f from L2. The vertical axis of P2 is y_2 and the horizontal axis is x_2 . Two lines from the upper and lower ends of L2 point toward the center of plane P2. A line from the base of the prism P points toward the bottom center of P2 and the line is extended as a dotted line beyond P2 and a horizontal dotted line is drawn from the bottom center of P2 at an angle theta from the previously mentioned dotted line. A line is shown from the center of the prism to a point in upper portion of P2 which is slightly away from the center.

The total intensity incident at each point on the recording medium is determined by the interference of the two mutually coherent amplitude distributions present. The tilted plane wave incident from the prism produces a field distribution

$$U_r(x_2, y_2) = r_o \exp(-j2\pi\alpha y_2),$$

$$U_r(x_2, y_2) = r_o \exp(-j2\pi\alpha y_2),$$

(10-7)

where the spatial frequency α is given by

$$\alpha = \sin\theta\lambda.$$

$$\alpha = \frac{\sin\theta}{\lambda}.$$

(10-8)

The total intensity distribution may therefore be written

$$\begin{aligned} I(x_2, y_2) &= r_o \exp(-j2\pi\alpha y_2) + 1/\lambda f H x_2 \lambda f, y_2 \lambda f^2 = r_o^2 + 1/\lambda f^2 H x_2 \lambda f, y_2 \lambda f^2 + r_o \lambda f H x_2 \lambda f, y_2 \lambda f \exp(j2\pi\alpha y_2) + r_o \lambda f H^* x_2 \lambda f, \\ &\quad y_2 \lambda f \exp(-j2\pi\alpha y_2). \end{aligned}$$

$$\begin{aligned} I(x_2, y_2) &= \left| r_o \exp(-j2\pi\alpha y_2) + \frac{1}{\lambda f} H \left(\frac{x_2}{\lambda f}, \frac{y_2}{\lambda f} \right) \right|^2 \\ &= r_o^2 + \frac{1}{\lambda^2 f^2} \left| H \left(\frac{x_2}{\lambda f}, \frac{y_2}{\lambda f} \right) \right|^2 + \frac{r_o}{\lambda f} H \left(\frac{x_2}{\lambda f}, \frac{y_2}{\lambda f} \right) \exp(j2\pi\alpha y_2) \\ &\quad + \frac{r_o}{\lambda f} H^* \left(\frac{x_2}{\lambda f}, \frac{y_2}{\lambda f} \right) \exp(-j2\pi\alpha y_2). \end{aligned}$$

(10-9)

Note that if the complex function H has an amplitude distribution A and a phase distribution ψ , that is, if

$$H x_2 \lambda f, y_2 \lambda f = A x_2 \lambda f, y_2 \lambda f \exp j \psi x_2 \lambda f, y_2 \lambda f,$$

$$H\left(\frac{x_2}{\lambda f}, \frac{y_2}{\lambda f}\right) = A\left(\frac{x_2}{\lambda f}, \frac{y_2}{\lambda f}\right) \exp\left[j\psi\left(\frac{x_2}{\lambda f}, \frac{y_2}{\lambda f}\right)\right],$$

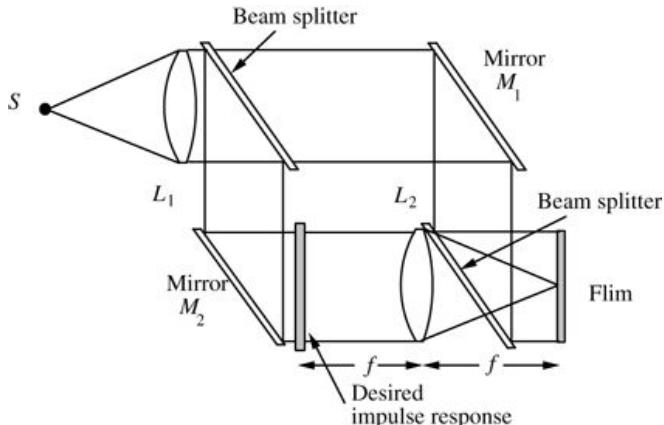
then the expression for \mathcal{I} can be rewritten in the form

$$\mathcal{I}(x_2, y_2) = r_o^2 + \frac{1}{\lambda^2 f^2} A^2 \left(\frac{x_2}{\lambda f}, \frac{y_2}{\lambda f} \right)$$

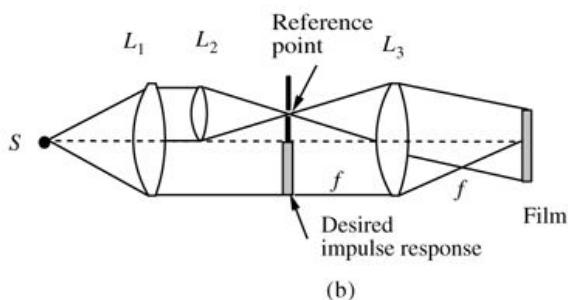
$$\begin{aligned} \mathcal{I}(x_2, y_2) &= r_o^2 + \frac{1}{\lambda^2 f^2} A^2 \left(\frac{x_2}{\lambda f}, \frac{y_2}{\lambda f} \right) \\ &\quad + \frac{2r_o}{\lambda f} A\left(\frac{x_2}{\lambda f}, \frac{y_2}{\lambda f}\right) \cos\left[2\pi\alpha y_2 + \psi\left(\frac{x_2}{\lambda f}, \frac{y_2}{\lambda f}\right)\right]. \end{aligned} \quad (10-10)$$

This form illustrates the means by which the interferometric process allows the recording of a complex function H on an intensity-sensitive detector: amplitude and phase information are recorded, respectively, as amplitude and phase modulations of a *high-frequency carrier* that is introduced by the relative angular tilt of the “reference” wave from the prism.

There are, of course, other optical systems that will produce the same intensity distribution as that of (10-10). [Figure 10.10](#) illustrates two additional possibilities. System (a) consists of a modified Mach-Zehnder interferometer. By tilting the mirror M_1 , a tilted plane wave is produced at the film plane. In the lower arm of the interferometer, the lens L_2 again Fourier transforms the desired impulse response. The final beam splitter allows the addition of these two waves at the recording plane.



(a)



(b)

Figure 10.10

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 10.10 Two alternative systems for producing the frequency-plane transparency (a) Modified Mach-Zehnder interferometer; (b) modified Rayleigh interferometer.

Illustration a shows a point on the left extreme labeled S. Two lines emerge from S and end at the upper and lower ends of Lens L1, respectively. The beam splitter is shown as a vertical lengthy rectangular strip in a sloping position. Two lines connect the upper and lower ends of L1 with the upper and lower ends of beam splitter. To the right of beam splitter, mirror M1 is also shown as a vertical lengthy rectangular strip in a sloping position. Two lines connect the upper and lower ends of beam splitter with the upper and lower ends of mirror M1. Mirror M2 is shown below the beam splitter. Two lines connect the upper and lower ends of the beam splitter with the upper and lower ends of mirror M2. A vertical rectangular bar is shown near the mirror M2 and is labeled “desired impulse response.” Lens L2 is shown at a distance f from the vertical rectangular bar. Two lines connect the top and bottom ends of the mirror M2 with the vertical rectangular bar and two lines from the rectangular bar point toward the upper and lower ends of Lens L2. Another beam splitter is shown adjoining L2. Two lines connect the upper and lower ends of mirror M1 with the second beam splitter. A film represented as a vertical rectangular bar is placed at a distance of f from L2. Two lines connect the top and bottom ends of the beam splitter with the top and bottom ends of film. Two lines from the upper and bottom portions of Lens L2 point toward the center of the film. Illustration b shows a point on the left extreme marked S. Two lines from S point toward the upper and lower ends of Lens L1. Another small Lens L2 is shown to the right of L1 and it is half the size of L1. A line from the top end of L1 points toward the top end of L2. A line from the center of L1 points toward the bottom end of L2. A horizontal dotted line runs from the point S up to the

film dividing the system in two equal halves. A vertical rectangular bar is shown next to Lens L2 which starts from the horizontal line and is labeled desired impulse response. The reference point is a thick vertical line above the vertical rectangular bar. A line starts from the upper portion of L2 and passes through the reference point and point toward the center of Lens L3. A line starts from the lower portion of L2 and passes through the reference point and point toward the upper end of Lens L3. A line starts from the lower end of L1 and ends at the end of the vertical rectangular bar and a line from the bar ends at the lower end of Lens L3. A line starts from the top end of L3 and ends at the upper end of the film and a line from the lower portion of L3 ends at the center of film. A line from a point slightly below the center of L3 points toward the bottom end of film. The distance between desired impulse response bar and L3 is f and the distance between L3 and film is f .

System (b), which is a modified Rayleigh interferometer, provides a third means for producing the same intensity distribution. The collimating lens L_1 is followed by a smaller lens L_2 , which focuses a portion of the collimated light to a bright spot in the front focal plane of lens L_3 . When the spherical wave generated by this “reference point” passes through L_3 , it is collimated to produce a tilted plane wave at the recording plane. The amplitude transmitted by the impulse response mask is Fourier transformed in the usual fashion. Thus an intensity distribution similar to (10-10) is again produced at the recording plane.

As a final step in the synthesis of the frequency-plane mask, the exposed film is developed to produce a transparency which has an amplitude transmittance that is proportional to the intensity distribution that was incident during exposure. Thus the amplitude transmittance of the filter is of the form

$$t_A(x_2, y_2) \propto r_o^2 + 1/\lambda^2 f^2 |H|^2 + r_o \lambda f H \exp(j2\pi a y_2) + r_o \lambda f H^* \exp(-j2\pi a y_2).$$

$$\begin{aligned} t_A(x_2, y_2) &\propto r_o^2 + \frac{1}{\lambda^2 f^2} \left| H \right|^2 + \frac{r_o}{\lambda f} H \exp(j2\pi a y_2) \\ &\quad + \frac{r_o}{\lambda f} H^* \exp(-j2\pi a y_2). \end{aligned}$$

(10-11)

Note that, aside from the simple complex-exponential factor, the third term of the amplitude transmittance is proportional to H and therefore exactly the form required to synthesize a filter with impulse response h . It remains to be demonstrated how that particular term of the transmittance can be utilized and the other terms excluded.

10.3.2 Processing the Input Data

Once the frequency-plane mask has been synthesized, it may be inserted in any of the processing systems shown previously in Fig. 10.6. To be specific, we focus on the system shown in part (a) of that figure. If the input to be filtered is $g(x_1, y_1)$, then incident on the frequency-plane

mask is a complex amplitude distribution given by $\frac{1}{\lambda f} G\left(\frac{x_2}{\lambda f}, \frac{y_2}{\lambda f}\right)$. The field strength transmitted by the mask then obeys the proportionality

$$U_2 \propto r_o^2 G + \frac{1}{\lambda^3 f^3} \left| H \right|^2 G + \frac{r_o}{\lambda^2 f^2} H G \exp(j2\pi\alpha y_2) + r_o \frac{H^*}{\lambda^2 f^2} G \exp(-j2\pi\alpha y_2).$$

$$\begin{aligned} U_2 &\propto \frac{r_o^2 G}{\lambda f} + \frac{1}{\lambda^3 f^3} \left| H \right|^2 G + \frac{r_o}{\lambda^2 f^2} H G \exp(j2\pi\alpha y_2) \\ &\quad + \frac{r_o}{\lambda^2 f^2} H^* G \exp(-j2\pi\alpha y_2). \end{aligned}$$

The final lens L_3 of Fig. 10.6(a) optically Fourier transforms U_2 . Taking note of the reflected coordinate system in plane P_3 as well as the scaling constants present in the Fourier transform operation, the field strength in that plane is found to obey the proportionality

$$U_3(x_3, y_3) \propto r_o^2 g(x_3, y_3) + 1/\lambda^2 f^2 h(x_3, y_3)^* h^*(-x_3, -y_3)^* g(x_3, y_3) + r_o \lambda f h(x_3, y_3)^* g(x_3, y_3) \delta(x_3, y_3 + \alpha \lambda f) + r_o \lambda f h^*(-x_3, -y_3)^* g(x_3, y_3) \delta(x_3, y_3 - \alpha \lambda f).$$

$$\begin{aligned} U_3(x_3, y_3) &\propto r_o^2 g(x_3, y_3) + \frac{1}{\lambda^2 f^2} [h(x_3, y_3)^* h^*(-x_3, -y_3)^* g(x_3, y_3)] \\ &\quad + \frac{r_o}{\lambda f} [h(x_3, y_3)^* g(x_3, y_3)^* \delta(x_3, y_3 + \alpha \lambda f)] \\ &\quad + \frac{r_o}{\lambda f} [h^*(-x_3, -y_3)^* g(x_3, y_3)^* \delta(x_3, y_3 - \alpha \lambda f)]. \end{aligned}$$

(10-12)

The third and fourth terms of this expression are of particular interest. Noting that

$$h(x_3, y_3)^* g(x_3, y_3)^* \delta(x_3, y_3 + \alpha \lambda f) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x_3 - \xi, y_3 + \alpha \lambda f - \eta) g(\xi, \eta) d\xi d\eta,$$

$$\begin{aligned} h(x_3, y_3)^* g(x_3, y_3)^* \delta(x_3, y_3 + \alpha \lambda f) \\ = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x_3 - \xi, y_3 + \alpha \lambda f - \eta) g(\xi, \eta) d\xi d\eta, \end{aligned}$$

(10-13)

we see that the third output term yields a *convolution* of h and g , centered at coordinates $(0, -\alpha \lambda f)$ in the (x_3, y_3) plane. Similarly, the fourth term may be rewritten as

$$h^*(-x_3, -y_3)^* g(x_3, y_3)^* \delta(x_3, y_3 - \alpha \lambda f) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(\xi, \eta) h^*(\xi - x_3, \eta - y_3 + \alpha \lambda f) d\xi d\eta,$$

$$\begin{aligned} h^*(-x_3, -y_3)^* g(x_3, y_3)^* \delta(x_3, y_3 - \alpha \lambda f) \\ = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(\xi, \eta) h^*(\xi - x_3, \eta - y_3 + \alpha \lambda f) d\xi d\eta, \end{aligned}$$

(10-14)

which is the *crosscorrelation* of g and h , centered at coordinates $(0, \alpha \lambda f)$ in the (x_3, y_3) plane.

Note that the first and second terms of (10-12), which are of no particular utility in the usual filtering operations, are centered at the origin of the (x_3, y_3) plane. Thus it is clear that if the “carrier frequency” $\alpha\alpha$ is chosen sufficiently high, or equivalently if the reference wave is introduced at a sufficiently steep angle, the convolution and crosscorrelation terms will be deflected (in opposite directions) sufficiently far off-axis to be viewed independently. To find the convolution of $h h$ and $g g$, the observer simply examines the distribution of light centered about the coordinates $(0, -\alpha\lambda f)$. To find the crosscorrelation of $h h$ and $g g$, the observation is centered at coordinates $(0, \alpha\lambda f)$.

To illustrate the requirements placed on $\alpha\alpha$ more precisely, consider the widths of the various output terms illustrated in Fig. 10.11. If the maximum width of $h h$ in the y^y direction is W_h and that of $g g$ is W_g , then the widths of the various output terms are as follows:

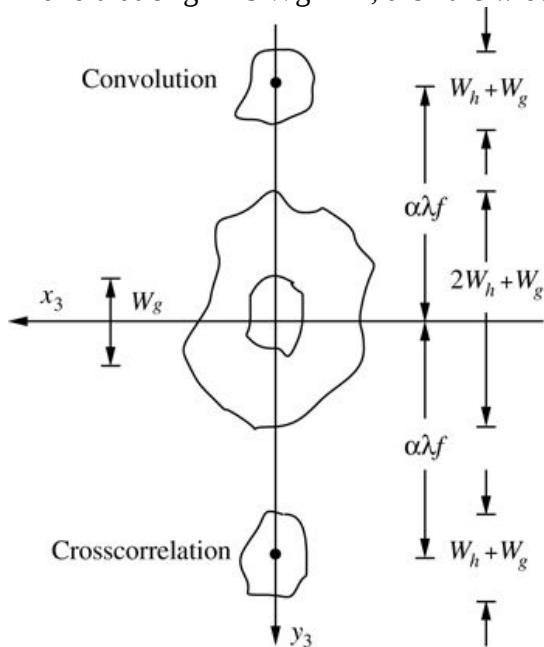


Figure 10.11
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 10.11 Locations of the various terms of the processor output.

The graph has the horizontal axis labeled x_3 in the negative axis and the horizontal axis labeled y_3 in the negative axis. Four planes are shown with their center on the vertical axis. The first plane is at the top with a dot at the center. The plane is labeled convolution. The length of the plane is marked as $W_h + W_g$. Another plane is centered at the origin. The length of the plane is W_g . A big plane surrounding the small one is shown also centered at the origin. The length of the plane is marked as $2W_h + W_g$. Another small plane is shown at the bottom of the vertical axis with a dot at its center. The plane is labeled “Cross-correlation.” The length of the plane is $W_h + W_g$. The distance between the center of the first plane to the center of the second plane is $\alpha\lambda f$. The distance between the center of the fourth plane to the center of the second plane is $\alpha\lambda f$.

1. $r_o^2 g(x_3, y_3) \rightarrow W_g$
2. $1/\lambda^2 f^2 h(x_3, y_3) * h^*(-x_3, -y_3) * g(x_3, y_3) \rightarrow 2W_h + W_g$
- $\frac{1}{\lambda^2 f^2} [h(x_3, y_3) * h^*(-x_3, -y_3) * g(x_3, y_3)] \rightarrow 2W_h + W_g$
3. $r_o \lambda f h(x_3, y_3) * g(x_3, y_3) * \delta(x_3, y_3 + \alpha \lambda f) \rightarrow Wh + W_g$
- $\frac{r_o}{\lambda f} [h(x_3, y_3) * g(x_3, y_3) * \delta(x_3, y_3 + \alpha \lambda f)] \rightarrow W_h + W_g$
4. $r_o \lambda f h^*(-x_3, -y_3) * g(x_3, y_3) * \delta(x_3, y_3 - \alpha \lambda f) \rightarrow Wh + W_g$
- $\frac{r_o}{\lambda f} [h^*(-x_3, -y_3) * g(x_3, y_3) * \delta(x_3, y_3 - \alpha \lambda f)] \rightarrow W_h + W_g$

From the figure it is clear that complete separation will be achieved if

$$\alpha > 1/\lambda f^2 (W_h + W_g),$$

$$\alpha > \frac{1}{\lambda f} \left(\frac{3W_h}{2} + W_g \right),$$

or equivalently, if

$$\theta > 32W_h f + W_g f,$$

$$\theta > \frac{3W_h}{2f} + \frac{W_g}{f},$$

(10-15)

where the small-angle approximation $\sin \theta \approx \theta$ has been used.

10.3.3 Advantages of the VanderLugt Filter

The use of a VanderLugt filter removes the two most serious limitations to conventional coherent optical processors. First, when a specified impulse response is desired, the task of finding the associated transfer function is eliminated; the impulse response is Fourier transformed *optically* by the system that synthesizes the frequency-plane mask. Second, the generally complicated complex-valued transfer function is synthesized with a single *absorbing* mask; the phase transmittance through the frequency plane need no longer be controlled in a complicated manner. The absorbing mask is simply immersed in a liquid gate to eliminate all relative phase shifts.

The VanderLugt filter remains very sensitive to the exact position of the frequency-plane mask, but no more sensitive than the conventional coherent processor. The recording of the modulated high-frequency carrier requires a higher-resolution film than might otherwise be used to synthesize the mask, but films with adequate resolution are readily available, and this requirement poses no particular problem in principle.

Note that the VanderLugt technique offers an important new flexibility to coherent processing. Whereas previously the realization of the frequency-plane mask was the major practical problem, the difficulties are now transferred back to the *space domain*. The difficulties are in general much less severe in the space domain, for the impulse responses required are often

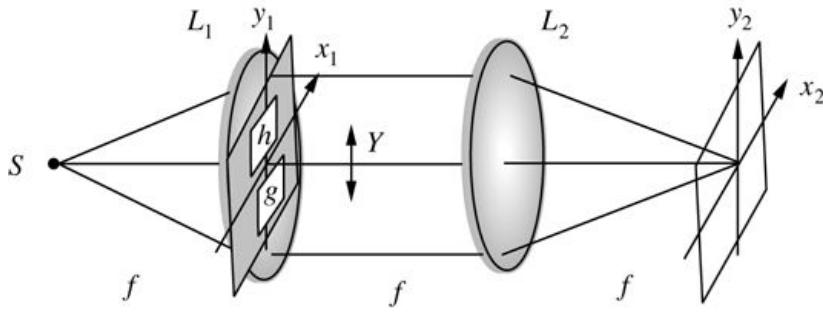
simple, and the necessary masks can be constructed by conventional photographic techniques. Thus the VanderLugt filter extends the use of coherent processors to an otherwise unattainable realm of operations. Many of the most promising applications fall in this realm.

10.4 The Joint Transform Correlator

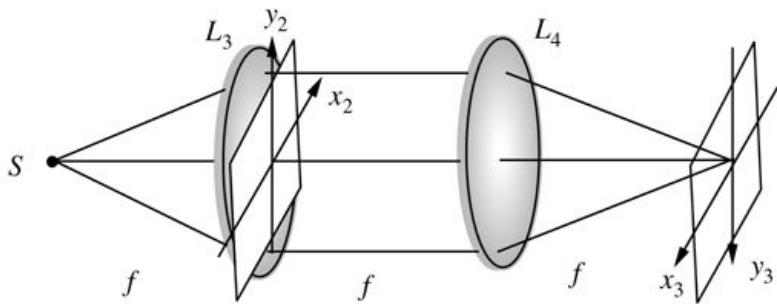
Before considering applications of coherent optical processing, an alternative method for performing complex filtering using a spatial carrier for encoding amplitude and phase information is considered. This method is due to [Weaver and Goodman \[360\]](#), and has become known as the *joint transform correlator*, although like the VanderLugt filter, it is equally capable of performing convolutions and correlations.

This type of filter differs from the VanderLugt filter in that *both* the desired impulse response *and* the data to be filtered are presented simultaneously during the recording process, rather than just presenting the desired impulse response. The transparency so constructed is then illuminated with a simple plane wave or spherical wave to obtain the filter outputs.

Consider the recording in [Fig. 10.12\(a\)](#). Lens L_1 collimates the light from the point source S . This collimated light then illuminates a pair of transparencies residing in the same plane, designated in the figure by their amplitude transmittances, h for the desired impulse response and g for the data to be filtered. For simplicity this input plane is taken to be the front focal plane of the Fourier transforming lens L_2 , but in fact this distance is arbitrary (vignetting will be eliminated if the inputs are placed in contact with lens, rather than in front of it). The Fourier transform of the composite input appears in the rear focal plane of L_2 , where the incident intensity is detected by either a photographic medium or a photosensitive spatial light modulator.



(a)



(b)

Figure 10.12Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company**Figure 10.12** The joint transform correlator. (a) Recording the filter, (b) obtaining the filtered output.

Illustration a shows a point S at the extreme left. Two lines from S point toward the upper and lower ends of Lens L1. The distance between the point S and L1 is marked as f . A plane and Lens L1 are adjoining. The vertical axis of the plane is y_1 and the horizontal axis is x_1 . Two small sloping parallelograms are shown, one in the upper half of the plane, labeled h , and another in the lower half of the plane, labeled g . Lens L2 is placed at a distance f from plane 1. A line connects the upper center of the plane with the upper portion of Lens L2. A line connects the lower center of the plane with the bottom portion of Lens L2. The second plane is placed at a distance f from Lens L2. The vertical axis of the plane is marked y_2 and the horizontal axis is marked x_2 . Two lines from the top and bottom portions of L2 point towards the center of the second plane. A horizontal line runs from the point S up to the center of second plane. A bidirectional vertical arrow is shown on the horizontal line between the first plane and Lens L2 and is marked Y .

Illustration b shows a point S at the extreme left. Two lines from S point toward the upper and lower ends of Lens L3. The distance between the point S and L3 is f . A plane and Lens L3 are adjoining. The vertical axis of the plane is y_2 and the horizontal axis is x_2 . Lens L4 is placed at a distance f from plane 1. A line connects the upper center of the plane with the upper portion of Lens L2. A line connects the lower center of the plane with the bottom portion of Lens L2. The second plane is placed at a distance of f from Lens L2. The vertical axis of the plane is marked y_3 in downward direction and the horizontal axis points toward the left and is marked X_3 . Two lines from the top and bottom portions of L2 point towards the center of second plane. A horizontal line runs from the point S up to the center of the second plane.

The field transmitted through the front focal plane is given by

$$U_1(x_1, y_1) = h(x_1, y_1 - Y/2) + g(x_1, y_1 + Y/2)$$

$$U_1(x_1, y_1) = h(x_1, y_1 - Y/2) + g(x_1, y_1 + Y/2)$$

where the separation between the centers of the two inputs is Y . In the rear focal plane of the lens we find the Fourier transform of this field,

$$U_2(x_2, y_2) = 1/\lambda f Hx_2 \lambda f, y_2 \lambda f e^{-j\pi y_2 Y/\lambda f} + 1/\lambda f Gx_2 \lambda f, y_2 \lambda f e^{+j\pi y_2 Y/\lambda f}.$$

$$U_2(x_2, y_2) = \frac{1}{\lambda f} H\left(\frac{x_2}{\lambda f}, \frac{y_2}{\lambda f}\right) e^{-j\pi y_2 Y/\lambda f} + \frac{1}{\lambda f} G\left(\frac{x_2}{\lambda f}, \frac{y_2}{\lambda f}\right) e^{+j\pi y_2 Y/\lambda f}.$$

Taking the squared magnitude of this field, the intensity incident on the recording plane is found to be

$$\mathcal{J}(x_2, y_2) = 1/\lambda^2 f^2 Hx_2 \lambda f, y_2 \lambda f^2 + Gx_2 \lambda f, y_2 \lambda f^2 + Hx_2 \lambda f, y_2 \lambda f G^* x_2 \lambda f, y_2 \lambda f e^{-j2\pi y_2 Y/\lambda f} + H^* x_2 \lambda f, y_2 \lambda f Gx_2 \lambda f, y_2 \lambda f e^{+j2\pi y_2 Y/\lambda f}.$$

$$\begin{aligned} \mathcal{J}(x_2, y_2) &= \frac{1}{\lambda^2 f^2} \left[\left| H\left(\frac{x_2}{\lambda f}, \frac{y_2}{\lambda f}\right) \right|^2 + \left| G\left(\frac{x_2}{\lambda f}, \frac{y_2}{\lambda f}\right) \right|^2 \right. \\ &\quad + H\left(\frac{x_2}{\lambda f}, \frac{y_2}{\lambda f}\right) G^*\left(\frac{x_2}{\lambda f}, \frac{y_2}{\lambda f}\right) e^{-j2\pi y_2 Y/\lambda f} \\ &\quad \left. + H^*\left(\frac{x_2}{\lambda f}, \frac{y_2}{\lambda f}\right) G\left(\frac{x_2}{\lambda f}, \frac{y_2}{\lambda f}\right) e^{+j2\pi y_2 Y/\lambda f} \right]. \end{aligned}$$

(10-16)

The transparency that results from this recording is assumed to have an amplitude transmittance that is proportional to the intensity that exposed it. After processing, this transparency is illuminated by collimated light and the transmitted field is Fourier transformed by a positive lens L_4 , assumed to have the same focal length f as the lens used in the recording process (see Fig. 10.12(b)). The field in the front focal plane of this final Fourier transforming lens L_4 consists of four terms, each of which is proportional to one of the terms in (10-16). Taking account of scaling factors and coordinate inversions, the field in the rear focal plane of L_4 is

$$U_3(x_3, y_3) = 1/\lambda f h(x_3, y_3) * h^*(-x_3, -y_3) + g(x_3, y_3) * g^*(-x_3, -y_3) + h(x_3, y_3) * g^*(-x_3, -y_3) * \delta(x_3, y_3 - Y) + h^*(-x_3, -y_3) * g(x_3, y_3) * \delta(x_3, y_3 + Y).$$

$$\begin{aligned} U_3(x_3, y_3) &= \frac{1}{\lambda f} [h(x_3, y_3) * h^*(-x_3, -y_3) + g(x_3, y_3) * g^*(-x_3, -y_3) \\ &\quad + h(x_3, y_3) * g^*(-x_3, -y_3) * \delta(x_3, y_3 - Y) \\ &\quad + h^*(-x_3, -y_3) * g(x_3, y_3) * \delta(x_3, y_3 + Y)]. \end{aligned}$$

(10-17)

Again it is the third and fourth terms of the expression for the output that are of most interest. We can rewrite them as

$$h(x_3, y_3) * g^*(-x_3, -y_3) * \delta(x_3, y_3 - Y) = \int_{-\infty}^{\infty} h(\xi, \eta) g^*(\xi - x_3, \eta - y_3 + Y) d\xi d\eta$$

$$h(x_3, y_3) \int_{-\infty}^{\infty} g(\xi, \eta) h^*(\xi - x_3, \eta - y_3 + Y) d\xi d\eta$$

(10-18)

and

$$h^*(-x_3, -y_3) * g(x_3, y_3) * \delta(x_3, y_3 + Y) = \int_{-\infty}^{\infty} g(\xi, \eta) h^*(\xi - x_3, \eta - y_3 - Y) d\xi d\eta.$$

$$h^*(-x_3, \int_{-\infty}^{\infty} y) * g(x_3, y_3) * \delta(x_3, y_3 + Y) d\xi d\eta.$$

(10-19)

Both of these expressions are crosscorrelations of the functions g^g and h^h . One output is centered at coordinates $(0, -Y)$ and the other at coordinates $(0, Y)$. The second output is a mirror reflection of the first about the optical axis.

To obtain a *convolution* of the functions h^h and g^g , it is necessary that one of them (and only one) be introduced in the processor of Fig. 10.12(a) with a mirror reflection about its own origin.⁶ For example, if originally we introduced the function $h(x_1, y_1 - Y/2)$, this input should be changed to $h(-x_1, -y_1 + Y/2)$, which is again centered at $Y/2$ but now is reflected about its own origin. The result will be two output terms, centered at $(0, Y)$ and $(0, -Y)$ in the output plane, each of which is a convolution of g^g and h^h . One term is identical with the other, but reflected about the optical axis.

Separation of the correlation (or convolution) terms from the uninteresting on-axis terms requires adequate separation of the two inputs at the start. If W_h represents the width of h^h and W_g is the width of g^g , both measured in the y direction, then separation of the desired terms can be shown to occur if

$$Y > \max(W_h, W_g) + W_g/2 + W_h/2,$$

$$Y > \max\{W_h, W_g\} + \frac{W_g}{2} + \frac{W_h}{2},$$

(10-20)

as is to be shown in Prob. 10-13.

The joint transform correlator is in some cases more convenient than the VanderLugt geometry, although both are widely used. Precise alignment of the filter transparency is required for the VanderLugt geometry, while no such alignment is necessary for the joint transform correlator. In addition, the joint transform approach has been found advantageous for real-time systems, i.e. systems that are required to rapidly change the filter impulse response. The price paid

for the joint transform geometry is generally a reduction of the space-bandwidth product of the input transducer that can be devoted to the data to be filtered, since a portion of that space-bandwidth product must be assigned to the filter impulse response. See [234] for further comparison of the two approaches.

10.5 Application to Character Recognition

A particular application of optical information processing that has been of interest for many years is found in the field of *character recognition*. As we shall see, this application affords an excellent example of desired processing operations with simple impulse responses but not necessarily simple transfer functions. The carrier-frequency filter synthesis methods are therefore particularly well suited for this application.

10.5.1 The Matched Filter

The concept of the *matched filter* plays an important role in pattern recognition problems. By way of definition, a linear space-invariant filter is said to be *matched* to a particular signal $s(x,y)$ if its impulse response $h(x,y)$ is given by

$$h(x,y) = s^*(-x,-y).$$

$$h(x, y) = s^* (-x, -y).$$

(10-21)

If an input $g(x,y)$ is applied to a filter matched to $s(x,y)$, then the output $v(x,y)$ is found to be

$$v(x,y) = \iint_{-\infty}^{\infty} h(x-\xi, y-\eta) g(\xi, \eta) d\xi d\eta = \iint_{-\infty}^{\infty} g(\xi, \eta) s^*(\xi-x, \eta-y) d\xi d\eta,$$

$$\begin{aligned} v(x,y) &= \int_{-\infty}^{\infty} h(x-\xi, y-\eta) g(\xi, \eta) d\xi d\eta \\ &= \int_{-\infty}^{\infty} g(\xi, \eta) s^*(\xi-x, \eta-y) d\xi d\eta, \end{aligned}$$

(10-22)

which is recognized to be the crosscorrelation function of g and s .

Historically the concept of the matched filter first arose in the field of signal detection; if a signal of known form, buried in “white” noise, is to be detected, then a matched filter provides the linear operation which maximizes the ratio of instantaneous signal power (at a particular time) to average noise power [350]. However, in the present application, the input patterns or characters will be assumed noiseless, and the use of a particular filtering operation must be justified on other grounds.

Considerable insight into the matched filtering operation is provided by an optical interpretation, as illustrated in [Fig. 10.13](#). Suppose that a filter, matched to the input signal $s(x,y)$, is to be synthesized by means of a frequency-plane mask in the usual coherent processing geometry. Fourier transformation of the impulse response (10-21) shows that the required transfer function is

$$H(fX, fY) = S^*(fX, fY),$$

$$H(f_X, f_Y) = S^*(f_X, f_Y),$$

(10-23)

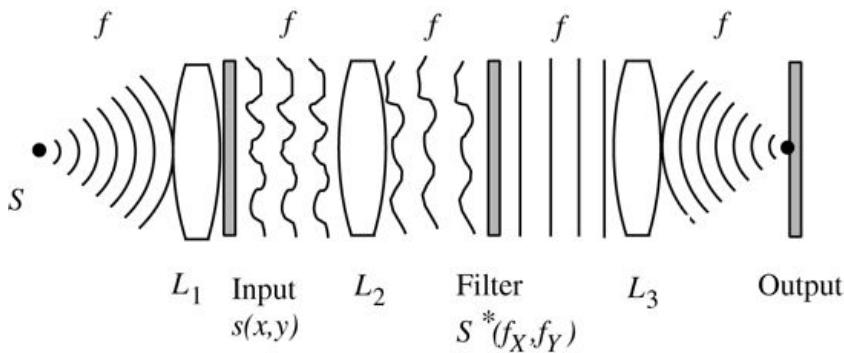


Figure 10.13

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 10.13 Optical interpretation of the matched-filtering operation.

The illustration shows a dark spot S on the left extreme from which waves move rightward to the convex curvature of Lens L_1 . Next to L_1 is a vertical thin rectangular strip labeled Input $s(x, y)$ followed by three irregularly curved vertical lines. Lens L_2 is next followed by three irregularly curved vertical lines. Next to the lines is a thin vertical rectangular strip. The strip is labeled “Filter S asterisk (f_X, f_Y).” Next there are four straight vertical lines followed by Lens L_3 . A vertical rectangular bar is shown at the extreme right and is marked “output.” At the center of the bar is a dark spot from which waves move leftward to L_3 . The distances between S and L_1 , between input and L_2 , between filter and L_3 , and between L_3 and output are each f .

where $H = \mathcal{F}_h H = \mathcal{F}\{h\}$ and $S = \mathcal{F}_s S = \mathcal{F}\{s\}$. Thus the frequency plane filter should have an amplitude transmittance proportional to $S^* S^*$.

Consider now the particular nature of the field distribution transmitted by the mask when the signal s^S (to which the filter is matched) is present at the input. Incident on the filter is a field distribution proportional to S^S , and transmitted by the filter is a field distribution proportional to $S S^* S S^*$. This latter quantity is entirely *real*, which implies that the frequency-plane filter exactly cancels all the curvature of the incident wavefront S^S . Thus the transmitted field consists of a *plane wave* (generally of nonuniform intensity), which is brought to a bright focus by the final transforming lens. When an input signal other than $s(x, y)$ is present, the wavefront curvature will in general *not* be canceled by the frequency-plane filter, and the transmitted light will *not* be brought to a bright focus by the final lens. Thus the presence of the signal s^S can conceivably be detected by measuring the intensity of the light at the focal point of the final transforming lens.

If the input s^S is not centered on the origin, the bright point in the output plane simply shifts by a distance equal to the misregistration distance, a consequence of the space invariance of the matched filter (cf. [Prob. 10-12](#)).

10.5.2 A Character-Recognition Problem

Consider the following character-recognition problem: The input g to a processing system may consist of any one of N possible alphanumeric characters, represented by s_1, s_2, \dots, s_N , and the particular character present is to be determined by the processor. As will now be demonstrated, the identification process can be realized by applying the input to a bank of N filters, each matched to one of the possible input characters.

A block diagram of the recognition machine is shown in Fig. 10.14. The input is simultaneously (or sequentially) applied to the N matched filters with transfer functions $S_1^*, S_2^*, \dots, S_N^*$. The response of each filter is normalized by the square root of the total energy in the character to which it is matched. This normalization, which can be accomplished electronically after detection of the filter outputs, takes account of the fact that the various input characters will generally not be of equal energy. Finally, the squared moduli of the outputs $|v_1|^2, |v_2|^2, \dots, |v_N|^2$ are compared at the particular points where their maximum outputs would be anticipated (assuming that the character to which they are matched is present in each case). As will now be demonstrated, if the particular character

$$g(x,y) = sk(x,y)$$

$$g(x, y) = s_k(x, y)$$

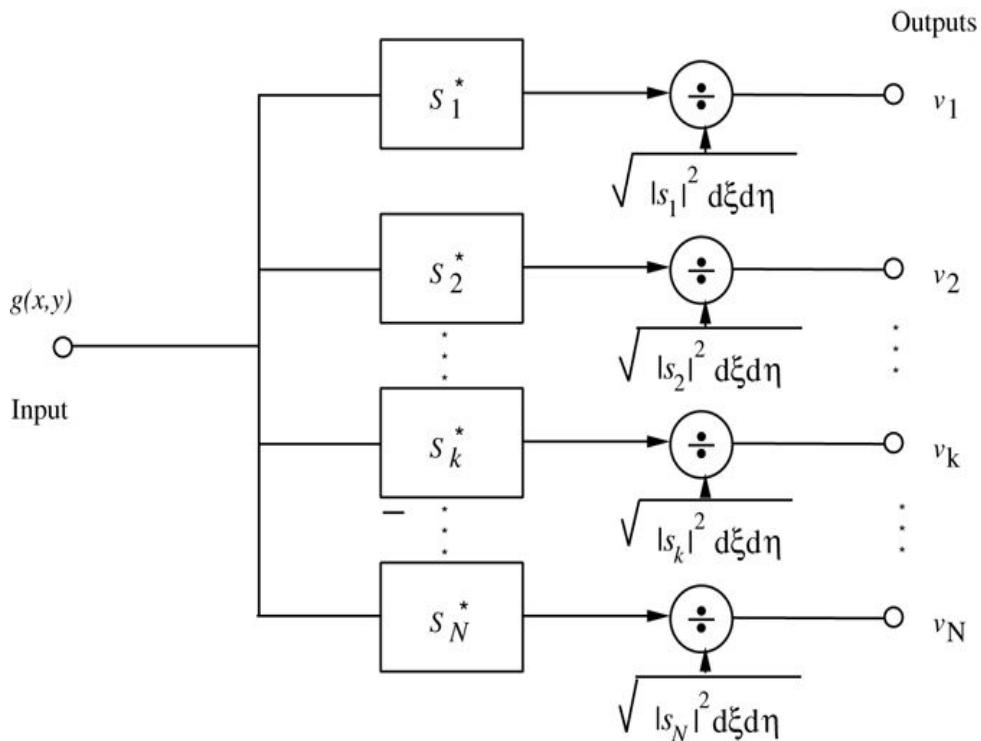


Figure 10.14

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 10.14 Block diagram of a character-recognition system.

The illustration starts with a node labeled Input $g(x, y)$. The root splits into four. The first node leads to a box labeled S_1^* . The node further leads to a circle with a division symbol inside

and then ends at a small circular structure labeled V1. Right below the division symbol, the following formula is given: Square root of (square of determinant of S1 dksi deta). An arrow from the formula points toward the division symbol. The second node leads to a box labeled S2 asterisk. The node further leads to a circle with a division symbol inside and then ends at a small circular structure labeled V2. Right below the division symbol, the following formula is given: Square root of (square of determinant of S2 dksi deta). An arrow from the formula points toward the division symbol. The third node leads to a box labeled SK asterisk. The node further leads to a circle with a division symbol inside and then ends at a small circular structure labeled Vk. Right below the division symbol, the following formula is given: Square root of (square of determinant of Sk dksi deta). An arrow from the formula points toward the division symbol. The third node leads to a box labeled SN asterisk. The node further leads to a circle with a division symbol inside and then ends at a small circular structure labeled VN. Right below the division symbol, the following formula is given: Square root of (square of determinant of SN dksi deta). An arrow from the formula points toward the division symbol. A sequence of vertical dots is shown between the box, S2 asterisk, and Sk asterisk and another sequence of vertical dots are shown between the box, Sk asterisk, and SN asterisk. There is a sequence of vertical dots between V2 and Vk and another between Vk and Vn.

is actually present at the input, then the particular output $|v_k|^2$ will be the largest of the N responses.

To prove this assertion, first note that, from (10-22), the peak output $|v_k|^2$ of the correct matched filter is given by

$$|v_k|^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |s_k|^2 d\xi d\eta = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |s_k|^2 d\xi d\eta. \quad (10-24)$$

$$\left| v_k \right|^2 = \frac{1}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |s_k|^2 d\xi d\eta} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |s_k|^2 d\xi d\eta.$$

(10-24)

On the other hand, the response $|v_n|^2 (n \neq k)$ of an incorrect matched filter is given by

$$|v_n|^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |s_k s_n^*|^2 d\xi d\eta = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |s_n|^2 d\xi d\eta.$$

$$\left| v_n \right|^2 = \frac{\left[\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} s_k s_n^* d\xi d\eta \right]^2}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |s_n|^2 d\xi d\eta}.$$

(10-25)

However, from Schwarz's inequality, we have

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |s_k s_n^*|^2 d\xi d\eta \leq \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |s_k|^2 d\xi d\eta \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |s_n|^2 d\xi d\eta.$$

$$\left| \int_{-\infty}^{\infty} \int s_k s_n^* d\xi d\eta \right|^2 \leq \int_{-\infty}^{\infty} \int |s_k|^2 d\xi d\eta \int_{-\infty}^{\infty} \int |s_n|^2 d\xi d\eta.$$

It follows directly that

$$|v_n|^2 \leq \int_{-\infty}^{\infty} \int |s_k|^2 d\xi d\eta = |v_k|^2,$$

$$|v_n|^2 \leq \int_{-\infty}^{\infty} \int |s_k|^2 d\xi d\eta = |v_k|^2,$$

(10-26)

with equality if and only if

$$s_n(x, y) = \kappa s_k(x, y).$$

$$s_n(x, y) = \kappa s_k(x, y).$$

From this result it is evident that the matched filter does provide *one* means of recognizing which character, of a set of possible characters, is actually being presented to the system. It should be emphasized that this capability is not unique to the matched filter. In fact it is often possible to modify (mismatch) all the filters in such a way that the discrimination between characters is improved. Examples of such modifications include: (1) overexposing the low-frequency portion of a VanderLugt filter transparency so as to suppress the influence of those frequencies in the decision process (see, for example, [344], pp. 130-133); (2) eliminating the amplitude portion of the transfer functions of the matched filters and retaining only phase information [176]; and (3) modifying the nonlinearity of the normally square-law detection process in the joint transform correlator to enhance discrimination between patterns [181], [182].

Not all pattern-recognition problems are of the type described above. For example, rather than trying to distinguish between several possible known patterns, we may wish simply to detect the presence or absence of a single known object in a larger image. Such a problem is closer to what the matched filter is known to do well, namely detect a known pattern in the presence of background noise, but has the added difficulty that the orientation and possibly the scale size of the target may not be under the same level of control that is present in the character recognition problem. We return in [Section 10.5.4](#) to discussing some of the difficulties of the matched filter approach to such problems.

10.5.3 Optical Synthesis of a Character-Recognition Machine

The matched filter operation can readily be synthesized by means of either the VanderLugt technique or the joint transform technique discussed earlier. Our discussion here is directed at the VanderLugt-type system, but the reader may wish to contemplate how the equivalent system could be realized with the joint transform geometry.

Recall that one of the outputs of the VanderLugt filtering operation is itself the crosscorrelation of the input pattern with the original pattern from which the filter was synthesized. By restricting attention to the proper region of the output space, the matched filter output is readily observed.

[Figure 10.15\(a\)](#) shows a photograph of the impulse response of a VanderLugt filter which has been synthesized for the character P. The upper portion of response will generate the convolution of the input data with the symbol P, while the lower response will generate the crosscorrelation of the input with the letter P. The central portion of the response is undesired and not of interest.

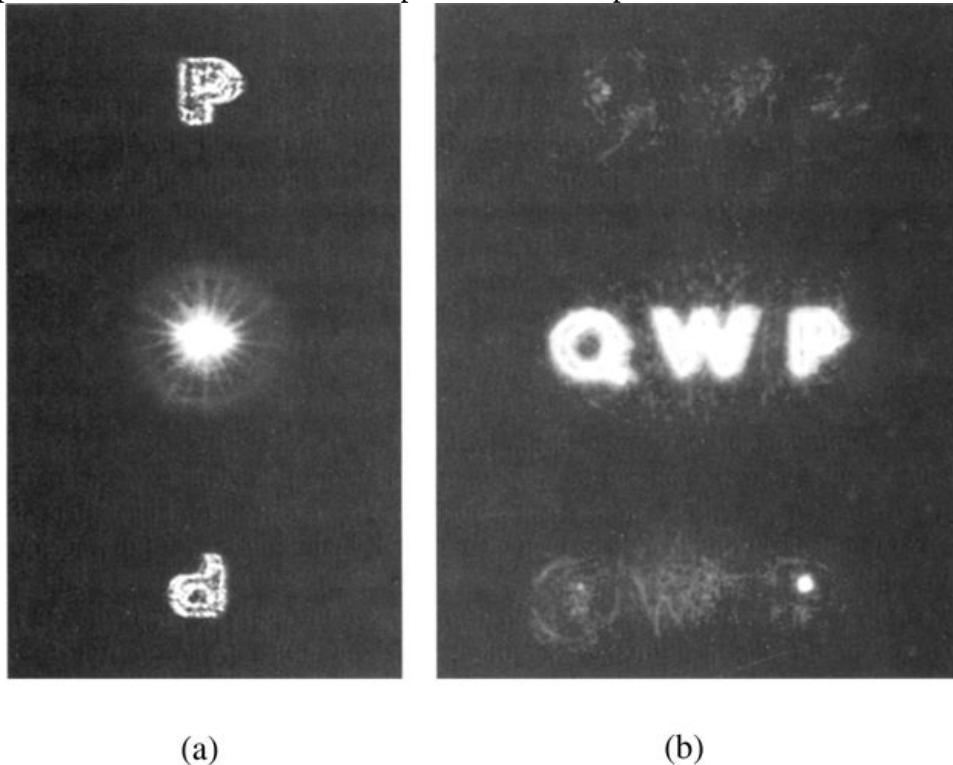


Figure 10.15

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

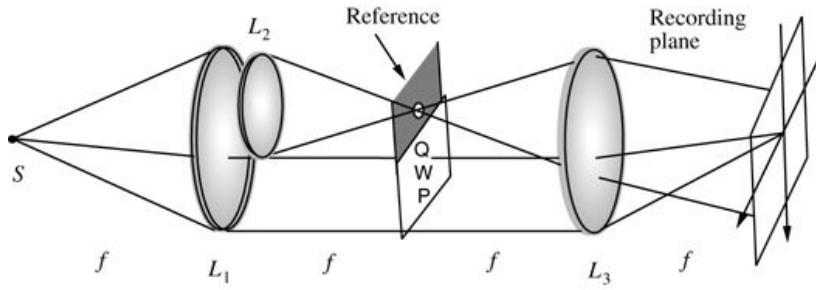
[Figure 10.15](#) Photographs of (a) the impulse response of a VanderLugt filter, and (b) the response of the filter to the letters Q, W, and P.

Image a shows a vertical rectangular patch of darkness with a bright circular spot at its center. The letter P is shown above the circular spot and below the circular spot, an inverted P is shown. Image b shows a vertical rectangular patch of darkness broader than image a. At its center are three letters Q, W, and P.

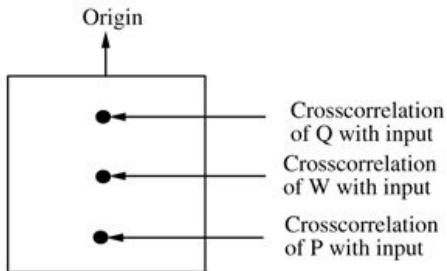
[Figure 10.15\(b\)](#) shows the response of the filter to the letters Q, W, and P. Note the presence of the bright point of light in the response to P, indicating the high correlation between the input letter and the letter to which it is matched.

To realize the entire *bank* of matched filters illustrated in [Fig. 10.14](#), it would be possible to synthesize N^N separate VanderLugt filters, applying the input to each filter sequentially.

Alternatively, if N^N is not too large, it is possible to synthesize the entire bank of filters on a single frequency-plane filter. This can be done by frequency-multiplexing, or recording the various frequency-plane filters with different carrier frequencies on a single transparency. [Figure 10.16\(a\)](#) illustrates one way of recording the multiplexed filter. The letters Q, W, and P are at different angles with respect to the reference point, and as a consequence, the crosscorrelations of Q, W, and P with the input character appear at different distances from the origin, as illustrated in [Fig. 10.16\(b\)](#).



(a)



(b)

Figure 10.16

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 10.16 Synthesis of a bank of matched filters with a single frequency-plane filter. (a) Recording the frequency-plane filter; (b) format of the matched filter portion of the output.

In illustration a, point S is in the extreme left. Lens L1 is placed at a distance f from S. Two lines emerging from S point toward the upper and lower ends of L1. Lens L2 is placed adjoining L1. L2 is half the size of L1 and is placed near the upper half of L1. A reference frame is placed at a distance f from L1. The upper half of the reference frame is shaded, with a hole at its center and three letters, Q, W, and P, stacked vertically in the top down direction. Lens L3 is placed at a distance f from the reference frame and a recording plane is placed at a distance f from L3. A line from the top of L2 passes through the hole of reference plane and points toward the lower end of L3 and a line from the bottom of L2 passes through the hole of reference plane and points toward the upper end of L3. A line from the bottom of L1 points toward the bottom of L3. A line from the upper end of L3 points at the upper center of the recording plane. A line from the lower end of L3 points toward the centre of the recording plane. A line connects a point slightly lower than the center of L3 and a point slightly above the bottom end of the recording plane. The vertical axis of L3 is in downward direction and horizontal axis of L3 points toward the left. A horizontal line starts from S and passes through the center, thereby dividing the system in two equal halves and ends at the center of recording plane.

Illustration b shows a square inside which three dots are stacked vertically at the center. The first dot is cross correlation of Q with input, the second dot is cross correlation of W with input, and the third dot is cross correlation of P with input. An arrow labeled origin emerges from the upper center of the square and points upward.

The number of different filters that can be realized by this technique is limited by the dynamic range that can be achieved in the frequency-plane filter. Synthesis of nine separate

impulse responses in a single mask was demonstrated by VanderLugt at an early date (see [344], pp. 133-139).

10.5.4 Sensitivity to Scale Size and Rotation

The coherent optical pattern-recognition technique described above suffers from certain deficiencies that are shared by all matched-filter approaches to the pattern recognition problem. Specifically, such filters are too sensitive to scale size changes and rotations of input patterns. When an input pattern is presented with an angular orientation or a scale size that is different from those of the pattern to which the filter is matched, the response of the correct matched filter is reduced, and errors arise in the pattern recognition process. The degree of sensitivity of a matched filter to rotation and scale-size depends to a large extent on the structure of the pattern to which it is matched. For example, a matched filter for the character L is obviously much more rotation-sensitive than that for the letter O. One solution that has been used is to make a bank of matched filters, each of which is matched to the pattern of interest with a different rotation and/or scale size. If any of these matched filters have a large output, then the pattern of interest is known to have been presented to the input. For other approaches to invariant coherent optical pattern recognition, see [11], [54] and [55].

10.6 Image Restoration

A common problem in image processing, and one that has been studied extensively in the context of optical information processing, is *image restoration*, by which we mean the restoration of an image that has been blurred by a known linear, invariant point-spread function. In this section we summarize some of the past work on this problem. The reason for doing so is only partly because of the extensive past work. Equally important, there are lessons that have been learned in this application, particularly about clever use of the properties of wavefront modulation devices, that can be applied to other unrelated problems.

10.6.1 The Inverse Filter

Let $o(x, y)$ represent the intensity distribution associated with an incoherent object, and let $i(x, y)$ represent the intensity distribution associated with a blurred image of that object. For simplicity we assume that the magnification of the imaging system is unity and we define the image coordinates in such a way as to remove any effects of image inversion.

We assume that the blur the image has been subjected to is a linear, space-invariant transformation, describable by a *known* space-invariant point-spread function $s(x, y)$. Thus, in the simplest description of the problem, the object and image are related by

$$i(x, y) = \iint_{-\infty}^{\infty} o(\xi, \eta) s(x - \xi, y - \eta) d\xi d\eta.$$

$$i(x, y) = \int_{-\infty}^{\infty} o(\xi, \eta) s(x - \xi, y - \eta) d\xi d\eta.$$

(10-27)

We seek to obtain an estimate $\hat{o}(x, y)$ of $o(x, y)$, based on the measured image intensity $i(x, y)$ and the known point-spread function $s(x, y)$. In other words, we wish to invert the blurring operation and recover the original object.

An unsophisticated solution to this problem is quite straightforward. Given the relationship between object and image in the frequency domain,

$$\mathcal{F}i(x, y) = \mathcal{F}s(x, y) * o(x, y) = S(f_X, f_Y) O(f_X, f_Y),$$

$$\mathcal{F}\{i(x, y)\} = \mathcal{F}\{s(x, y) * o(x, y)\} = S(f_X, f_Y) O(f_X, f_Y),$$

(10-28)

it seems obvious that the spectrum of the original object can be obtained by simply dividing the image spectrum by the known OTF of the imaging system,

$$\hat{o}(f_X, f_Y) = I(f_X, f_Y) S(f_X, f_Y).$$

$$\hat{o}(f_X, f_Y) = \frac{I(f_X, f_Y)}{S(f_X, f_Y)}.$$

(10-29)

An equivalent statement of this solution is that we should pass the detected image $i(x, y)$ through a linear space-invariant filter with transfer function

$$H(f_X, f_Y) = S(f_X, f_Y).$$

$$H(f_X, f_Y) = \frac{1}{S(f_X, f_Y)}.$$

(10-30)

Such a filter is commonly referred to as an “inverse filter,” for obvious reasons.

This straightforward solution has several serious defects:

1. Diffraction limits the set of frequencies over which the transfer function $S(f_X, f_Y)$ is nonzero to a finite range. Outside this range, $S=0$ and its inverse is ill defined. For this reason, it is necessary to limit the application of the inverse filter to those frequencies lying within the diffraction-limited passband.
2. Within the range of frequencies for which the diffraction-limited transfer function is nonzero, it is possible (indeed likely) that transfer function S will have isolated zeros. Such is the case for both a serious defocusing error and for many kinds of motion blur (see [Prob. 10-14](#)). The value of the restoration filter is undefined at the frequencies where these isolated zeros occur. Another way of stating this problem is that the restoration filter would need a transfer function with infinite dynamic range in order to properly compensate the spectrum of the image.
3. The inverse filter takes no account of the fact that there is inevitably noise present in the detected image, along with the desired signal. The inverse filter boosts the most those frequency components that have the worst signal-to-noise ratios, with the result that the recovered image is usually dominated by noise.

The only solution to the last of the problems raised above is to adopt a new approach to determining the desired restoration filter, an approach that includes the effects of noise. One such approach is described in the following, and it will be seen to solve the first two problems as well.

10.6.2 The Wiener Filter, or the Least-Mean-Square-Error Filter

A new model for the imaging process is now adopted, one that takes into account explicitly the presence of noise. The detected image is now represented by

$$i(x, y) = o(x, y) * s(x, y) + n(x, y),$$

$$i(x, y) = o(x, y) * s(x, y) + n(x, y),$$

(10-31)

where $n(x, y)$ is the noise associated with the detection process. In addition to the presence of the noise term, which must be regarded as a random process, we also treat the object $o(x, y)$ as a random process in this formulation (if we knew what the object is, we would have no need to form an image of it, so the object that is present is regarded as one realization of a random process). We assume that the power spectral densities⁷ (i.e. the distributions of average power over frequency) of the object and the noise are known, and are represented by $\Phi_o(f_X, f_Y)$ and $\Phi_n(f_X, f_Y)$. Finally, the goal is to produce a linear restoration filter that minimizes the mean-square difference between the true object $o(x, y)$ and the estimate of the object $\hat{o}(x, y)$, i.e. to minimize

$$\epsilon^2 = \text{Average}|o - \hat{o}|^2.$$

$$\epsilon^2 = \text{Average}[|o - \hat{o}|^2].$$

(10-32)

The derivation of the optimum filter would take us too far afield, so we content ourselves with presenting the result and referring the reader to another source [131]. The transfer function of the optimum restoration filter is given by

$$H(f_X, f_Y) = S^*(f_X, f_Y) |S(f_X, f_Y)|^2 + \Phi_n(f_X, f_Y) \Phi_o(f_X, f_Y).$$

$$H(f_X, f_Y) = \frac{S^*(f_X, f_Y)}{\left|S(f_X, f_Y)\right|^2 + \frac{\Phi_n(f_X, f_Y)}{\Phi_o(f_X, f_Y)}}.$$

(10-33)

This type of filter is often referred to as a *Wiener filter*, after its inventor, Norbert Wiener.

Note that at frequencies where the signal-to-noise ratio is high ($\Phi_n/\Phi_o \ll 1$), the optimum filter approximates an inverse filter,

$$H \approx S^* |S|^2 = 1/S,$$

$$H \approx \frac{S^*}{|S|^2} = \frac{1}{S},$$

while at frequencies where the signal-to-noise ratio is low ($\Phi_n/\Phi_o \gg 1$), it reduces to a strongly attenuating matched filter,

$$H \approx \Phi_o \Phi_n S^*.$$

$$H \approx \frac{\Phi_o}{\Phi_n} S^*.$$

[Figure 10.17](#) shows plots of the magnitude of the transfer function of the restoration filter under the assumption of a severe focusing error and white (i.e. flat) power spectra for the signal and the

noise. Several different signal-to-noise ratios are represented. Note that at high signal-to-noise ratio, the Wiener filter reduces the relative strength of the low frequencies and boosts the relative strength of the high frequencies. At low signal-to-noise ratio, all frequencies are reduced.

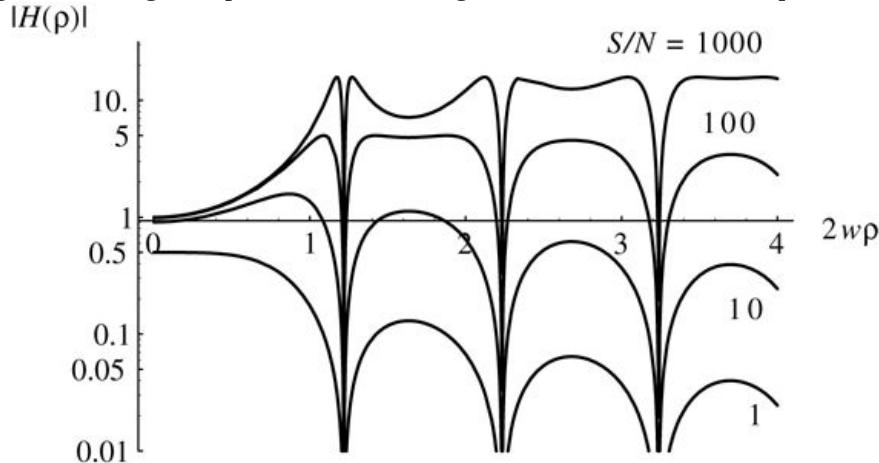


Figure 10.17

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 10.17 Magnitudes of the transfer function of a Wiener filter. The image is assumed to have been blurred by a point-spread function consisting of a circular disk of radius w . The signal-to-noise ratio is varied from 1000 to 1. The phase of the filter changes between 0 and π radians between alternate zeros of this transfer function.

The graph plots values of determinant of $H(\rho)$ along the vertical axis marked from 0.01 to 10. Values of $2 w \rho$ are plotted along the vertical axis marked from 0 to 4. Four curves are shown. The first curve starts at 0 on the vertical axis and runs parallel to the horizontal axis up to 0.4 after which it shows a steep decrease and reaches a value of 0.01 on the vertical axis which corresponds to 1.2 on horizontal axis. The curve shows a gradual increase and reaches 0.1 on vertical axis which corresponds to 1.4 on horizontal axis and again decreases to 0.01 on vertical axis which corresponds to 2.2 on horizontal axis. The curve shows a similar pattern and reaches 0.05 on vertical axis which corresponds to 2.4 on horizontal axis and decreases again to 0.01 on vertical axis which corresponds to 3.2 on horizontal axis. The curve again increases and reaches 0.04 on vertical axis which corresponds to 3.4 on horizontal axis and ends at 0.02 on vertical axis which corresponds to 4 on the horizontal axis. The first curve is labeled 1.

The second curve starts at 1 on the vertical axis and runs on horizontal axis up to 0.38 after which it shows an increase and reaches 1.2 on the vertical axis which corresponds to 0.9 on horizontal axis. The curve shows a steep decrease and reaches 0.01 on vertical axis which corresponds to 1.2 on horizontal axis and again increases and reaches 1.1 on vertical axis which corresponds to 1.4 on horizontal axis and again shows a decrease and reaches 0.01 on vertical axis which corresponds to 2.2 on horizontal axis. The curve shows a similar pattern and reaches the value of 0.06 on vertical axis which corresponds to 2.4 on horizontal axis and decreases again to the value 0.01 on vertical axis which corresponds to 3.2 on horizontal axis. The curve again increases and reaches a value of 0.4 on the vertical axis which corresponds to 3.4 on horizontal axis and ends at 0.2 on vertical axis which corresponds to 4 on the horizontal axis. The second curve is labeled 10.

The third curve starts at 1 on the vertical axis and runs on the horizontal axis up to 0.38 after which it shows an increase and reaches 4 on the vertical axis which corresponds to 1.1 on horizontal axis. The curve shows a steep decrease and reaches 0.01 on vertical axis which corresponds to 1.2 on horizontal axis and again increases steeply and reaches 4 on vertical axis which corresponds to 1.4 on horizontal axis and runs parallel to horizontal axis till the point 2 and

again shows a decrease and reaches 0.01 on vertical axis which corresponds to 2.2 on horizontal axis. The curve shows a similar pattern and reaches the value of 3 on vertical axis which corresponds to 2.4 on horizontal axis and shows a decrease and reaches the value 0.01 on vertical axis which corresponds to 3.2 on horizontal axis. The curve again increases and reaches a value of 2.75 on the vertical axis which corresponds to 3.4 on horizontal axis and ends at 2.5 on vertical axis which corresponds to 4 on the horizontal axis. The third curve is labeled 100.

The fourth curve starts at 1 on the vertical axis and runs on the horizontal axis up to 0.38 after which it shows an increase and reaches the point 11 on the vertical axis which corresponds to 1.18 on horizontal axis. The curve shows a steep decrease and reaches 0.01 on vertical axis which corresponds to 1.2 on horizontal axis and again increases steeply and reaches 11 on vertical axis which corresponds to 1.2 on horizontal axis and shows a dip reaching 6 on the vertical axis which corresponds to 1.6 on horizontal axis and again increases to point 11 on the vertical axis which corresponds to 2.2 after which it shows a steep decrease reaching 0.01 on vertical axis at 2.2 on horizontal axis. The curve shows a similar pattern and reaches the value of 11 on vertical axis which corresponds to 2.25 on horizontal axis and runs parallel to horizontal axis with a slight dip at 2.6 on horizontal axis and again shows a steep decrease and reaches 0.01 on vertical axis at 3.2 on horizontal axis. The curve again increases and reaches a value of 11 on the vertical axis which corresponds to 3.25 on horizontal axis and runs parallel to horizontal axis and ends at 4. The fourth curve is labeled S/N=1000. The values of the graphs mentioned above are approximate.

Note that at frequencies outside the diffraction-limited passband of the imaging system, no object information is present, and therefore the noise-to-signal ratio is infinite. Hence the Wiener filter makes no attempt to restore object frequency components that are simply not present in the image, a very sensible strategy.

10.6.3 Filter Realization

Many methods exist for optically realizing inverse and Wiener restoration filters. We discuss only two such methods, one relatively obvious, the other not at all obvious. Both depend on the use of VanderLugt-type filters. In both cases we suppose that the known impulse response $s(x, y)$ of the blurred system has been recorded, and therefore is known. This recording could have been obtained by imaging a point source through the blurred system, or could have been generated by computer. We assume that the filter transparencies are made using photographic film or plate, but they could in fact be realized with any spatial light modulator if operated properly to achieve the desired amplitude transmittances.

Inverse Filter

The first method is one that attempts to realize an inverse filter [333]. Using the recording of the blur, we record two transparencies which will be sandwiched (i.e. placed in close contact) to form the frequency plane filter. Referring back to Fig. 10.16(a), one component of the filter is of the VanderLugt type, recorded interferometrically as shown, but with an input that consists only of the known blur function s . This filter captures both the amplitude and phase associated with the transfer function of the blur, S . A second transparency is recorded in the same geometry, but with the reference point source blocked, thus capturing information only about the intensity $|S|^2$.

The transmittance of the VanderLugt filter consists of four terms, as before, and only one of these is of interest in this problem. We again focus on the term proportional to $S^* S^*$, the same

term that was of interest in the case of the matched filter. With exposure in the linear region of the t_A versus E curve and with proper processing, this component of amplitude transmittance can be written

$$t_{A1} \propto S^*(f_X, f_Y).$$

$$t_{A1} \propto S^*(f_X, f_Y).$$

The second transparency is exposed in the linear region of the H&D curve and processed with a photographic γ equal to 2. The result is an amplitude transmittance

$$t_{A2} \propto 1|S(f_X, f_Y)|^2.$$

$$t_{A2} \propto \frac{1}{|S(f_X, f_Y)|^2}.$$

When these two transparencies are placed in close contact, the amplitude transmittance of the pair is

$$t_A = t_{A1} t_{A2} = S^*(f_X, f_Y) |S(f_X, f_Y)|^2 = 1|S(f_X, f_Y)|,$$

$$t_A = t_{A1} t_{A2} = \frac{S^*(f_X, f_Y)}{|S(f_X, f_Y)|^2} = \frac{1}{|S(f_X, f_Y)|},$$

which is the transfer function of an inverse filter.

In addition to all the difficulties associated with an inverse filter that were mentioned earlier, this method suffers from other problems related to the photographic medium or spatial light modulator. The dynamic range of amplitude transmittance over which this filter can function properly is quite limited. The problem is evident if we consider only the second filter, which was recorded in the linear region of the H&D curve. If we wish this filter to behave as desired over a 10:1 dynamic range of $|S|^{1/2}$, this requires proper behavior over a 100:1 range of $1/|S|^2$. But since the amplitude transmittance of this filter is proportional to $1/|S|^2$, the intensity transmittance is proportional to $1/|S|^4$, and a 10:1 change of S implies a 10,000:1 change of intensity transmittance. To properly control this filter over the range of interest would require controlling the density accurately over a range of 0 to 4. Densities as high as 4 can seldom be achieved in practice, and even a density of 3 requires some special effort. For this reason, the dynamic range of $|S|^{1/2}$ over which the filter functions properly is severely limited in practice.

Wiener Filter

A superior approach to realizing an image restoration filter is one that generates a Wiener filter, and does so with considerably more dynamic range than the previous method afforded. Such a method was introduced by [Ragnarsson \[292\]](#). There are several novel aspects to this approach to filter realization:

1. Diffraction, rather than absorption, is used to attenuate frequency components.

2. Only a single interferometrically generated filter is required, albeit one with an unusual set of recording parameters.
3. The filter is bleached and therefore introduces only phase shifts in the transmitted light.

Certain postulates underlie this method of recording a filter. First, it is assumed that the maximum phase shift introduced by the filter is much smaller than 2π radians, and therefore

$$tA = ej\phi \approx 1 + j\phi.$$

$$t_A = e^{j\phi} \approx 1 + j\phi.$$

In addition, it is assumed that the phase shift of the transparency after bleaching is linearly proportional to the silver density present before bleaching,

$$\phi \propto D.$$

$$\phi \propto D.$$

This assumption is true to a very good approximation if a nontanning bleach is used, for such a bleach returns metallic silver to a transparent silver salt, and the density of that transparent material determines the phase shift introduced by the bleached transparency. Finally, it is assumed that the filter is exposed and processed such that operation is in the linear part of the H&D curve, where density is linearly proportional to the logarithm of exposure, i.e. where

$$D = \gamma \log E - D_o.$$

$$D = \gamma \log E - D_o.$$

Note that this is not the usual region of operation used for other interferometrically generated filters, which are typically recorded in the linear portion of the t_A versus E curve.

The three postulates above lead to certain conclusions regarding the mathematical relationship between changes of exposure and resulting changes of amplitude transmittance. To discover this relationship, first note that a change of logarithmic exposure induces a proportional change of amplitude transmittance, as evidenced by the chain

$$\Delta tA \propto \Delta\phi \propto \Delta D \propto \Delta(\log E),$$

$$\Delta t_A \propto \Delta\phi \propto \Delta D \propto \Delta(\log E),$$

which is implied by the above hypotheses. In addition, if the exposure pattern consists of a strong average exposure E^- and a weaker varying exposure ΔE , then

$$\Delta(\log E) \approx \Delta E E^-,$$

$$\Delta(\log E) \approx \frac{\Delta E}{E},$$

making

$$\Delta tA \propto \Delta E E^-.$$

$$\Delta t_A \propto \frac{\Delta E}{E}.$$

(10-34)

With the above information as background, attention is turned to the process of recording the deblurring filter. The recording geometry is that of a VanderLugt filter, exactly as illustrated previously in [Figs. 10.9 or 10.10](#), but with only the function $s(x,y)$ present in the input transparency. The exposure produced by this interferometric recording is

$$E(x,y) = TA^2 + a^2 S(x,y) \cos[2\pi\alpha x + \phi(x,y)],$$

$$E(x,y) = T \left\{ A^2 + a^2 \left| S \left(\frac{x}{\lambda f}, \frac{y}{\lambda f} \right) \right|^2 + 2Aa \left| S \left(\frac{x}{\lambda f}, \frac{y}{\lambda f} \right) \right| \cos [2\pi\alpha x + \phi \left(\frac{x}{\lambda f}, \frac{y}{\lambda f} \right)] \right\},$$

(10-35)

where A is the square root of the intensity of the reference wave at the film plane, a is the square root of the intensity of the object wave at the origin of the film plane, α is again the carrier frequency introduced by the off-axis reference wave, ϕ is the phase distribution associated with the blur transfer function S , and T is the exposure time.

An additional unusual attribute of the Ragnarsson filter is the fact that it is recorded with the object wave much stronger at the origin of the film plane than the reference wave, i.e.

$$a^2 \gg A^2.$$

$$a^2 \gg A^2.$$

Because of this condition, we make the following associations with the average exposure E^- and the varying component of exposure ΔE ,

$$E^- = A^2 + a^2 S(x,y) 2T, \quad \Delta E = 2Aa TS(x,y) \cos[2\pi\alpha x + \phi(x,y)].$$

$$\begin{aligned} E^- &= \left[A^2 + a^2 \left| S \left(\frac{x}{\lambda f}, \frac{y}{\lambda f} \right) \right|^2 \right] T, \\ \Delta E &= 2Aa T \left| S \left(\frac{x}{\lambda f}, \frac{y}{\lambda f} \right) \right| \cos \left[2\pi\alpha x + \phi \left(\frac{x}{\lambda f}, \frac{y}{\lambda f} \right) \right]. \end{aligned}$$

Choosing the term of transmittance of the processed transparency that is proportional to S^* , we have

$$\Delta t_A \propto \Delta E E^- \propto S^* K + |S|^2,$$

$$\Delta t_A \propto \frac{\Delta E}{E} \propto \frac{S^*}{K + |S|^2},$$

(10-36)

where

$$K = A^2 a^2,$$

$$K = \frac{A^2}{a^2},$$

(often called the *beam ratio*), which is precisely the amplitude transmittance required for a Wiener filter when the signal and noise have flat power spectra with a ratio of noise power to signal power of K at all frequencies.

Both [Ragnarsson \[292\]](#) and [Tichenor and Goodman \[343\]](#) have demonstrated restorations with dynamic ranges of 100:1 in $|S|^{-1}|S|$ using this technique. [Figure 10.18](#) shows photographs of the blur impulse response, the magnitude of the deblur impulse response, and the impulse response of the cascaded blur and deblur filters, illustrating the restoration of a blurred point source. The deblurring operation becomes highly sensitive to optical noise at the input of the processor as the dynamic range of the deblurring operation increases. For example, dust specks and small phase perturbations on the input transparency generate deblur impulse responses in the output image which eventually mask the desired image detail [\[134\]](#).

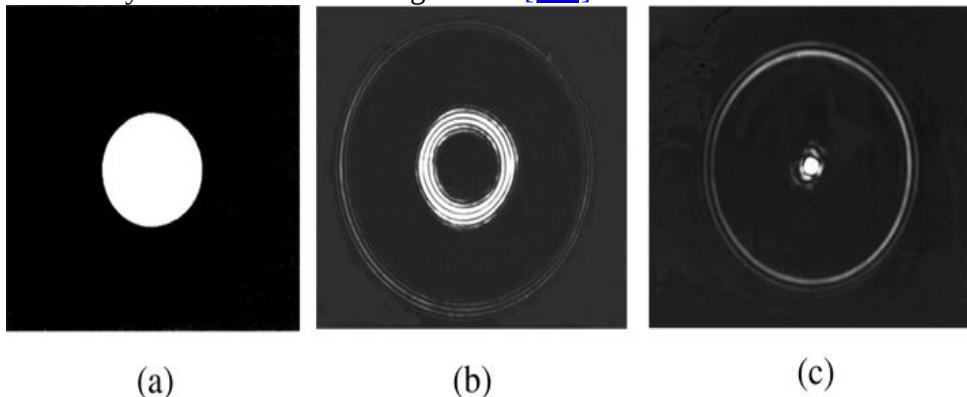


Figure 10.18
Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 10.18 Deblurring of the blur point-spread function. (a) The original blur, (b) the magnitude of the deblur point-spread function, and (c) the point-spread function of the blur-deblur sequence. [From [\[343\]](#). Copyright 1975 by the Optical Society of America, Inc., reprinted with permission.]

Image a shows a square patch of darkness with a big bright circular spot at the center. Image b shows a square patch of darkness where the intensity of darkness is less than that of Image a. A bright small circular ring is shown at the center and a big circular ring with reduced intensity of brightness surrounds the small ring. Image c shows a square patch of darkness with the intensity of darkness more than that of Image b but less than Image a. A small bright circular spot is shown at the center and two concentric circles with reduced brightness surrounds the circular spot.

The technique that has been described can be viewed as utilizing an amplitude transmittance saturation phenomenon that takes place when the object intensity exceeds the reference intensity in the average exposure $E^- E$. A similar phenomenon has been used in [\[266\]](#) to detect defects in periodic structures using a photorefractive crystal as the spatial light modulator.

10.7 Acousto-Optic Signal Processing Systems

The means by which a temporal electrical signal can be converted into a moving spatial optical signal with the help of an acousto-optic cell was discussed in [Section 9.5](#). Attention is turned here to the use of such cells as input transducers for various types of signal processing systems. Since virtually all modern work in this area utilizes microwave signals in crystals, we focus attention exclusively on *Bragg cells* as the input transducers. While systems based on Raman-Nath diffraction were important in the early days of acousto-optic signal processing [321], [9], they are virtually non-existent today.

Our discussion is of necessity brief, but we will describe three different system architectures. One, the Bragg cell spectrum analyzer, can be used to analyze the spectrum of broadband microwave signals. Attention is then turned to two types of acousto-optic correlators, the space-integrating correlator and the time-integrating correlator.

10.7.1 Bragg Cell Spectrum Analyzer

The ease with which Fourier transforms can be performed in coherent light suggests that a system that combines an acousto-optic input transducer with a coherent optical Fourier transform system can function as a spectrum analyzer for wideband and high-frequency electrical signals. [Figure 10.19](#) shows the basic structure of such a spectrum analyzer.

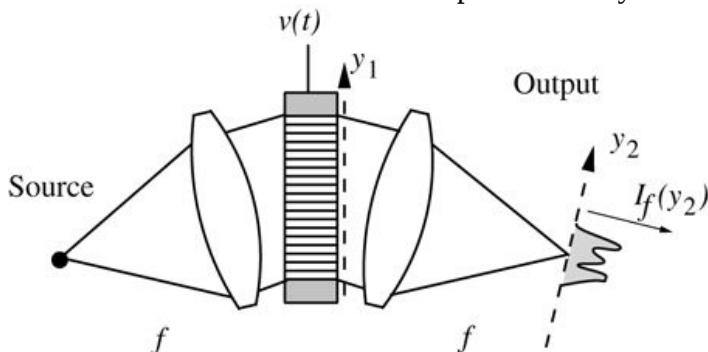


Figure 10.19

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 10.19 Bragg cell spectrum analyzer.

The illustration shows a dark spot labeled source at the extreme left. To the right of the source, is a downward sloping lens which is at a distance f from the source. Rays diverge from the source to the lens. From the lens parallel rays extend in an upward slope to the horizontally striped part of a rectangular bar whose longer side is oriented vertically. An arrow labeled $v(t)$ emerges from the upper center of the bar. Beyond the stripes, the upper and lower ends of the bar are shaded. A vertical arrow is shown along the length of the bar and is labeled y_1 . To the right of the vertical rectangular bar, is an upward sloping lens. Parallel downward sloping rays from the bar extend to a lens, beyond which the rays converge to give the output on an upward sloping dotted output arrow labeled y_2 , at a distance f from the second lens. A wavelike pattern is at the center of the lens. A rightward arrow perpendicular to y_2 is labeled $I_f(y_2)$.

Consider a high-frequency signal represented by the electrical voltage

$$v(t) = A(t) \cos[2\pi f_c t - \psi(t)] = \operatorname{Re} A(t) e^{j\psi(t)} e^{-j2\pi f_c t} = \operatorname{Re} s(t) e^{-j2\pi f_c t},$$

$$v(t) = A(t) \cos[2\pi f_c t - \psi(t)] = \operatorname{Re} \left\{ A(t) e^{j\psi(t)} e^{-j2\pi f_c t} \right\} = \operatorname{Re} \left\{ s(t) e^{-j2\pi f_c t} \right\},$$

(10-37)

where $s(t) = A(t) e^{j\psi(t)}$ is the complex representation of the signal.

With reference to (9-32) and Fig. 10.19, let the coordinate y_1 refer to the plane where the transmitted field exits the Bragg cell, and let the coordinate y_2 refer to the plane in the rear focal plane of the Fourier transforming lens. When the above signal is applied to an acousto-optic cell, and the cell is illuminated at the Bragg angle by a collimated, monochromatic wave, there results a transmitted wavefront into the -1^{-1} diffracted order given by

$$U(y_1; t) = C s(y_1; t) e^{-j2\pi y_1 / \Lambda} \operatorname{rect} \frac{y_1}{L}$$

$$U(y_1; t) = C s \left(\frac{y_1}{V} + t - \tau_o \right) e^{-j2\pi y_1 / \Lambda} \operatorname{rect} \frac{y_1}{L}$$

where C is a constant and we have neglected the temporal frequency up-shift by f_c , since it has no impact on our calculations.

This optical signal now passes through a positive lens tilted at the Bragg angle, as shown in Fig. 10.19. Noting that the linear phase factor in y_1 is canceled by the tilt of the lens, the spatial distribution of fields appearing in the back focal plane of the lens will be (aside from quadratic-phase factors in y_2 which we can neglect)

$$U_f(y_2; t) = C' \int_{-\infty}^{\infty} s \left(\frac{y_1}{V} + Vt - V\tau_o \right) \operatorname{rect} \frac{y_1}{L} \exp \left(-j \frac{2\pi y_1 y_2}{\lambda f} \right) dy_1.$$

$$U_f(y_2; t) = C' \int_{-\infty}^{\infty} s \left(\frac{y_1 + Vt - V\tau_o}{V} \right) \operatorname{rect} \frac{y_1}{L} \exp \left(-j \frac{2\pi y_1 y_2}{\lambda f} \right) dy_1.$$

(10-38)

This is a Fourier transform of a product of two functions, so the convolution theorem will apply, and we consider each of the two spectra individually. Consider the Fourier transform of the scaled signal first; we have

$$\begin{aligned} & \mathcal{F} \{ s \left(\frac{y_1 + Vt - V\tau_o}{V} \right) \} = V S(Vf_Y) \exp[j2\pi f_Y V(t - \tau_o)] \\ & \quad (10-39) \end{aligned}$$

where $S = \mathcal{F}_s \{ s \}$. The presence of the term depending on time t in this result is an indication that *every spatial frequency component is oscillating with a different optical frequency*. Considering the rect function next, we have

$$\mathcal{F}\text{rect}y_1L=L\text{sinc }LfY.$$

$$\mathcal{F}\left\{\text{rect}\frac{y_1}{L}\right\}=L \text{ sinc } Lf_Y.$$

For the moment, neglect the finite length of the Bragg cell, allowing L to become arbitrarily large, in which case the sinc function approaches a δ function. The optical intensity incident in the focal plane will then be (neglecting multiplicative constants)

$$I_f(y_2)=SVy2\lambda f\exp(j2\pi\lambda fV_y2(t-\tau_o))=SVy2\lambda f^2.$$

$$I_f(y_2)=\left|S\left(\frac{V_y2}{\lambda f}\right)\exp\left[j\frac{2\pi}{\lambda f}V_y2(t-\tau_o)\right]\right|^2=\left|S\left(\frac{V_y2}{\lambda f}\right)\right|^2.$$

(10-40)

The intensity distribution measured by an array of time-integrating detectors will therefore be proportional to the *power spectrum* of the input signal, and the acousto-optic system acts as a *spectrum analyzer*.

The relationship between position y_2 in the focal plane and temporal frequency f_t of the input signal can be found by first noting that the center frequency f_c of the electrical signal corresponds to the origin of the y_2 plane (we choose the origin to make this true). As we move in the positive y_2 direction, we are moving to lower temporal frequencies (zero temporal frequency corresponds to the direction of the zero order of the acoustic grating). From the scaling factors present in the equation above we find that the temporal input frequency corresponding to coordinate y_2 is

$$f_t=f_c-Vy2\lambda f.$$

$$f_t=f_c-\frac{V_y2}{\lambda f}.$$

However, when only the time integrated intensity of the light is detected, the temporal frequency of the light is of no consequence.

When the length of the Bragg cell is not infinite, convolution with the sinc function in the spectral domain cannot be neglected. This convolution in effect smooths the measured spectrum, establishing the frequency resolution obtainable. The minimum resolvable difference in temporal frequency is readily shown to be approximately the reciprocal of the total time delay stored in the Bragg cell window, i.e.

$$\Delta f_t=V/L.$$

$$\Delta f_t = V / L.$$

The technology of high-performance Bragg cell spectrum analyzers is well developed. Center frequencies lie in the hundreds of MHz to the 1- to 3-GHz range, and time-bandwidth products

(equal to the number of resolvable spectral elements) from several hundred to more than 1,000 have been reported.

10.7.2 Space-Integrating Correlator

Bragg cells can also be used as real-time inputs to convolvers and correlators. Historically the first such systems were based on what is now called a space-integrating architecture. Consider the acousto-optic system shown in Fig. 10.20. This system contains one Bragg cell, which is used for converting a temporal voltage $v_1(t)$ into a complex distribution of field $s_1 y_1 V + t - \tau_o$. Here s_1 is the complex representation of an amplitude and phase modulated voltage, analogous to the representation of (10-37). Due to the Bragg effect, the cell is assumed to transmit only the zero order and the -1 order.

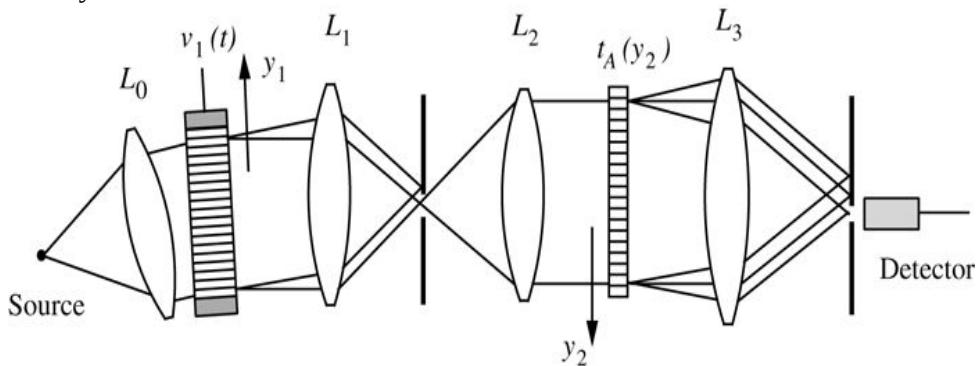


Figure 10.20
Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 10.20 Acousto-optic space-integrating correlator.

The illustration shows a dark spot labeled source at the extreme left. To the right of the source, is lens L_0 at a distance f from the source. Rays from source diverge to L_0 . To the right of the lens, is a vertical rectangular bar with horizontal stripes and an arrow labeled $v_1(t)$ emerges from the upper end of the bar. The upper and lower ends of the bar are shaded. To the right of the bar is a vertical upward arrow y_1 . To the right of the vertical rectangular bar, is Lens L_1 . Parallel rays run from L_0 to the striped section of the bar, on the other side of which, from the top and lower ends, rays diverge to the top and lower ends of lens L_1 . To the right of L_1 , a thick vertical line is shown which has an opening at the center. From the top end of L_1 two downward sloping rays reach the vertical line, the lower one passing through the opening and reaching the lower end of L_2 on the other side. Similarly, from the lower end of L_1 , two upward sloping rays reach the vertical line, the upper one passing through the opening and reaching the upper end of L_2 . The upper rays from both ends of L_1 meet at the same point just above the opening.

From L_2 parallel rightward rays run to a vertical rectangular bar labeled $t_A(y_2)$ with horizontal stripes, next to which is a downward arrow labeled y_2 . From the top and bottom ends of this bar, three rays diverge rightward to the top and bottom extremes of Lens L_3 .

To the right of L_3 , a thick vertical line is shown with an opening at its center, behind which is a box with a rightward connection, all together labeled "Detector." The topmost rays from the top and bottom ends of L_3 meet above the opening. The middle ones meet just above the opening. And the lowermost meet just below the opening.

The second input is provided by a fixed transparency which contains an amplitude and phase modulated grating. If $s_2 = B \exp(j\chi)$ represents the second signal, with which s_1 is to be correlated, then the amplitude transmittance of the transparency should ideally be chosen to be

$$t_A(y_2) = 1 + B(y_2) \cos(2\pi f_Y y_2 - \chi(y_2)) = 1 + B(y_2) e^{-j\chi(y_2)} e^{j2\pi f_Y y_2} + B(y_2) e^{j\chi(y_2)} e^{-j2\pi f_Y y_2}.$$

$$\begin{aligned} t_A(y_2) &= \frac{1}{2} [1 + B(y_2) \cos[2\pi f_Y y_2 - \chi(y_2)]] \\ &= \frac{1}{2} + \frac{B(y_2)}{4} e^{-j\chi(y_2)} e^{j2\pi f_Y y_2} + \frac{B(y_2)}{4} e^{j\chi(y_2)} e^{-j2\pi f_Y y_2}. \end{aligned}$$

(10-41)

Such a grating could be computer generated. Alternatively a transparency with the same two first-order grating components could be recorded interferometrically, in a manner analogous to that used for the VanderLugt filter. It is assumed to be a thin grating, so a zero order and two first orders are generated.

The optical system following the Bragg cell contains a stop in the spatial frequency domain that blocks the zero-order transmitted component and passes the $-1 - 1$ diffracted component. Lenses L_1 and L_2 together image the amplitude distribution corresponding to the first diffraction order onto the fixed transparency, with inversion. The y_2 coordinate system is inverted to account for the inversion associated with the imaging operation. Lens L_3 is used to bring the $-1 - 1$ order component diffracted by the fixed grating to focus on a pinhole, which is followed by a nonintegrating photodetector.

The operation performed by lens L_3 , the pinhole, and the detector can be expressed as a spatial integration of the product of the two complex functions of interest. The current generated by the detector is therefore (up to multiplicative constants)

$$id(t) = \int_{-\infty}^{\infty} s_1(y_2) s_2^*(y_2) \text{rect}\left(\frac{y_2}{L}\right) dy_2$$

$$i_d(t) = \left| \int_{-\infty}^{\infty} s_1\left(\frac{y_2 + Vt - V\tau_o}{V}\right) s_2^*(y_2) \text{rect}\left(\frac{y_2}{L}\right) dy_2 \right|^2$$

(10-42)

where L is again the length of the Bragg cell, and the complex conjugate of s_2 occurs because we have chosen the $-1 - 1$ diffraction order of the fixed grating. As time progresses, the scaled signal s_1 slides through the Bragg cell and the relative delay between s_1 and s_2 changes, thus providing the values of the correlation between the two signals for different delays. The correlation operation takes place only within the window provided by the finite length of the Bragg cell.

The distinguishing characteristics of the space-integrating correlator are that the correlation integration is over space and the various values of relative delay occur sequentially in time.

10.7.3 Time-Integrating Correlator

An entirely different approach to realization of an acousto-optic correlator is provided by a system that interchanges the roles of time and space vis-à-vis the space-integrating correlator. Such an approach was first conceived of by [Montgomery \[259\]](#). A different architecture that accomplishes a similar operation was demonstrated by Sprague and Koliopoulos [\[330\]](#). This general approach to correlation is known as “time-integrating correlation.”

[Figure 10.21](#) shows one architecture of such a correlator. Two RF voltages $v_1(t)$ and $v_2(t)$ are applied to different Bragg cells in close proximity, arranged so that the resulting acoustic signals propagate in opposite directions. The lenses L_1 and L_2 form a standard double Fourier transform imaging system. A light ray entering the first Bragg cell exits as a zero-order ray and a -1 order ray.⁹ Both the zero order and the -1 order transmitted by the first Bragg cell are split by the second cell, which itself applies zero-order and $+1$ order diffraction to each of those incident rays. A stop in the rear focal plane of L_1 passes only rays that have undergone the sequence $0 \rightarrow +1$ order diffractions or $-1 \rightarrow 0$ order diffractions, blocking the rays that have undergone two first-order diffractions and two 0-order diffractions. Note that the optical frequencies of the two beams passed are identical due to the opposite directions in which the acoustic waves are propagating. The two optical signals passed by the aperture-stop are then brought back together on an array of time-integrating detectors, which is situated in a plane where the product of the amplitude transmittances of the two Bragg cells is imaged.

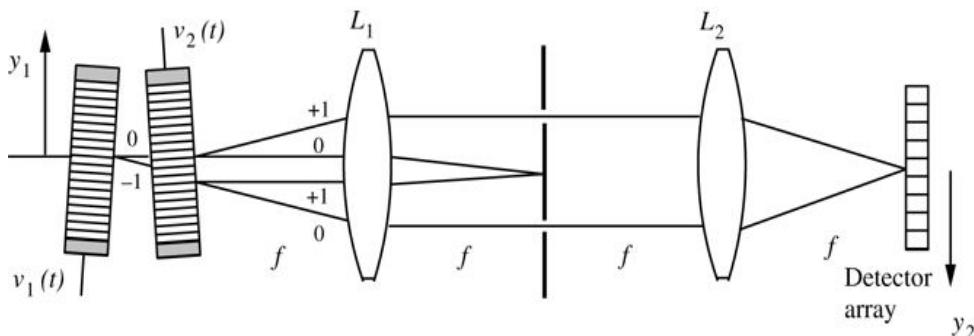


Figure 10.21

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 10.21 Time-integrating correlator.

The illustration shows a small horizontal line with an upward vertical arrow Y1 emerging from its centre and ends at the center of a vertical rectangular bar. The vertical bar has horizontal stripes and the upper and lower portions of the bar are shaded. A vertical line $v_1(t)$ emerges from the bottom center of the rectangular bar. To its right, is another similar rectangular bar which is tilted to the left. A line labeled 0 emerges from the centre of the first rectangular bar and points toward the center of the second rectangular bar. Another line labeled -1 emerges from the centre of the first rectangular bar and points toward the second rectangular bar at a point slightly below the center. A vertical line $v_2(t)$ emerges from the top center of the second rectangular bar. To the right of the second vertical rectangular bar, is Lens L_1 . A straight line starts from the center of the second vertical rectangular bar and points toward the center of L_1 . Another sloping line emerges

from the same center point of vertical rectangular bar and points toward L1 at a point above the center of L1. A straight line emerges from a point below the center of second vertical rectangular bar and points toward L1 at a point slightly below the midpoint of L1. A sloping line starts from the vertical rectangular bar at the same point as that off horizontal line and point toward L1 at a point slightly above the lower end. To the right of L1, a thick vertical line with two openings, one near the upper end and one near the lower end is shown. To its right is Lens L2. A straight line emerges from a point below the upper end of L1 and passes through the first opening of vertical line and ends at a point near the upper end of L2. A straight line emerges from a point above the lower end of L1 and passes through the second opening of vertical line and ends at a point near the upper end of L2. Two lines emerge from L1, one from a point above the midpoint of L1 and another from a point below the midpoint of L1 and both point toward the center of the vertical line. To the right of L2, is a vertical bar labeled detector array with horizontal lines inside. Two lines emerge from L2, one from a point slightly below the upper end and another above the lower end and both point toward the center of the detector array. A downward arrow y_2 to the right of detector array is shown. The distance between the second vertical rectangular bar and L1, between L1 and the vertical line, between vertical line and L2, between L2 and detector array are each f .

Note that each element of the detector array measures the intensity associated with a different vertical location on the two Bragg cells, and as a consequence, for each detector element there is a different relative time delay between the two signals driving the cells, due to the opposite

directions of acoustic wave propagation. If $s_1(y_1)$ is the complex representation of the signal in the first cell, and $s_2(y_1)$ is the complex representation of the signal in the second cell, a detector at location y_2 measures the finite time integral of the squared magnitude of the sum of the two fields (both conjugated) that have traveled the two different paths to the detector. Neglecting multiplicative constants, the integral in question is given by

$$E(y_2) = \int_{\Delta T} \left| s_1^* \left(-\frac{y_2}{V} + t + \tau_o \right) e^{-j2\pi\alpha_c y_2} + s_2^* \left(\frac{y_2}{V} + t + \tau_o \right) e^{j2\pi\alpha_c y_2} \right|^2 dt,$$

$$E(y_2) = \int_{\Delta T} \left| s_1^* \left(-\frac{y_2}{V} + t + \tau_o \right) e^{-j2\pi\alpha_c y_2} + s_2^* \left(\frac{y_2}{V} + t + \tau_o \right) e^{j2\pi\alpha_c y_2} \right|^2 dt, \quad (10-43)$$

where the linear exponential terms of opposite sign account for the opposite angles with which the two components arrive at the detector plane, and ΔT is the finite integration time.

Considering the various parts of this integral, the terms

$$E_1 = \int_{\Delta T} s_1 \left(-\frac{y_2}{V} + t + \tau_o \right)^2 dt$$

$$\begin{aligned} E_1 &= \int_{\Delta T} \left| s_1 \left(-\frac{y_2}{V} + t + \tau_o \right) \right|^2 dt \\ E_2 &= \int_{\Delta T} \left| s_2 \left(\frac{y_2}{V} + t + \tau_o \right) \right|^2 dt \end{aligned}$$

will approach constants as the integration time ΔT grows large. The remaining term is

$$E_3 = \int_{\Delta T} [\int_{\Delta T} s_1(t + \tau_o - \frac{y_2}{V}) e^{j2\pi\alpha_c y_2} s_2^*(t + \tau_o + \frac{y_2}{V}) e^{j2\pi\alpha_c y_2} + cc] dt$$

$$\begin{aligned} &= 2\text{Re} \left\{ e^{j4\pi\alpha_c y_2} \int_{\Delta T} s_1(t + \tau_o - \frac{y_2}{V}) s_2^*(t + \tau_o + \frac{y_2}{V}) dt \right\} \\ &= 2\text{Re} \left\{ e^{j4\pi\alpha_c y_2} \int_{\Delta T'} s_1(t') s_2^*(t' + \frac{2y_2}{V}) dt' \right\}, \end{aligned}$$

(10-44)

where a simple change of variables has been made in the last line, $\Delta T'$ is of the same duration as ΔT , but shifted in accord with the variable change, and cc stands for the complex conjugate of the previous term.

Let the complex function $c(\tau) e^{j\phi(\tau)}$ represent the complex finite-time crosscorrelation of s_1 and s_2 ,

$$c(\tau) = \int_{\Delta T} s_1(t) s_2^*(t + \tau) dt = |c(\tau)| e^{j\phi(\tau)}.$$

$$c(\tau) = \int_{\Delta T'} s_1(t') s_2^*(t' + \tau) dt' = |c(\tau)| e^{j\phi(\tau)}.$$

Then the last line of (10-44) becomes

$$E_3 = 2\text{Re} \left\{ e^{j4\pi\alpha_c y_2} c(2y_2/V) \right\} = 2|c(2y_2/V)| \cos[4\pi\alpha_c y_2 + \phi(2y_2/V)].$$

$$E_3 = 2\text{Re} \left\{ e^{j4\pi\alpha_c y_2} c(2y_2/V) \right\} = 2|c(2y_2/V)| \cos[4\pi\alpha_c y_2 + \phi(2y_2/V)].$$

(10-45)

If the detectors are small compared with the period $1/(2\alpha_c)$ of the spatial carrier frequency, the fringe pattern incident on the detector will be sampled at or above the Nyquist rate, and the complex correlation information will be captured by the detector array. If the detector array is of the charge-coupled-device (CCD) type, the measured intensities are read out serially from the array. As a result there is an AC output from the CCD array, which can be isolated from the DC output components with a highpass or bandpass filter. The amplitude of the AC component represents the magnitude of the complex correlation coefficient, and the phase is the phase of that coefficient. By using an envelope detector, the magnitude of the complex correlation can be measured. To measure the phase, synchronous detection must be used. Generally it is the magnitude information that is of most interest.

Note that for this architecture, each detector element measures the correlation for a different relative delay $2y_2/V$ between the two signals. The time-bandwidth product of the correlation measurement is determined by the integration time of the detector array, and is no longer limited to the delay time of the acoustic cell; rather, the delay time determines the range of relative delays that can be explored. In practice the integration time is limited by accumulation of

dark current and by detector saturation that is ultimately introduced by the constant terms E_1 and E_2 .

The architecture described above is that of [Montgomery \[259\]](#). The architecture of [Sprague and Koliopoulos \[330\]](#) differs in that only a single Bragg cell is used and the second signal is introduced by temporal modulation of the optical source. The reader should consult the reference for details.

10.7.4 Other Acousto-Optic Signal Processing Architectures

A multitude of other acousto-optic signal processing architectures exist, but they will not be covered here. We mention in particular various extensions of acousto-optic systems to two-dimensional processing (see [\[356\]](#), Chapter 15, for examples). Applications of acousto-optic processing to numerical or digital computation are omitted here.

10.8 Discrete Analog Optical Processors

Until now, we have considered only optical systems that process continuous analog optical signals. Attention is now turned to another class of systems, namely those that process *discrete* analog optical signals. Discrete signals arise in many different applications. For example, an array of sensors collects a discrete set of measurements. Those measurements may be changing continuously with time, but because there is a discrete array of sensors, only a discrete array of data is available at any one time. In addition, it is often necessary to discretize continuous data in order to subject it to processing. Thus discrete data can arise in a multitude of different ways.

The discreteness does not imply that the data is digital. Quite the contrary, the data of interest here has analog values, which have not been quantized, but there is a finite set of such data to be processed. All of the optical processing systems we shall describe are analog processing systems, in keeping with our earlier restrictions.

10.8.1 Discrete Representation of Signals and Systems

Any continuous signal s^S dependent on a time coordinate t^T and/or space coordinates (x, y) can be sampled in a discrete array of data representable by a *vector* of values

$$s \rightarrow = s_1 s_2 : s_N.$$

$$\vec{s} = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_N \end{bmatrix}.$$

(10-46)

If the sample values are taken sufficiently close together and if the function s^S is bandlimited or nearly bandlimited, we know that it would be possible to reconstruct s^S either exactly (in the bandlimited case) or with high accuracy (in the almost bandlimited case). Therefore the vector $s \rightarrow$

\vec{s} is a suitable representation of the original data. Note that if the signal s^S arose from a discrete array of sensors, then each component of $s \rightarrow \vec{s}$ may be a function of time.

For discrete signals, the superposition integral becomes a matrix-vector product (see [216], [Chapter 6](#)). Thus the output $g \rightarrow \vec{g}$ (with M^M samples) of a linear system having $s \rightarrow \vec{s}$ at the input (N^N samples) is represented by

$$g \rightarrow = H s \rightarrow,$$

$$\vec{g} = \mathbf{H} \vec{s},$$

(10-47)

where \mathbf{H} is a matrix with M rows and N columns,

$$\mathbf{H} = [h_{11} \ h_{12} \ \dots \ h_{1N} \ ; \ h_{21} \ h_{22} \ \dots \ h_{2N} \ ; \ \vdots \ ; \ h_{M1} \ h_{M2} \ \dots \ h_{MN}]$$

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1N} \\ h_{21} & h_{22} & \dots & h_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ h_{M1} & h_{M2} & \dots & h_{MN} \end{bmatrix}$$

(10-48)

and

$$\mathbf{g} \rightarrow = [g_1 \ g_2 \ \dots \ g_M]$$

$$\vec{\mathbf{g}} = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_M \end{bmatrix}$$

(10-49)

Note that there are $M \times N$ analog multiplications and $M \times N$ analog additions¹⁰ needed to perform the above computation.

Thus in discrete signal processing, the matrix-vector product is as fundamental as the superposition integral, or as its special case, the convolution integral. It is therefore of great interest to devise methods for performing such operations optically, preferably using the parallelism of optical systems to full advantage.

10.8.2 A Parallel Incoherent Matrix-Vector Multiplier

A fully parallel incoherent matrix-vector processor is illustrated in [Fig. 10.22 \[136\]](#) which temporarily omits the details of the optical elements used. This architecture has come to be known as the “Stanford matrix-vector multiplier.” This system is fundamentally parallel, due to the entry of *all* elements of the signal vector $\mathbf{s} \rightarrow \vec{\mathbf{s}}$ simultaneously, in one clock cycle.

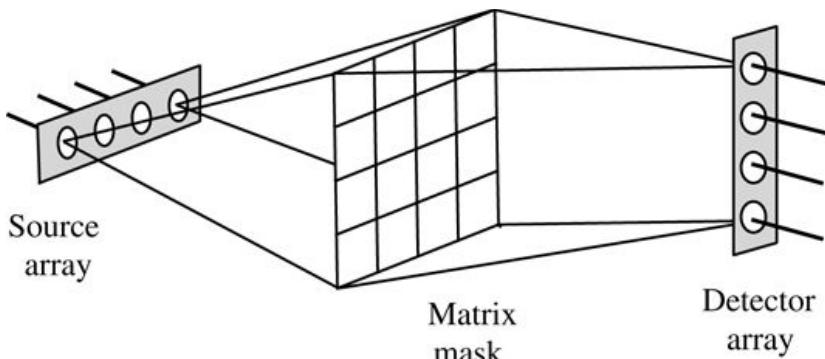


Figure 10.22

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 10.22 A fully parallel incoherent matrix-vector multiplier.

The illustration shows a rectangular bar labeled source array with four holes arranged horizontally along the length of the rectangle. Four lines pointing toward the left emerge from the bar. To the right of the source array, is a square grid labeled matrix mask. Two lines emerge from the hole at the left extreme with one pointing toward the top left corner and other one pointing toward the bottom left corner. Two lines emerge from the hole at the right extreme with one pointing toward the top right corner and other one pointing toward the bottom right corner. To the right of the matrix mask is the detector array which is a vertical rectangular bar with holes arranged vertically. A line emerges from each hole and points toward the right. Two lines emerge from the top left and bottom left corners and point towards the bottom hole of the detector array. Two lines emerge from the top right and bottom right corners and point towards the top hole of the detector array.

The optics before the mask are arranged so that the light from any one input source, which may be an LED or a laser diode, is spread vertically and imaged horizontally, so that it fills a single vertical column of the mask. Each source thus illuminates a different column. The optics following the mask are arranged so that light from each row of the matrix mask is focused horizontally and imaged vertically, so that it falls upon a single detector element in the output detector array. Thus the light transmitted by each row of the mask is summed optically on a unique detector element. The detectors used here do not integrate charge but rather respond as fast as possible, generating output signals that vary in unison with the variations of the light intensities falling upon them.

In effect, the input vector \vec{s} is spread vertically so that each output detector can measure an inner product of that vector with a different row vector stored in the matrix mask. For this reason, such a processor is sometimes called an “inner product processor.”

There are several different ways to construct an optical system that will achieve the operations indicated diagrammatically in Fig. 10.22. Figure 10.23 shows one such arrangement of elements. Note that because different operations are desired in the horizontal and vertical dimensions, both before and after the matrix mask, the optical system must be anamorphic, with combinations of spherical and cylindrical lenses. The operation of this optical system is as follows.

Each of the lenses, whether spherical or cylindrical, has a focal length f . The combination of a spherical and cylindrical lens in close contact has a focal length that is $f/2$ in the direction for which the cylinder has power, and f in the direction for which the cylinder has no power.

Thus such a pair will collimate light diverging in the direction with weaker power, and image light diverging in the direction of stronger power. As a result, the lens combination L_1 , L_2 L_2 collimates the light diverging vertically from an input source, but images in the horizontal direction, thereby illuminating a column of the matrix mask. Similarly, the lens combination L_3 L_3 , L_4 images a row of the mask onto the vertical position of a single detector element but collimates or spreads the light from a single column of the matrix mask. Ideally the detector elements should be long in the horizontal direction, allowing detection of most of the light across a row of the mask, but a long detector has high capacitance, and such capacitance limits the bandwidth of the electronic channel that follows.

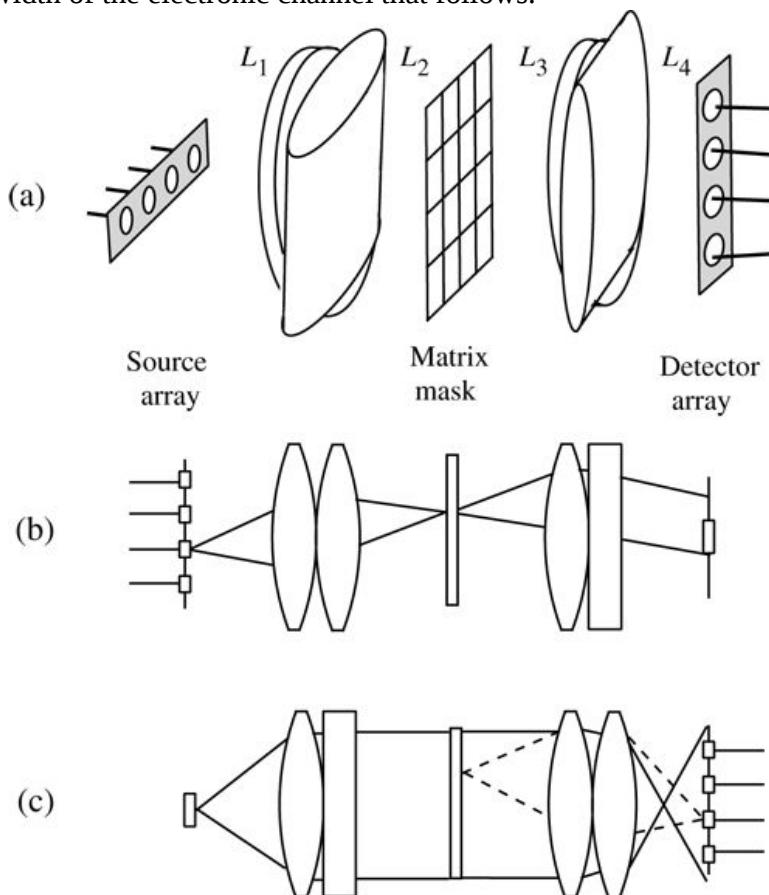


Figure 10.23

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 10.23 Optical elements comprising the parallel matrix-vector multiplier: (a) perspective view, (b) top view, and (c) side view.

Illustration a shows a rectangular bar labeled source array with four holes arranged horizontally along the length of the rectangle. Four lines pointing toward the left emerge from the bar. To the right of the rectangular bar is Lens L_1 . Lens L_2 adjoins L_1 . To the right of L_2 is a square grid labeled matrix mask. To its right, is a Lens L_3 . Lens L_4 adjoins L_3 . To the right of L_4 , is the detector array which is a vertical rectangular bar with holes arranged vertically one below the other. A line emerges from each hole and points toward the right.

Illustration b shows four small square boxes stacked vertically and a small vertical line connects one box with the other. Four small horizontal lines emerge from the center of each box and point toward the left. To the left of the boxes, is a lens. Two lines emerge from the third box with one pointing at a point above the midpoint of the lens and another at a point below the midpoint of the lens. Another lens adjoins this lens. To the right of the lens, is a vertical lengthy rectangular bar and next to it, is the third lens. The fourth lens in the shape of vertical rectangle adjoins the third lens. A line from a point slightly above the midpoint of the second lens passes through the vertical lengthy bar and ends at the center of the third lens. A line from a point slightly below the midpoint of the second lens passes through the vertical lengthy bar and ends at the midpoint of L3. Next to L4, is a short vertical rectangular bar with a line emerging from both the upper and lower ends. A line emerges from a point little below the upper end of third lens and passes through L and ends at the line above the vertical rectangular bar. Another line emerges from the midpoint of L4 and ends at the base of the vertical rectangular bar.

Illustration c shows a small vertical rectangular bar and to its right, is a lens. Another vertical rectangle shaped lens adjoins the first one. A line emerges from the center of the small rectangular bar and points toward a point slightly below the upper end of the first lens. Another line emerges from the center of the small rectangular bar and points toward a point slightly above the lower end of the first lens. Next to the second lens, is a lengthy thin rectangular bar. A line emerging from a point slightly below the upper end of the second lens points toward the upper end of the rectangular bar. A line emerging from a point slightly below the upper end of the second lens points toward the upper end of the rectangular bar. A line emerging from a point slightly above the lower end of the second lens points toward the lower end of the rectangular bar. Next to the rectangular bar, is the third lens. Another lens adjoins the third lens. A line emerges from the upper end of the vertical rectangular bar and points to a point slightly below the upper end of L3 and the line further leads to a point slightly below the upper end of L4. A line emerges from the lower end of the vertical rectangular bar and points to a point slightly above the lower end of L3 and the line further leads to a point slightly above the lower end of L4. Next to L4, are four small square boxes stacked vertically and a small vertical line connects one box with the other with a small vertical line emerging from the top and bottom ends. Four small horizontal lines emerge from the center of each box and point toward the right. A line emerges from a point slightly below the upper end of L4 and point towards the bottom end of the four small square boxes. Another line emerges from a point slightly above the lower end of L4 and point towards the upper end of the four small square boxes. A dotted line emerges from the vertical rectangular bar at a point slightly above the midpoint and points toward a point slightly below the upper end of L3 and another dotted line emerges from a point slightly below the upper end of L4 and points toward the third box. Another dotted line emerges from the vertical rectangular bar at the same point as the other dotted line and points toward a point slightly below the midpoint of L3 and another dotted line emerges from a point slightly below the midpoint of L4 and points toward the third box.

The parallel matrix-vector multiplier performs all $N \times M$ multiplications and additions in a single clock cycle. A clock cycle can be very short, for example 10 nsec, depending on the amount of light available from each source. Lasers can be used as the sources, in spite of the fact that incoherent addition is used by this system, due to the fact that all additions are of light from different lasers, which for most types of semiconductor lasers will be mutually incoherent on the time scale of a clock cycle.

Many applications of the parallel matrix-vector multiplier architecture have been proposed and demonstrated. These include the construction of an optical crossbar switch [91], [242], the

construction of Hopfield neural networks [106], and others. The architecture has been a useful workhorse of the optical information processing field in the past.

10.8.3 Methods for Handling Bipolar and Complex Data

Until now we have assumed that the elements of both the input vector and the system matrix are nonnegative real numbers, an assumption that ensures compatibility with the nonnegative real character of incoherent optical signals. To handle bipolar data or complex data, two different methods can be utilized, either individually or together. For the purposes of this discussion, we focus on the parallel matrix-vector multiplier, although the methods are more widely applicable.

The first method places all bipolar signals on a bias, with the bias chosen large enough so that all elements of the input vector and all elements of the system matrix remain nonnegative. The biasing operation can be represented mathematically by noting that the input vector is now the sum of the signal vector $s \rightarrow \vec{s}$ and a bias vector $b \rightarrow \vec{b}$ (all bias elements assumed identical), and the system matrix is likewise the sum of two matrices, $H \mathbf{H}$ and $B \mathbf{B}$. The output of the system now becomes

$$H + Bs \rightarrow +b \rightarrow = Hs \rightarrow +Hb \rightarrow +Bs \rightarrow +Bb \rightarrow .$$

$$(H + B)(\vec{s} + \vec{b}) = H \vec{s} + H \vec{b} + B \vec{s} + B \vec{b} .$$

(10-50)

If the bias matrix and the bias vector are known and constant over time, then the last term can be subtracted from the output electronically. In addition, the matrix $H \mathbf{H}$ is known *a priori*, so the product $Hb \rightarrow H \vec{b}$ can be calculated in advance and subtracted from any result. However, the vector $s \rightarrow \vec{s}$ is not known in advance, and therefore it is generally necessary to measure its inner product with a row vector of the bias matrix, perhaps by adding a simple extra bias row to the matrix $H \mathbf{H}$ and one extra element to the detector array.

An alternative approach to handling bipolar elements is to represent the input vector and the system matrix as the *difference* of two nonnegative vectors or two nonnegative matrices,

respectively. Thus $H = H_+ - H_-$ and $s \rightarrow = s_+ - s_-$, where $\vec{s} = \vec{s}_+ - \vec{s}_-$, where the matrix H_+ contains positive elements only in those locations where $H \mathbf{H}$ contains positive elements, and zero elsewhere, and H_- contains positive elements equal to the magnitude of any negative elements of $H \mathbf{H}$ and zero for all other elements, with a similar construction procedure for s_+ and s_- . In addition, the output vector $g \rightarrow \vec{g}$ can be similarly decomposed. It is now easily shown that the nonnegative components of the output vector are related to the similar components of the input vector and the system matrix by

$$g \rightarrow + = H_+ s_+ + H_- s_- - g \rightarrow - = H_- s_+ + H_+ s_- .$$

$$\vec{g}_+ = H_+ \vec{s}_+ + H_- \vec{s}_- .$$

$$\vec{g}_- = H_- \vec{s}_+ + H_+ \vec{s}_- .$$

(10-51)

A simpler way of stating this relation is to stack $s \rightarrow +$ \vec{s}^+ and $s \rightarrow -$ \vec{s}^- in a longer column vector, and to do the same for the two parts of $g \rightarrow \vec{g}$, yielding

$$g \rightarrow + g \rightarrow - = H + H - H - H + s \rightarrow + s \rightarrow -.$$

$$\begin{bmatrix} \vec{g}^+ \\ \vec{g}^- \end{bmatrix} = \begin{bmatrix} H_+ & H_- \\ H_- & H_+ \end{bmatrix} \begin{bmatrix} \vec{s}^+ \\ \vec{s}^- \end{bmatrix}.$$

(10-52)

From this result we can see that a doubling of the two dimensions of the matrix mask to accommodate the larger matrix above, and a doubling of the length of the input vector, will allow the two components of $g \rightarrow \vec{g}$ to be computed without the use of biases. Those two output vectors must then be subtracted electronically, element by element.

When complex elements of the input vector and matrix are important, then the most straightforward approach is to quadruple the dimensions of the input vector, the output vector, and the matrix, thus allowing positive and negative real parts and positive and negative imaginary parts to be handled properly. More efficient decompositions can also be found [\[143\]](#).

Problems - Chapter 10

1. 10-1. An object has a periodic amplitude transmittance described by

$$t_A(\xi, \eta) = t_A(\xi) \cdot 1$$

$$t_A(\xi, \eta) = t_A(\xi) \cdot 1$$

where $t_A(\xi)$ is shown in Fig. P10.1. The object is placed in the object plane of the optical system shown in Fig. 10.1, and a tiny completely opaque stop is introduced on the optical axis in the focal plane, blocking only the spot on the optical axis. Sketch the intensity distribution observed in the image plane.

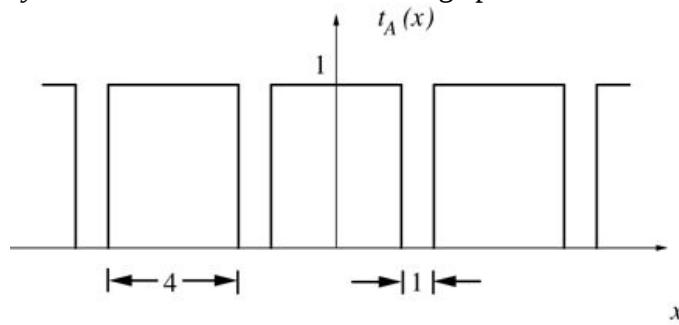


Figure P10.1
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure P10.1

The graph has the horizontal axis labeled x and the vertical axis labeled $t_A(x)$. A bar is on the horizontal axis with the vertical axis running at its center and it extends up to the point on the vertical axis. On the left of the center bar, is a bar of same height and the width of the bar is 4. And another bar is at a distance 1 from the bar at the center. On the left extreme, is a line of the same height as the bars and its end is bent toward the left. On the right extreme, is a line of the same height as the bars and its end is bent toward the right.

2. 10-2. The *central dark ground method* for observing phase objects is achieved by placing a tiny opaque stop on the optical axis in the focal plane to block the undiffracted light. Assuming that the variable component of phase shift through the object is always small compared with 2π radians, find the observed image intensity in terms of the object phase delay.
3. 10-3. The *schlieren* method for observing phase objects is achieved by introduction of a knife edge in the focal plane to block half of the diffracted light. The amplitude transmittance through the focal plane may be written

$$tf(x,y) = 12(1 + \text{sgn}x).$$

$$t_f(x, y) = \frac{1}{2}(1 + \operatorname{sgn} x).$$

1. Assuming a magnification of unity and neglecting image inversion, show that the image amplitude U_i is related to the object amplitude U_o by

$$U_i(u, v) = 12U_o(u, v) + j\pi \int_{-\infty}^{\infty} U_o(\xi, v) u - \xi d\xi.$$

$$U_i(u, v) = \frac{1}{2} \left[U_o(u, v) + \frac{j}{\pi} \int_{-\infty}^{\infty} \frac{U_o(\xi, v)}{u - \xi} d\xi \right].$$

2. Let the field transmitted by the object be of the form

$$U_o(\xi, \eta) = e^{j\phi_o} \exp[j\Delta\phi(\xi, \eta)]$$

$$U_o(\xi, \eta) = e^{j\phi_o} \exp[j\Delta\phi(\xi, \eta)]$$

where $\Delta\phi(\xi, \eta) \ll 2\pi$. Show that the image intensity can be approximated as

$$I_i(u, v) \approx 141 - 2\pi \int_{-\infty}^{\infty} \Delta\phi(\xi, v) u - \xi d\xi.$$

$$I_i(u, v) \approx \frac{1}{4} \left[1 - \frac{2}{\pi} \int_{-\infty}^{\infty} \frac{\Delta\phi(\xi, v)}{u - \xi} d\xi \right].$$

3. Find and sketch the image intensity distribution when

$$\Delta\phi = \Phi \operatorname{rect}\left(\frac{\xi}{U}\right)$$

$$\Delta\phi = \Phi \operatorname{rect}\left(\frac{\xi}{U}\right)$$

with the constant $\Phi \ll 2\pi$.

4. 10-4. Find an expression for the image intensity observed when the phase-shifting dot of the Zernike phase-contrast microscope is also partially absorbing, with intensity transmittance equal to α ($0 < \alpha < 1$).
5. 10-5. A certain coherent processing system has an input aperture that is 3 cm wide. The focal length of the initial transforming lens is 10 cm, and the wavelength of the light is $0.6328 \mu\text{m}$. With what accuracy must a frequency-plane mask be positioned in the focal plane, assuming that the mask has a structure comparable in scale size with the smallest structure in the spectrum of the input?
6. 10-6. It is desired to remove an additive periodic intensity interference of the form
- $$IN(\xi, \eta) = 121 + \cos 2\pi f_o \xi$$
- from a photograph taken by an imaging system. A coherent “4f” optical processing system will be used for that

removal. The wavelength of the coherent light is λ . The image was recorded on photographic film (size $L \times L$) using the linear region of the H&D curve. A purely absorbing positive transparency with a photographic gamma of -2 was made, and that transparency is to be inserted in the input plane of the optical processing system. Specify the absorbing mask you would place in the frequency plane of the coherent optical processor in order to remove the interference. Consider especially:

1. Where should the absorbing spots be placed?
2. What size would you make the absorbing spots?
3. What would you do at frequency ($f_X=0, f_Y=0$) ($f_X = 0, f_Y = 0$)?

Note: Neglect any effect the mask might have on the nonperiodic signal that is also present at the input.

7. 10-7. A grating with amplitude transmittance $t_A(x,y)=121+\cos(2\pi f_o x)$

$t_A(x, y) = \frac{1}{2}[1 + \cos(2\pi f_o x)]$ is placed at the input to a standard “4f” coherent optical processing system of the kind illustrated in [Fig. 10.6\(a\)](#). Specify the transfer function (as a function of f_X) of a *pure phase* spatial filter that will completely suppress the spatial frequency component of output *intensity* having spatial frequency f_o . Assume normally incident plane wave illumination, monochromatic light, and neglect the effects of finite lens apertures.

8. 10-8. A transparent object with complex amplitude transmittance $t_A(x,y)$ is placed immediately in front of a positive spherical lens. The object is normally illuminated with a monochromatic plane wave, and a photographic transparency records the intensity distribution across the back focal plane. A positive transparency with a gamma of -2 is produced. The developed transparency is then illuminated by a plane wave, and the same positive lens is inserted directly behind the transparency. What is the relationship between the amplitude transmittance of the original object and the intensity distribution observed across the back focal plane of the lens in the second step of the process?

9. 10-9. A phase object with amplitude transmittance $t_A(x_1,y_1)=\exp[j\phi(x_1,y_1)]$

$t_A(x_1, y_1) = \exp[j\phi(x_1, y_1)]$ is present in the object plane of a coherent imaging system. In the back focal plane of the system, an attenuating plate (of uniform thickness) with intensity transmittance

$$\tau(x_2,y_2)=\alpha(x_2^4+2x_2^2y_2^2+y_2^4)$$

$$\tau(x_2, y_2) = \alpha(x_2^4 + 2x_2^2y_2^2 + y_2^4)$$

is introduced. How is the resulting image intensity related to the object phase?

10. 10-10. Consider the optical system shown in [Fig. P10.10](#). A transparency with a real and non-negative *amplitude* transmittance $s_1(\xi, \eta)$ is placed in plane P_1 and coherently illuminated by a monochromatic, unit-intensity, normally incident plane wave. Lenses L_1 and L_2 are spherical with common focal length f . In plane P_2 , which is the rear

focal plane of L_1 , a moving diffuser is placed. The effect of the moving diffuser can be considered to be the conversion of spatially coherent incident light into spatially incoherent transmitted light, without changing the intensity distribution of the light in plane P_2 and without appreciably broadening the spectrum of the light. In plane P_3 , in contact with L_2 , is placed a second transparency, this one with amplitude transmittance $s_2(x,y)$. Find an expression for the intensity distribution incident on plane P_4 .

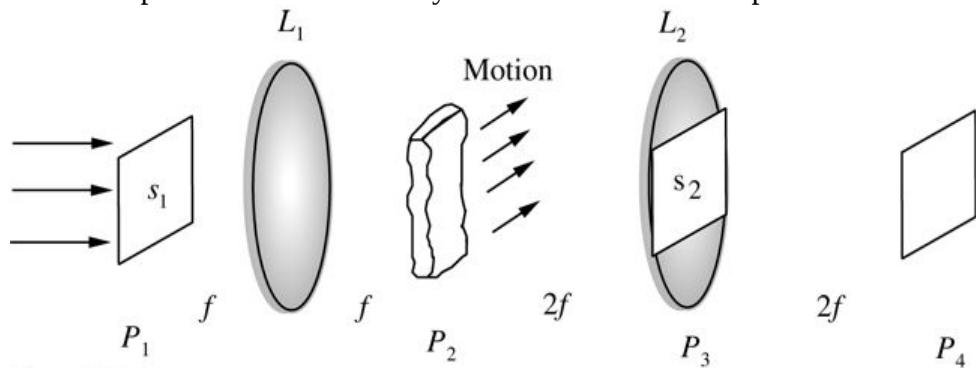


Figure P10.10

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure P10.10

The illustration shows a Plane P_1 , Lens L_1 , Plane P_2 , Lens L_3 , and Plane P_4 arranged from left to right. Four arrows from the left point toward P_1 , which is represented as a sloping parallelogram with the text S_1 inside. The distance between P_1 and L_1 is f . Plane P_2 is at a distance f from L_1 . P_2 is represented as vertical rectangle with curved edges and four arrows from P_2 point toward upper right direction. The text “motion” is shown above P_2 . Lens L_2 and Plane P_3 are adjoining. L_3 is at a distance $2f$ from P_2 . P_3 is represented as a sloping parallelogram with the text S_2 inside. P_4 is at a distance $2f$ from L_3 .

11. 10-11. The VanderLugt method is used to synthesize a frequency-plane filter. As shown in Fig. P10.11(a), a “signal” transparency with amplitude transmittance $s(x,y)$ is placed immediately against a positive lens (rather than in the front focal plane) and a photographic plate records the intensity in the back focal plane. The amplitude transmittance of the developed plate is made proportional to exposure, and the resulting transparency is placed in the system of part (b) of the figure. Assuming that the appropriate portions of the output plane are examined in each case, what should the distance d between the object plane and the first lens of the filtering system be in order to synthesize:

1. a filter with impulse response $s(x,y)$?
2. a filter with impulse response $s^*(-x, -y)$?

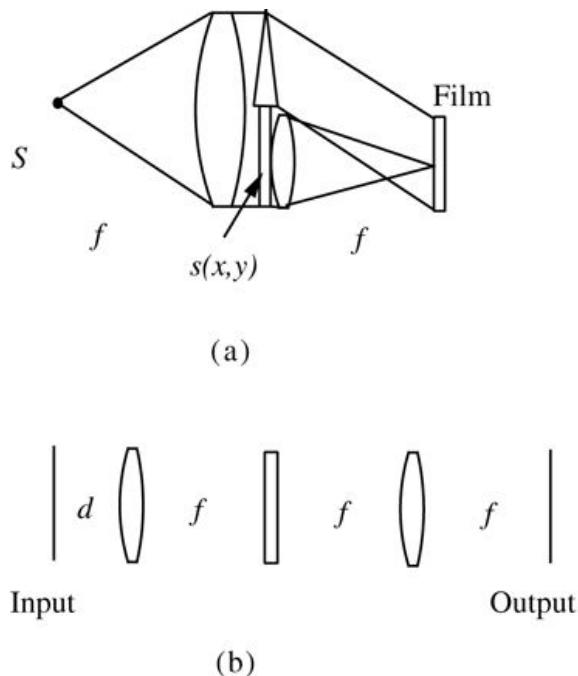


Figure P10.11
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure P10.11

Illustration a shows a dark spot S at the extreme left. To its right is a lens. Two lines emerge from S and point toward the upper and lower ends of the lens. To the right of the lens, is a thin vertical rectangular strip labeled $s(x, y)$ above which a prism is placed. To the right of the thin vertical rectangular strip, is a lens whose size is same as that of the rectangular strip. To its right is another thin vertical rectangular strip labeled film. A line from the top end of the first lens point toward the top end of the prism and a line from the top end of the prism points toward the top end of film. A line from the bottom end of the second lens point toward the bottom end of the thin vertical rectangular strip and a line from the bottom end of the thin vertical rectangular strip points toward the bottom end of film. Two lines from the top and bottom ends of the second lens point toward the center of film. The distance between source and first lens and between second lens and the film is each f .

Illustration b shows a vertical line labeled input on the extreme left. On its right, at a distance d is a lens. To the right of the lens, is a thin vertical rectangular strip at a distance f from lens. To the right of the thin vertical rectangular strip, is a lens which is at a distance f from the strip. To the right of the lens, is a vertical line labeled output which is at a distance f from the lens.

12. 10-12. Given a standard VanderLugt matched filtering system, prove that the output correlation spot shifts with any shift of the input signal against which we are correlating the reference. Assume that the magnification of the system is unity.
13. 10-13. Prove that the inequality of (10-20) must be satisfied if the various output terms of the joint transform correlator are to be separated.

14. 10-14. A certain image is blurred by camera motion such that the image incident on the recording film has moved linearly with velocity V on the film during a T -second exposure time.

1. Specify the point-spread function and the optical transfer function of the blur.
2. Specify and plot the magnitude of the transfer function of an inverse filter that will in principle remove the blur.
3. Assuming a constant ratio of signal power spectrum to noise power spectrum of 10, specify and plot the transfer function of a Wiener filter that will serve as a better deblurring filter than the simple inverse filter.
4. If you have access to a computer, calculate and plot the impulse response of the Wiener filter of part (c).

15. 10-15. Consider an ideal “perfect” periodic transmitting object with amplitude transmittance $p(x, y)$ having period L in both the x and y directions. On this object there exists an opaque defect, with a size much smaller than the period of the object, but nonetheless much larger than the smallest structure contained in $p(x, y)$. We wish to create an optical filter that will enhance the brightness of the defect with respect to the periodic object, thus enhancing our ability to detect the defect. Ideally we would like to completely suppress the periodic portion of the image and pass only a bright image of the defect.

1. Describe how you might make a spatial filter that would accomplish the task described above. Be as specific as possible.
2. Suppose that your filter were able to completely eliminate the discrete frequency components associated with the ideal periodic object, but also pass essentially all the light caused to leave these locations by the defect. Find an approximate expression for the image intensity that would be obtained at the output. As an aid to your analysis, let the amplitude transmittance of the defect be described by $1 - d(x, y)$, where the function $d(x, y)$ is unity within the defect and zero outside it. Remember that the defect and the periodic object should be treated as two separate diffracting structures in close contact with one another. You may neglect the finite sizes of the lenses and any vignetting effects.

16. 10-16. You are to construct a coherent optical “grade change” filter that will change the letter F into the letter A. The filtering system is of the standard “ $4f$ ” type. Describe in detail how you would construct such a filter. Be specific. How would you expose the photographic plate in making the filter? What behavior of the photographic transparency would you try to achieve? Where would you look in the output plane of the processor? Give as many details as you can.

17. 10-17. With reference to (9-32), show that the optical frequency of the light at coordinate y_2 in the spatial frequency domain of the Bragg cell spectrum analyzer is offset from the optical frequency of the source by an amount that is exactly equal to the temporal frequency of the RF spectral component represented at that coordinate.

11 Holography

In 1948, [Dennis Gabor \[120\]](#) proposed a novel two-step, lensless imaging process which he called *wavefront reconstruction* and which we now know as *holography*. Gabor recognized that when a suitable coherent reference wave is present simultaneously with the light diffracted by or scattered from an object, then information about both the amplitude and phase of the diffracted or scattered waves can be recorded, in spite of the fact that recording media respond only to light intensity. He demonstrated that, from such a recorded interference pattern (which he called a *hologram*, meaning a “total recording”), an image of the original object can ultimately be obtained.

While Gabor’s imaging technique received only mild interest in its early days, the 1960s saw dramatic improvements in both the concept and the technology, improvements that vastly extended its applicability and practicality. In 1971 Gabor received the Nobel prize in physics for his invention.

In this chapter we examine the basic principles behind holography, explore the many modern variations upon Gabor’s original theme, and survey some of the important applications that have been found for this novel imaging technique. Several excellent books devoted to holography exist. The classic text is that of [Collier, Burckhardt and Lin \[75\]](#). For another excellent and authoritative treatment see the book by [Hariharan \[160\]](#). Other broad books include those by [Smith \[322\]](#), [Develis and Reynolds \[90\]](#), [Caulfield \[58\]](#), and [Saxby \[307\]](#). For modern treatments see [Benton and Bove \[24\]](#) and [Toal \[345\]](#).

11.1 Historical Introduction

Gabor was influenced in his early studies of holography by previous work of W.L. Bragg in X-ray crystallography (see, for example, [39]), but was primarily motivated by possible applications of his newfound technique to electron holography. Gabor followed his original proposal with two more lengthy papers ([121], [122]) published in 1949 and 1951, considering the possible application of holography to microscopy. While for practical reasons he was unable to realize his envisioned application, the improvements developed in the 1960s led to many applications that Gabor could not possibly have foreseen.

In the 1950s, a number of authors, including [G.L. Rogers \[298\]](#), [H.M.A. El-Sum \[103\]](#), and [A.W. Lohmann \[230\]](#), significantly extended the theory and understanding of holography. It was not, however, until the early 1960s that a revolution in holography began. It was workers at the University of Michigan's Radar Laboratory, in particular [E.N. Leith and J. Upatnieks \[222\]](#), who recognized the similarity of Gabor's lensless imaging process to the synthetic-aperture-radar problem and suggested a modification of his original technique that greatly improved the process. At virtually the same time, [Y.N. Denisyuk \[89\]](#), working in what was then the Soviet Union, created a remarkable synthesis of the ideas of both Gabor and French physicist G. Lippmann to invent the thick reflection hologram, which he perfected to an advanced state.

The Michigan workers soon coupled their new developments with the emerging technology of lasers in order to perform lensless three-dimensional photography [224]. The quality and realism of the three-dimensional images obtained by holography were largely responsible for the development of a great popular interest in the field. Today it is common to find museums or galleries specializing in holography in several of the great cities of the world. However, contrary to popular impression, many of the most interesting and useful properties of holography are quite independent and separate from the three-dimensional imaging capability, as we shall see in some detail in later sections.

11.2 The Wavefront Reconstruction Problem

The fundamental problem addressed by holography is that of recording, and later reconstructing, both the amplitude and the phase of an optical wave arriving from a coherently illuminated object. This problem is sufficiently general to be of interest for electromagnetic waves in all regions of the spectrum, as well as for acoustic and seismic waves. Our considerations here, however, will be largely restricted to the optical problem.

11.2.1 Recording Amplitude and Phase

As indicated above, the wavefront-reconstruction problem must consist of two distinct operations: a recording or detection step, and a reconstruction step. For the moment we focus on the first of these two operations.

Since the wavefronts of concern are coherent, it is necessary to detect information about both the amplitude and phase of the waves. However, all recording media respond only to light intensity. It is therefore required that the phase information somehow be converted to intensity variations for recording purposes. A standard technique for accomplishing this task is *interferometry*; that is, a second wavefront, mutually coherent with the first and of known amplitude and phase, is added to the unknown wavefront, as shown in [Fig. 11.1](#). The intensity of the sum of two complex fields then depends on both the amplitude and phase of the unknown field. Thus if

$$a(x,y) = |a(x,y)| \exp[j\phi(x,y)]$$

$$a(x, y) = |a(x, y)| \exp [j\phi(x, y)]$$

(11-1)

represents the wavefront to be detected and reconstructed, and if

$$A(x,y) = |A(x,y)| \exp[j\psi(x,y)]$$

$$A(x, y) = |A(x, y)| \exp [j\psi(x, y)]$$

(11-2)

represents the “reference” wave with which $a(x,y)$ $a(x, y)$ interferes, the intensity of the sum is given by

$$\mathcal{I}(x,y) = |A(x,y)|^2 + |a(x,y)|^2 + 2|A(x,y)||a(x,y)| \cos[\phi(x,y) - \psi(x,y)].$$

$$\mathcal{I}(x, y) = |A(x, y)|^2 + |a(x, y)|^2 + 2|A(x, y)| |a(x, y)| \cos[\phi(x, y) - \psi(x, y)].$$

(11-3)

While the first two terms of this expression depend only on the intensities of the individual waves, the third depends on their relative phases. Thus information about both the amplitude and phase of

$a(x, y)$ has been recorded. The issue as to whether it is sufficient information to reconstruct the original wavefront remains to be dealt with. At this point we have not specified any detailed character of the reference wave $A(x, y)$. Properties that the reference wave must satisfy in order to enable reconstruction of $a(x, y)$ will become evident as the discussion progresses. The recording of the pattern of interference between an “object” wave and a “reference” wave may be regarded as a hologram.

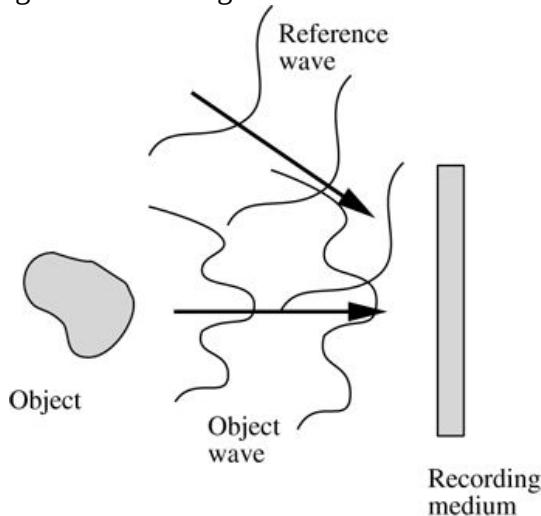


Figure 11.1
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.1 Interferometric recording.

The illustration shows an irregular shaped object on the left extreme from which complex object waves move toward the recording medium. The recording medium is a thin vertical rectangular strip. A horizontal arrow from the object points toward the recording medium. Two curved reference waves from upper left direction move toward the recording medium. An arrow from upper left direction points toward the recording medium.

11.2.2 The Recording Medium

The material used to record the pattern of interference described above will be assumed to provide a linear mapping of intensity incident during the detection process into amplitude transmitted by or reflected from the material during the reconstruction process. Usually both light detection and wavefront modulation are performed by photographic film or plate. The linear relation required is then provided by operation in the linear portion of the t_A versus E curve of the emulsion. However, many other materials suitable for holography exist, including photopolymers, dichromated gelatin, photorefractive materials, and others (see [Section 11.8](#)). It is even possible to detect the interference pattern electronically and reconstruct the wavefront with a digital computer. However, photographic materials remain the most important and widely used recording medium in holography.

Thus we assume that the variations of exposure in the interference pattern remain within a linear region of the t_A versus E curve. In addition, it is assumed that the MTF of the

recording material extends to sufficiently high spatial frequencies to record all the incident spatial structure (effects of removing some of these ideal assumptions are examined in [Section 11.10](#)).

Finally we assume that the intensity $|A|^2 |A|^2$ of the reference wave is uniform across the recording material, in which case the amplitude transmittance of the developed film or plate can be written

$$\begin{aligned} t_A(x,y) &= tb + \beta' |a|^2 + A^* a + A a^*, \\ t_A(x, y) &= t_b + \beta' (|a|^2 + A^* a + A a^*), \end{aligned} \quad (11-4)$$

where t_b is a uniform “bias” transmittance established by the constant reference exposure, and β' is the product of the slope β of the t_A versus E curve at the bias point and the exposure time. Note that, as in [Section 9.1](#), β' is a negative number for a negative transparency, and a positive number for a positive transparency.

11.2.3 Reconstruction of the Original Wavefront

Once the amplitude and phase information about the object wave $a(x,y)$ have been recorded, it remains to reconstruct that wave. Suppose that the developed transparency is illuminated by a coherent *reconstruction* wave $B(x,y)$. The light transmitted by the transparency is evidently

$$\begin{aligned} B(x,y) t_A(x,y) &= tbB + \beta' aa^* B + \beta' A^* Ba + \beta' ABa^* = U_1 + U_2 + U_3 + U_4. \\ B(x, y) t_A(x, y) &= t_b B + \beta' aa^* B + \beta' A^* Ba + \beta' ABa^* \\ &= U_1 + U_2 + U_3 + U_4. \end{aligned} \quad (11-5)$$

Note that if B is simply an exact duplication of the original uniform reference wavefront A , the third term of this equation becomes

$$U_3(x,y) = \beta' |A|^2 a(x,y).$$

$$U_3(x, y) = \beta' |A|^2 a(x, y).$$

(11-6)

Since the intensity of the reference wave is uniform, it is clear that reconstructed wave component U_3 is, up to a multiplicative constant, an exact duplication of the original wavefront $a(x,y)$, as shown in [Fig. 11.2\(a\)](#).

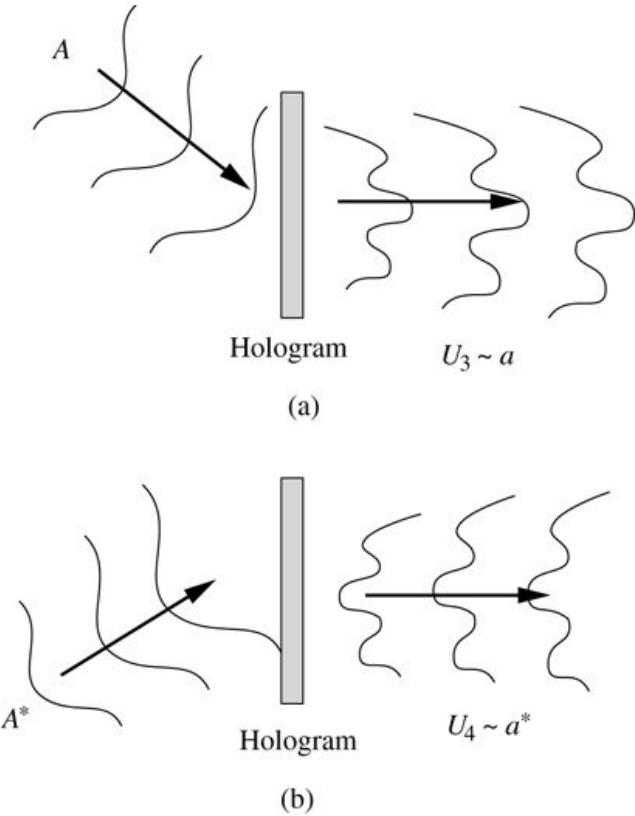


Figure 11.2
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.2 Wavefront reconstruction with (a) the original reference wave A as illumination, and (b) the conjugate reference wave A^* as illumination.

Illustration a shows a thin vertical rectangular bar labeled hologram. Three waves curved toward right labeled A start from the upper left direction and move toward hologram. Three complex waves labeled “ U_3 approximately equal to a ” move away from hologram and a horizontal arrow from hologram points toward the right.

Illustration b shows a thin vertical rectangular bar labeled hologram. Three curved waves curved toward left labeled A^* start from the lower left direction and move toward hologram. Three complex waves labeled “ U_4 approximately equal to a^* ” move away from hologram and a horizontal arrow from hologram points toward the right.

In a similar fashion, if $B(x, y)$ happens to be chosen as the *conjugate* of the original reference wave, i.e. as $A^*(x, y)$, the fourth term of the reconstructed field becomes

$$U_4(x, y) = \beta' |A|^2 a^*(x, y),$$

$$U_4(x, y) = \beta' |A|^2 a^*(x, y),$$

(11-7)

which is proportional to the *conjugate* of the original wavefront. This case is illustrated in Fig. 11.2(b).

Note that in either case, the particular field component of interest (that is, U_3 when $B=A$ and U_4 when $B=A^*$) is accompanied by three additional field components, each of which may be regarded as extraneous interference. Evidently, if a usable duplication of the object wave $a(x,y)$ ($a^*(x,y)$) is to be obtained, some method for separating the various wave components of transmitted light is required.

11.2.4 Linearity of the Holographic Process

The characteristic behavior hypothesized for the recording material in (11-4) corresponds to a highly *nonlinear* mapping of fields incident during exposure into fields transmitted after development. It would therefore appear, at first glance, that linear systems concepts can play no role in the theory of holography. While the overall mapping introduced by the film is nonlinear, nonetheless the mapping of object field $a(x,y)$ into the transmitted field component $U_3(x,y)$ is entirely linear, as evidenced by the proportionality of (11-6). Similarly, the mapping of $a(x,y)$ into the transmitted field component $U_4(x,y)$, as represented by (11-7), is a linear one. Thus if the object field $a(x,y)$ is regarded as an input, and the transmitted field component $U_3(x,y)$ ($U_4(x,y)$) is regarded as an output, the system so defined is a linear one. The nonlinearity of the detection process manifests itself in the generation of several output terms, but there is no nonlinear distortion of the one term of interest, assuming that the exposure variations remain in the linear region of the t_A versus E curve.

11.2.5 Image Formation by Holography

To this point we have considered only the problem of reconstructing a wavefront which arrived at a recording medium from a coherently illuminated object. It requires but a small change in point of view to regard the wavefront reconstruction process as a means of *image formation*.

To adopt this point of view, note that the wave component $U_3(x,y)$ of (11-6), being simply a duplication of the original object wavefront $a(x,y)$, must appear to the observer to be diverging from the original object, in spite of the fact that the object has long since been removed. Thus when the reference wave $A(x,y)$ is used as the illumination during reconstruction, the transmitted wave component $U_3(x,y)$ may be regarded as generating a *virtual image* of the original object. This case is illustrated in Fig. 11.3(a) and (b) for the particular case of a simple point-source object.

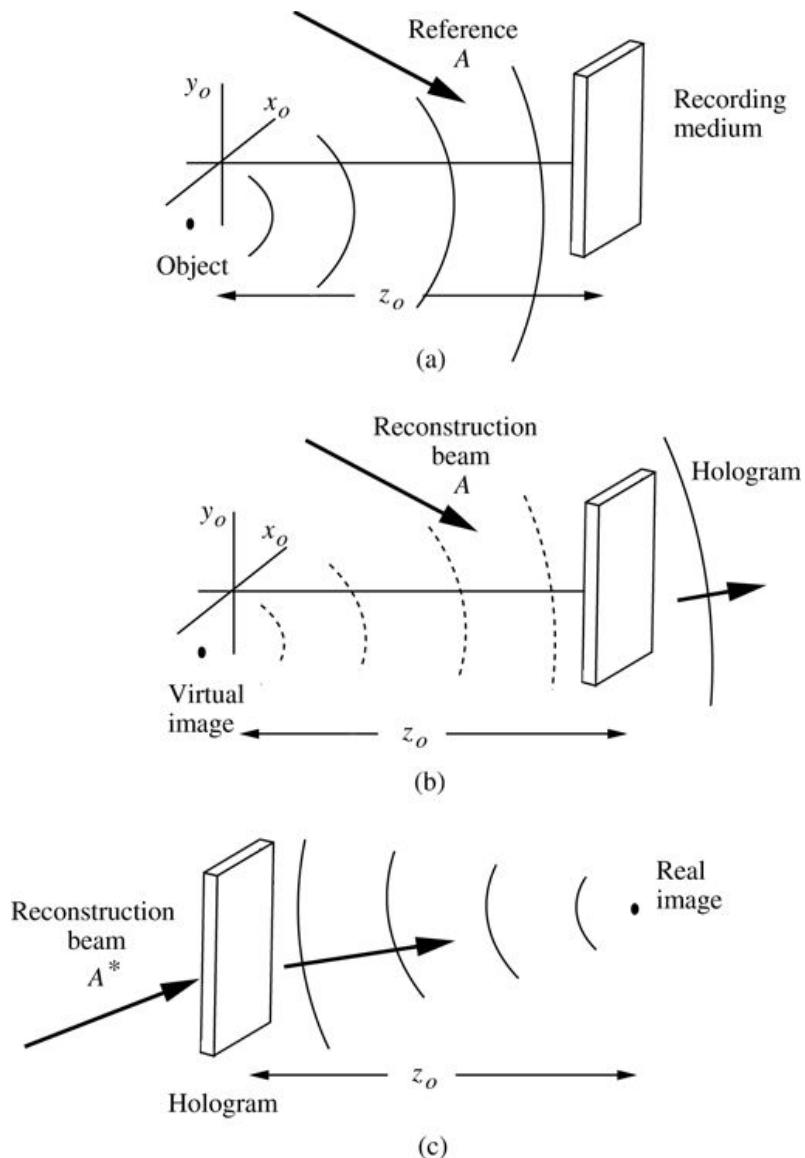


Figure 11.3
Goodman, Introduction to Fourier Optics, 4e,
 © 2017 W. H. Freeman and Company

Figure 11.3 Imaging by wavefront reconstruction. (a) Recording the hologram of a point-source object; (b) generation of the virtual image; (c) generation of the real image.

Illustration a depicts a graphical representation with vertical axis labeled y_0 and a third axis labeled x_0 . At the right end of the horizontal axis, is the recording medium which is a vertical horizontal structure. The distance between the vertical axis and recording medium is z_0 . A dark spot labeled object is on the left of the vertical axis below the horizontal axis. Waves from the object move rightward toward the recording medium. An arrow labeled Reference A from upper right direction point toward the recording medium.

Illustration b depicts a graphical representation with vertical axis labeled y_0 and a third axis labeled x_0 . At the right end of the horizontal axis, is the hologram which is a vertical horizontal structure. The distance between the vertical axis and recording medium is z_0 . A dark spot labeled virtual image is on the left of the vertical axis below the horizontal axis. Dotted waves from the

object move rightward toward the hologram and a wave which is represented with a solid line is shown beyond the hologram and moves rightward. An arrow labeled Reconstruction beam A starts from upper right and point toward the hologram and an arrow emerging from hologram moves in the rightward direction.

Illustration c shows a dark spot at the extreme right and a vertical rectangular bar labeled hologram is at the extreme left. Waves from the real image move toward the hologram. The hologram is at a distance of z_0 from the real image. An arrow labeled “Reconstruction beam A asterisk” starts from the left and points toward hologram and an arrow from the hologram points toward the real image.

In a similar fashion, when the conjugate of the reference wave, $A^*(x, y)$, is used as the illumination during reconstruction, the wave component $U_4(x, y)$ of (11-7) also generates an image, but this time it is a *real image* which corresponds to an actual focusing of light in space. To prove this assertion, we invoke the linearity property discussed above, considering an object which consists of a single point source. The corresponding result for a more complicated object may then be found by linear superposition of point-source solutions.

Incident on the recording medium we have the sum of the reference wave $A(x, y)$ and a simple spherical object wave,

$$a(x, y) = a_o \exp[jkz_o] + (x - \hat{x}_o)^2 + (y - \hat{y}_o)^2$$

$$a(x, y) = a_o \exp \left[jk \sqrt{z_o^2 + (x - \hat{x}_o)^2 + (y - \hat{y}_o)^2} \right]$$

(11-8)

where (\hat{x}_o, \hat{y}_o) are the (x, y) coordinates of the object point, and z_o is its normal distance from the recording plane. Illuminating the developed hologram with a reconstruction wave $A^*(x, y)$, we obtain the transmitted wave component

$$U_4(x, y) = \beta' |A|^2 a^*(x, y) = \beta' |A|^2 a_o \exp[-jkz_o] + (x - \hat{x}_o)^2 + (y - \hat{y}_o)^2$$

$$\begin{aligned} U_4(x, y) &= \beta' |A|^2 a^*(x, y) \\ &= \beta' |A|^2 a_o^* \exp \left[-jk \sqrt{z_o^2 + (x - \hat{x}_o)^2 + (y - \hat{y}_o)^2} \right], \end{aligned}$$

(11-9)

which is a spherical wave that *converges* towards a real focus at distance z_o to the right of the hologram, as shown in [Fig. 11.3\(c\)](#). A more complicated object may be considered to be a multitude of point sources of various amplitudes and phases, and by the linearity property, each such point source generates its own real image as above. Thus a real image of the entire object is formed in this fashion.

Note that the amplitude of the wave described by (11-9) is proportional to a_o^* , the conjugate of the original object point-source amplitude. Similarly, for a more complicated object, the real image generated by the hologram is always the complex conjugate of the original object amplitude. Such a change of phase does not affect image intensity, but it can be important in certain applications that utilize both the amplitude and phase of the image. In addition, for three-dimensional objects, it can have an unusual effect that we will elaborate on later.

It should again be emphasized that we have considered, in each case, only one of the four wave components transmitted by the hologram. This approach is acceptable if, by proper choice of reference wave, the undesired components are suppressed or are separated from the image of interest. When this is not the case, the interference of the various components of transmitted light must be taken into account.

11.3 The Gabor Hologram

Keeping in mind the preceding general discussion, we now consider the wavefront-reconstruction process in the form originally proposed and demonstrated by Gabor. In [Section 11.4](#), we turn to modifications of the process which improve its imaging capabilities.

11.3.1 Origin of the Reference Wave

The geometry required for recording a *Gabor hologram* is illustrated in [Fig. 11.4](#). The object is assumed to be highly transmissive, with an amplitude transmittance

$$t(x_0, y_0) = t_o + \Delta t(x_0, y_0),$$

$$t(x_o, y_o) = t_o + \Delta t(x_o, y_o),$$

(11-10)

where t_o is a high average level of transmittance, Δt represents the variations about this average, and

$$|\Delta t| \ll |t_o|.$$

$$|\Delta t| \ll |t_o|.$$

(11-11)

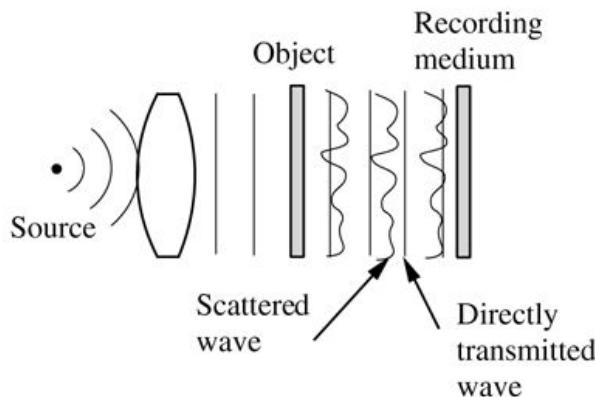


Figure 11.14

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.4 Recording a Gabor hologram.

The illustration shows a dark spot S on the left extreme from which waves move rightward to the convex curvature of the lens. Next to the lens are four straight vertical lines followed by a thin vertical rectangular strip labeled output. Next to it, are four vertical lines labeled directly transmitted wave and in between the lines are three irregularly curved vertical lines labeled scattered wave. Next to the waves is a vertical rectangular bar labeled recording medium.

When such an object is coherently illuminated by the collimated wave shown in Fig. 11.4, the transmitted light consists of two components: (1) a strong uniform plane wave passed by the term to t_o , and (2) a weak scattered wave generated by the transmittance variations $\Delta t(x_0, y_0)$ $\Delta t(x_o, y_o)$. The intensity of the light incident on the recording medium at distance z_0 from the object may be written

$$J(x, y) = A + a(x, y) |^2 = |A|^2 + |a(x, y)|^2 + A^* a(x, y) + A a^*(x, y),$$

$$\begin{aligned} \mathcal{I}(x, y) &= |A + a(x, y)|^2 \\ &= |A|^2 + |a(x, y)|^2 + A^* a(x, y) + A a^*(x, y), \end{aligned}$$

(11-12)

where A is the amplitude of the plane wave, and $a(x, y)$ is the amplitude of the scattered light at the recording plane.

Thus the object has, in a sense, supplied the required reference wave itself through the high average transmittance to t_o . The interference of the directly transmitted light with the scattered light results in a pattern of intensity that depends on both the amplitude and the phase of the scattered wave $a(x, y)$.

11.3.2 The Twin Images

The developed hologram is assumed to have an amplitude transmittance that is proportional to exposure. Thus

$$\begin{aligned} t_A(x, y) &= t_b + \beta' |a|^2 + A^* a + A a^*. \\ t_A(x, y) &= t_b + \beta' (|a|^2 + A^* a + A a^*). \end{aligned}$$

(11-13)

If the transparency is now illuminated by a normally incident plane wave with uniform amplitude B , the resulting transmitted field amplitude consists of a sum of four terms:

$$\begin{aligned} B t_A &= B t_b + \beta' B |a(x, y)|^2 + \beta' A^* B a(x, y) + \beta' A B a^*(x, y). \\ B t_A &= B t_b + \beta' B |a(x, y)|^2 + \beta' A^* B a(x, y) + \beta' A B a^*(x, y). \end{aligned}$$

(11-14)

The first term is a plane wave which passes directly through the transparency, suffering uniform attenuation but without scattering. The second term may be dropped as negligible by virtue of our assumption (11-11), which implies that

$$|a(x, y)| \ll A.$$

$$|a(x, y)| \ll A.$$

(11-15)

The third term represents a field component that is proportional to the original scattered wave $a(x, y) a^*(x, y)$. This wave appears to originate from a virtual image of the original object located at distance z_o from the transparency, as shown in Fig. 11.5. Similarly, the fourth term is proportional to $a^*(x, y) a^*(x, y)$ and, in accord with our earlier discussions, leads to the formation of a real image at distance z_o on the opposite side of the transparency from the virtual image (again, see Fig. 11.5).

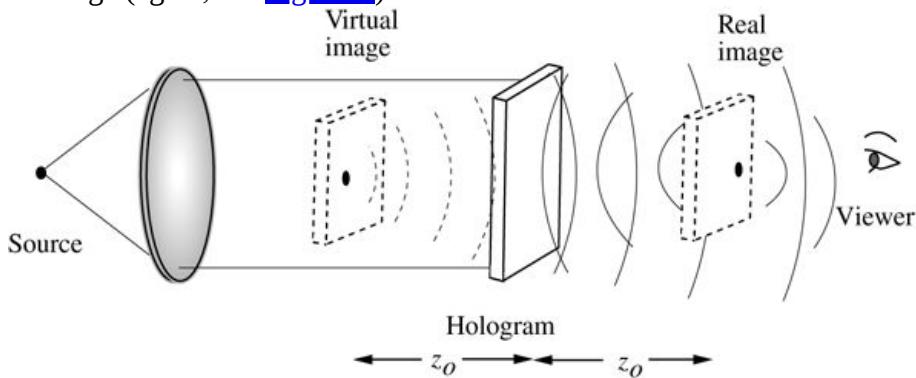


Figure 11.5

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 11.5 Formation of twin images from a Gabor hologram.

The illustration shows a dark spot labeled source on the left extreme. To its right, is a lens. Two lines from source point toward the upper and lower ends of the lens. To its right, is vertical rectangular bar labeled hologram. Two lines from the upper and lower ends of the lens point toward the upper and lower ends of hologram, respectively. A dotted vertical rectangular bar labeled virtual image is shown between the lens and hologram. A dark spot is at the center of the dotted vertical rectangular bar. Dotted waves emerge from virtual image and move toward the hologram. To the right of hologram, is another dotted vertical rectangular bar labeled real image. Waves emerge from virtual image and move toward the hologram. To the right of the real image is a human eye labeled viewer. Waves from real image move toward the viewer. The distance between hologram and virtual image and between hologram and real image are each f .

Thus the Gabor hologram generates simultaneous real and virtual images of the object transmittance variations Δt , both images being centered on the hologram axis. These so-called *twin images* are separated by the axial distance $2z_o$, and are accompanied by a coherent background Bt_b .

Note from (11-14) that positive and negative transparencies yield different signs for the image-forming waves with respect to the background (β' is positive for a positive transparency and negative for a negative transparency). In addition, for any one of these two cases, the real image wave is the conjugate of the virtual image wave, and depending on the phase structure of the object, further contrast reversals are possible when one of these waves interferes with the constant background. For an object with constant phase, a positive hologram transparency is found to produce a positive image, and a negative hologram transparency is found to produce a negative image.

11.3.3 Limitations of the Gabor Hologram

The Gabor hologram is found to suffer from certain limitations which restrict the extent of its applicability. Perhaps the most important limitation is inherent in the assumption of a highly transparent object and the consequent conclusion ([11-15](#)) that followed. If this assumption is not adopted, there exists an additional wave component

$$\begin{aligned} U_2(x,y) &= \beta' B |a(x,y)|^2 \\ U_2(x, y) &= \beta' B |a(x, y)|^2 \end{aligned}$$

(11-16)

transmitted by the hologram which can no longer be dropped as negligible. In fact, if the object is of low average transmittance, this particular wave component may be the largest transmitted term and as a consequence may entirely obliterate the weaker images. Thus with a Gabor hologram it is possible to image an object consisting of, for example, opaque letters on a transparent background, but not transparent letters on an opaque background. This restriction seriously hampers the use of Gabor holograms in many potential applications.

A second serious limitation lies in the generation of overlapping twin images, rather than a single image. The problem lies not with the presence of twin images per se, but rather with their inseparability. When the real image is brought to focus, it is always accompanied by an out-of-focus virtual image. Likewise an observer viewing the virtual image sees simultaneously a defocused image arising from the real-image term. Thus, even for highly transparent objects, the quality of the images is reduced by the twin image problem. A number of methods have been proposed for eliminating or reducing the twin-image problem, (e.g. see [\[230\]](#)), including one technique originated by [Gabor himself \[124\]](#). The most successful of these methods has been that of [Leith and Upatnieks \[222\]](#), which we discuss in detail in the next section.

11.4 The Leith-Upatnieks Hologram

Leith and Upatnieks suggested and demonstrated a modification of Gabor's original recording geometry that solved the twin image problem and vastly extended the applicability of holography. This type of hologram will be called the *Leith-Upatnieks* hologram, and is also known as an *offset-reference* hologram. The major change between this type of hologram and the Gabor hologram is that, rather than depending on the light directly transmitted by the object to serve as a reference wave, a separate and distinct reference wave is introduced. Furthermore the reference is introduced at an offset angle, rather than being collinear with the object-film axis.

The first successful demonstration of this type of hologram, reported in [222], was carried out without a laser source. However, it was not until the technique was combined with highly coherent laser illumination that its full potential became evident [224], [223].

11.4.1 Recording the Hologram

One possible geometry for recording a Leith-Upatnieks hologram is illustrated in Fig. 11.6. The light from a point source of illumination is collimated by the lens L . A portion of the resulting plane wave strikes the object, which is taken to be a transparency with a general amplitude transmittance $t(x_o, y_o)$. A second portion of the plane wave strikes a prism P located above the object and is deflected downwards at angle 2Θ with respect to the normal to the recording plane.¹ Thus at the recording surface we find the sum of two mutually coherent waves, one consisting of light transmitted by the object, and the other consisting of a tilted plane wave. The amplitude distribution incident on the recording plane may be written

$$\begin{aligned} U(x, y) &= A \exp(-j2\pi\alpha y) + a(x, y), \\ U(x, y) &= A \exp(-j2\pi\alpha y) + a(x, y), \end{aligned} \tag{11-17}$$

where the spatial frequency α of the reference wave is given by

$$\alpha = \sin 2\Theta \lambda.$$

$$\alpha = \frac{\sin 2\Theta}{\lambda}.$$

$$(11-18)$$

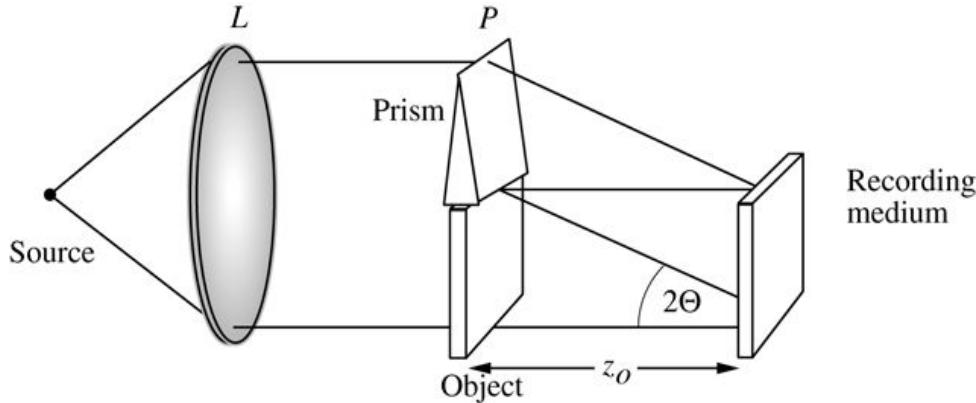


Figure 11.6

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 11.6 Recording a Leith-Upatnieks hologram.

The illustration shows a dark spot labeled source on the extreme left and right to it is a Lens L. Two lines from the source point toward the upper and lower end of Lens L. To its right, is a square shaped bar labeled object. Above the rectangle is a Prism P. To the right of the object, is another square bar labeled recording medium. Two lines from the upper and lower ends of L point toward the upper center of Prism P and the bottom center of object, respectively. A line from the top center of Prism P points toward the top center of recording medium. A line from the bottom center of object points toward the bottom center of the recording medium. A line at an angle of 20 degrees from the horizontal line starts from a point slightly above the lower end of the recording medium and points toward the top center of the object. The distance between object and recording medium is z_0 .

The intensity distribution across the recording plane is evidently

$$I(x, y) = |A|^2 + |a(x, y)|^2 + A * a(x, y) \exp(j2\pi\alpha y) + A * a(x, y) \exp(-j2\pi\alpha y).$$

$$\begin{aligned} \mathcal{I}(x, y) &= |A|^2 + |a(x, y)|^2 \\ &+ A^* a(x, y) \exp(j2\pi\alpha y) + A a^*(x, y) \exp(-j2\pi\alpha y). \end{aligned} \quad (11-19)$$

An alternative more revealing form may be obtained by writing $a(x, y)$ explicitly as an amplitude and phase distribution,

$$a(x, y) = |a(x, y)| \exp[j\phi(x, y)]$$

$$a(x, y) = |a(x, y)| \exp[j\phi(x, y)]$$

$$(11-20)$$

and combining the last two terms of (11-19) to yield

$$\mathcal{I}(x, y) = |A|^2 + |a(x, y)|^2 + 2|A||a(x, y)| \cos[2\pi\alpha y + \phi(x, y)].$$

$$\mathcal{I}(x, y) = |A|^2 + |a(x, y)|^2 + 2|A||a(x, y)| \cos[2\pi\alpha y + \phi(x, y)].$$

(11-21)

This expression demonstrates that the amplitude and phase of the light arriving from the object have been recorded, respectively, as amplitude and phase modulations of a spatial carrier of frequency α . If the carrier frequency is sufficiently high (we shall see shortly just how high it must be), the amplitude and phase distributions can be unambiguously recovered from this pattern of interference.

11.4.2 Obtaining the Reconstructed Images

In the usual fashion, the photographic plate is developed to yield a transparency with an amplitude transmittance proportional to exposure. Thus the film transmittance may be written

$$tA(x,y)=tb+\beta'|a(x,y)|^2+A^*a(x,y)\exp(j2\pi\alpha y)+Aa^*(x,y)\exp(-j2\pi\alpha y).$$

$$t_A(x, y) = t_b + \beta \left[|a(x, y)|^2 + A^* a(x, y) \exp(j2\pi\alpha y) + A a^*(x, y) \exp(-j2\pi\alpha y) \right].$$

(11-22)

For convenience we represent the four terms of transmittance by

$$t1=tbt3=\beta'A^*a(x,y)\exp(j2\pi\alpha y)t2=\beta'|a(x,y)|^2t4=\beta'Aa^*(x,y)\exp(-j2\pi\alpha y).$$

$$\begin{aligned} t_1 &= t_b & t_3 &= \beta' A^* a(x, y) \exp(j2\pi\alpha y) \\ t_2 &= \beta' |a(x, y)|^2 & t_4 &= \beta' A a^*(x, y) \exp(-j2\pi\alpha y). \end{aligned}$$

(11-23)

For the present we assume that the hologram is illuminated by a normally incident, uniform plane wave of amplitude B , as illustrated in [Fig. 11.7](#). The field transmitted by the hologram has four distinct components, each generated by one of the transmittance terms of [\(11-23\)](#):

$$U1=tbBU3=\beta'BA^*a(x,y)\exp(j2\pi\alpha y)U2=\beta'B|a(x,y)|^2U4=\beta'BAa^*(x,y)\exp(-j2\pi\alpha y).$$

$$\begin{aligned} U_1 &= t_b B & U_3 &= \beta' B A^* a(x, y) \exp(j2\pi\alpha y) \\ U_2 &= \beta' B |a(x, y)|^2 & U_4 &= \beta' B A a^*(x, y) \exp(-j2\pi\alpha y). \end{aligned}$$

(11-24)

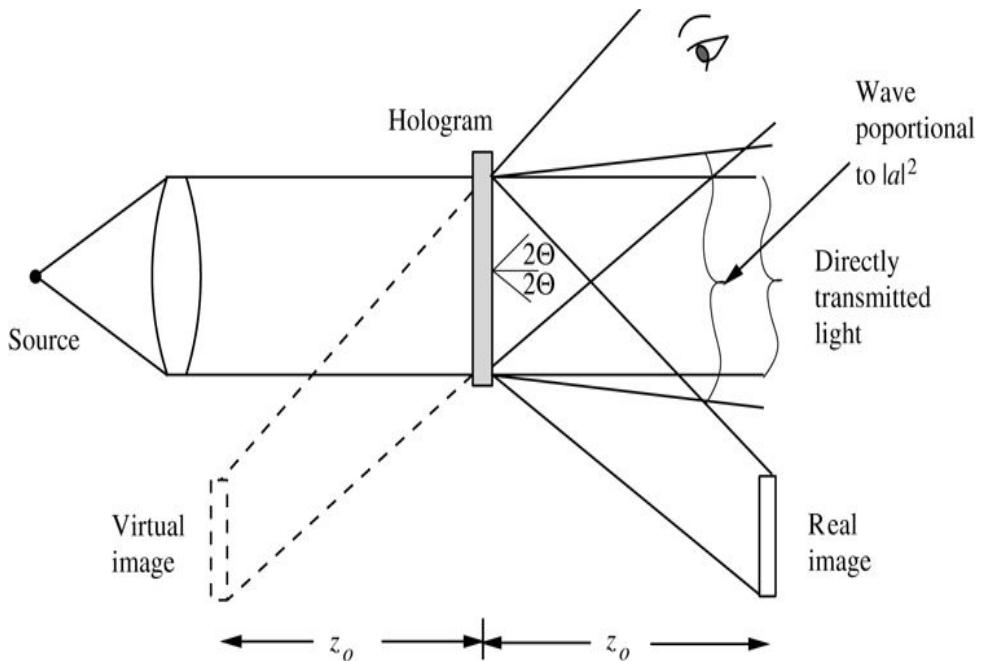


Figure 11.7

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 11.7 Reconstruction of images from a Leith-Upatnieks hologram.

The illustration shows a dark spot labeled source on the left extreme. To the right of it is a lens followed by a thin vertical rectangular bar. A line from the upper end of the lens points toward hologram at a point slightly below the upper end. A line from the lower end of the lens points toward hologram at a point slightly above the lower end. A straight horizontal line starts from a point slightly below the upper end of hologram and extends rightward to a distance of z_0 . A slanting horizontal line starts at an angle of 20 degrees from the same point on hologram and extends rightward to a distance of z_0 . A straight horizontal line starts from a point slightly above the lower end of hologram and extends rightward to a distance of z_0 . A slanting horizontal line starts at an angle of 20 degrees from the same point on hologram and extends rightward to a distance of z_0 . The distance between two horizontal lines is labeled "Directly transmitted Light." The distance between two slanting lines is labeled "Wave proportional to square of determinant of a." A small straight horizontal line starts from the center of the hologram and moves rightward. Two small slanting lines at an angle of 20 degrees start from the same point and extend rightward. A line emerges from the upper right direction and points toward a point slightly below the upper end of hologram. Another line emerges from the upper right direction and points toward a point slightly above the lower end of hologram. A human eye is shown at the upper right between the two lines. At the bottom towards the left is a thin vertical rectangular bar with a dotted outline labeled virtual image. A dotted line emerges from the upper end of virtual image and points toward a point slightly below the upper end of hologram. A dotted line emerges from the lower end of virtual image and points toward a point slightly above the lower end of hologram. At the bottom towards the right, is a thin vertical rectangular bar labeled real image. A line emerges from the upper end of virtual image and points toward a point slightly below the upper end of hologram. A line emerges from the lower end of virtual image and points toward a point slightly above the lower end of hologram.

The field component U_1 is simply an attenuated version of the incident reconstruction illumination, and therefore represents a plane wave traveling down the optical axis. The second term U_2 is spatially varying and therefore has plane wave components traveling at various angles with respect to the optical axis. However, as we shall see in more detail shortly, if the bandwidth of $a(x,y)$ is sufficiently small compared with the carrier frequency α , the energy in this wave component remains sufficiently close to the optical axis to be spatially separated from the images of interest.

The wave component U_3 is proportional to the original object wavefront a multiplied by a linear exponential factor. Proportionality to a implies that this term generates a virtual image of the object at distance z_o to the left of the transparency, while the linear exponential factor $\exp(j2\pi\alpha y)$ indicates that this image is deflected away from the optical axis at angle 2Θ , as shown in [Fig. 11.7](#). Similarly, wave component U_4 is proportional to the conjugate wavefront $a^* a^*$, which indicates that a real image forms at distance z_o to the right of the transparency. The presence of the linear exponential factor $\exp(-j2\pi\alpha y)$ indicates that the real image is deflected at angle -2Θ from the optical axis, as again can be seen in [Fig. 11.7](#).

The most important observation to be derived from these results is that, while twin images are again generated by the holographic process, they have been angularly separated from each other and from the wave components U_1 and U_2 . This separation comes about due to the use of a reference wave with an angular offset; indeed, successful isolation of each of the twin images requires the use of an angle between object and reference which is chosen larger than some lower limit (the minimum reference angle will be discussed in more detail shortly). When this angle exceeds the minimum allowable angle, the twin images are not contaminated by each other nor by other wave components.

Note in addition that since the images may be viewed without the presence of a coherent background generated by the object transparency, the particular sign associated with the wave components U_3 and U_4 of (11-24) is immaterial. The transparency may be either a positive or a negative; in each case a positive image is obtained. For practical reasons it is generally preferable to use negatives directly, thus avoiding the two-step process usually required for making a positive transparency.

Finally we should point out that we have chosen to illuminate the hologram with a normally incident plane wave, which is neither a duplication of the original reference wave nor its complex conjugate, yet we have obtained a real and a virtual image simultaneously. Evidently our conditions concerning the required nature of the reconstruction illumination were overly restrictive. However, when we consider the effects of the thickness of the emulsion on the reconstructed wavefronts, the exact nature of the reconstruction illumination will become more important. As will be discussed in [Section 11.7](#), it then becomes critical that the hologram be illuminated with a duplicate of the original reference wave to obtain one image, and the complex conjugate of the reference wave to obtain the other image.

11.4.3 The Minimum Reference Angle

Returning to the reconstruction geometry of Fig. 11.7, if the twin images are to be separated from each other and from the light transmitted with directions close to the optical axis, the offset angle 2Θ of the reference beam with respect to the object beam must be greater than some minimum angle $2\Theta_{min}$. To find this minimum, it suffices to determine the minimum carrier frequency α for which the spatial frequency spectra of t_3 and t_4 (which is the virtual-image and real-image terms of hologram transmittance) do not overlap each other and do not overlap the spectra of t_1 and t_2 . If there is no overlap, then in principle the hologram amplitude transmittance can be Fourier transformed with the help of a positive lens, the unwanted spectral components can be removed with appropriate stops in the focal plane, and a second Fourier transformation can be performed to yield just that portion of the transmitted light that leads to the twin images.²

Consider the spatial frequency spectra of the various terms of transmittance listed in Eq. (11-23). Neglecting the finite extent of the hologram aperture, we have directly that

$$\begin{aligned} G_1(f_X, f_Y) &= \mathcal{F}\{t_1(x, y)\} = t_b \delta(f_X, f_Y), \\ G_1(f_X, f_Y) &= \mathcal{F}\{t_1(x, y)\} = t_b \delta(f_X, f_Y). \end{aligned} \quad (11-25)$$

Using the autocorrelation theorem, we also have

$$\begin{aligned} G_2(f_X, f_Y) &= \mathcal{F}\{t_2(x, y)\} = \beta' G_a(f_X, f_Y) \star G_a(f_X, f_Y) \\ G_2(f_X, f_Y) &= \mathcal{F}\{t_2(x, y)\} = \beta' G_a(f_X, f_Y) \star G_a(f_X, f_Y) \end{aligned} \quad (11-26)$$

where $G_a(f_X, f_Y) = \mathcal{F}\{a(x, y)\}$ and the \star indicates the autocorrelation operation. Finally we have

$$\begin{aligned} G_3(f_X, f_Y) &= \mathcal{F}\{t_3(x, y)\} = \beta' A^* G_a(f_X, f_Y - \alpha) \\ G_3(f_X, f_Y) &= \mathcal{F}\{t_3(x, y)\} = \beta' A^* G_a(f_X, f_Y - \alpha) \\ G_4(f_X, f_Y) &= \mathcal{F}\{t_4(x, y)\} = \beta' A G_a^*(-f_X, -f_Y - \alpha). \end{aligned}$$

Now note that the bandwidth of G_a is identical with the bandwidth of the object, for the two spectra differ only by the transfer function of the propagation phenomenon, which (neglecting the evanescent wave cutoff) is the pure phase function of (3-74). Suppose that the object has no spatial frequency components higher than $B/2$ cycles/mm. Thus the spectrum $|G_a|$ might be as shown in Fig. 11.8(a). The corresponding spectrum of the hologram transmittance is illustrated in Fig. 11.8(b). The term $|G_1|$ is simply a δ function at the origin in the (f_X, f_Y) plane. The term $|G_2|$, being proportional to the autocorrelation function of $|G_a|$, extends to frequencies as high as $\pm B$. Finally, $|G_3|$ is simply proportional to $|G_a|$.

$|G_a|$, displaced to a center frequency $(0, \alpha)$ $(0, \alpha)$, while $|G_4| |G_4|$ is proportional to a reflected version of $|G_a| |G_a|$ centered at frequency $(0, -\alpha)$ $(0, -\alpha)$.

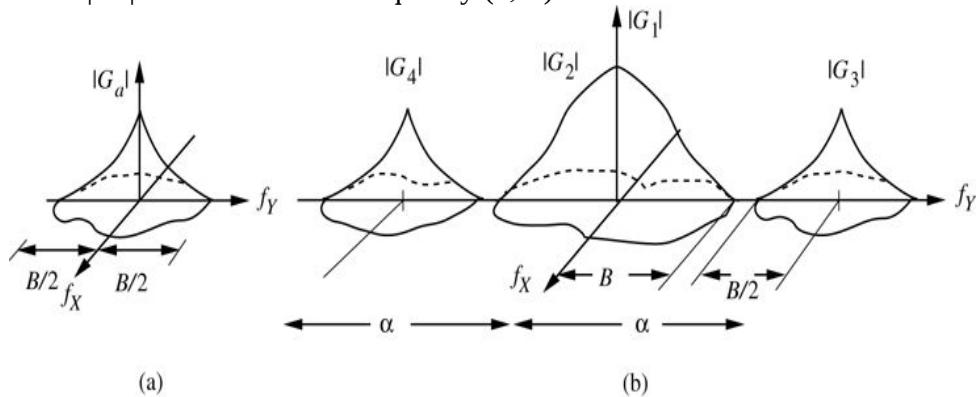


Figure 11.6

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 11.8 Spectra of (a) the object and (b) the hologram.

Illustration a shows the horizontal axis labeled f_Y and vertical axis labeled determinant of G_a . The curve is triangular shaped with the vertical axis running at the center and has an irregularly curved base below the horizontal axis. A slanting line labeled f_X from upper right moves toward the bottom left passing through the meeting point of horizontal and vertical axis. A dotted curve starts at the left end of the curve slightly above the horizontal axis and points toward the right end of the curve. The distance between left extreme and f_X and between f_X and the point where the dotted curve ends are $B/2$ each.

Illustration b shows the horizontal axis labeled f_Y and vertical axis labeled determinant of G_1 . Three triangular shaped curves are shown with the center one slightly big than the other two. The center curve is labeled determinant of G_2 with the vertical axis running at the center and has an irregularly curved base which is below the horizontal axis. A slanting line labeled f_X from a point at the right center on the curve moves toward the bottom left passing through the meeting point of horizontal and vertical axis. A dotted curve starts at the left end of the curve slightly above the horizontal axis and points toward the right end of the curve with a slight dip at a point slightly rightward to the center. The distance between f_X and the end of the curve is B .

The triangular curve on the extreme left is labeled determinant of G_3 and the base of the curve is irregular and is below the horizontal axis. The center point of the curve is marked on the horizontal axis with a small vertical line and a slanting line starts from the point and moves toward the bottom left direction. A dotted curve starts at the left end of the curve slightly above the horizontal axis and points toward the right end of the curve.

The triangular curve on the right is labeled determinant of G_4 and the base of the curve is irregular and is below the horizontal axis. The center point of the curve is marked on the horizontal axis and a slanting line starts from the point and moves toward the bottom left direction. The distance between the line and the end of curve is $B/2$. A dotted curve starts at the left end of the curve slightly above the horizontal axis and points toward the right end of the curve. The distance between the left end of the horizontal axis and f_X and the distance between f_X and the starting point of third curve are α each.

Examination of [Fig. 11.8\(b\)](#) shows that $|G_3|$ and $|G_4|$ can be isolated from $|G_2|$ if

$$\alpha \geq 3B/2$$

$$\alpha \geq 3B/2$$

(11-27)

or equivalently if

$$\sin 2\Theta \geq 3B\lambda/2.$$

$$\sin 2\Theta \geq 3B\lambda/2.$$

(11-28)

Evidently the minimum allowable reference angle is given by

$$2\Theta_{min} = \sin^{-1}(3B\lambda/2).$$

$$2\Theta_{min} = \sin^{-1}(3B\lambda/2).$$

(11-29)

When the reference wave is much stronger than the object wave, this requirement can be relaxed somewhat. The term G_2 is generated physically by interference of light from each object point with light from all other object points, while G_3 and G_4 arise from interference between the object and reference waves. When the object wave is much weaker than the reference wave (i.e. when $|a| \ll |A|$), the term G_2 is of much smaller magnitude than G_1 , G_3 , or G_4 and can be dropped as negligible. In this case the minimum reference angle is that which barely separates G_3 and G_4 from each other, or

$$2\Theta_{min} = \sin^{-1}(B\lambda/2).$$

$$2\Theta_{min} = \sin^{-1}(B\lambda/2).$$

(11-30)

11.4.4 Holography of Three-Dimensional Scenes

In 1964, Leith and Upatnieks reported the first successful extension of holography to three-dimensional imagery [\[224\]](#). Success in this endeavor rested to a large degree on the availability of the HeNe laser, with its excellent temporal and spatial coherence.

[Figure 11.9\(a\)](#) illustrates the general geometry used for recording holograms of three-dimensional scenes. Coherent light illuminates the scene of interest. In addition, a portion of the illumination strikes a “reference” mirror placed next to the scene. Light is reflected from the

mirror directly to the photographic plate, where it serves as a reference wave, interfering with light reflected from the scene itself. Thus the photographic plate records a hologram of the three-dimensional scene.

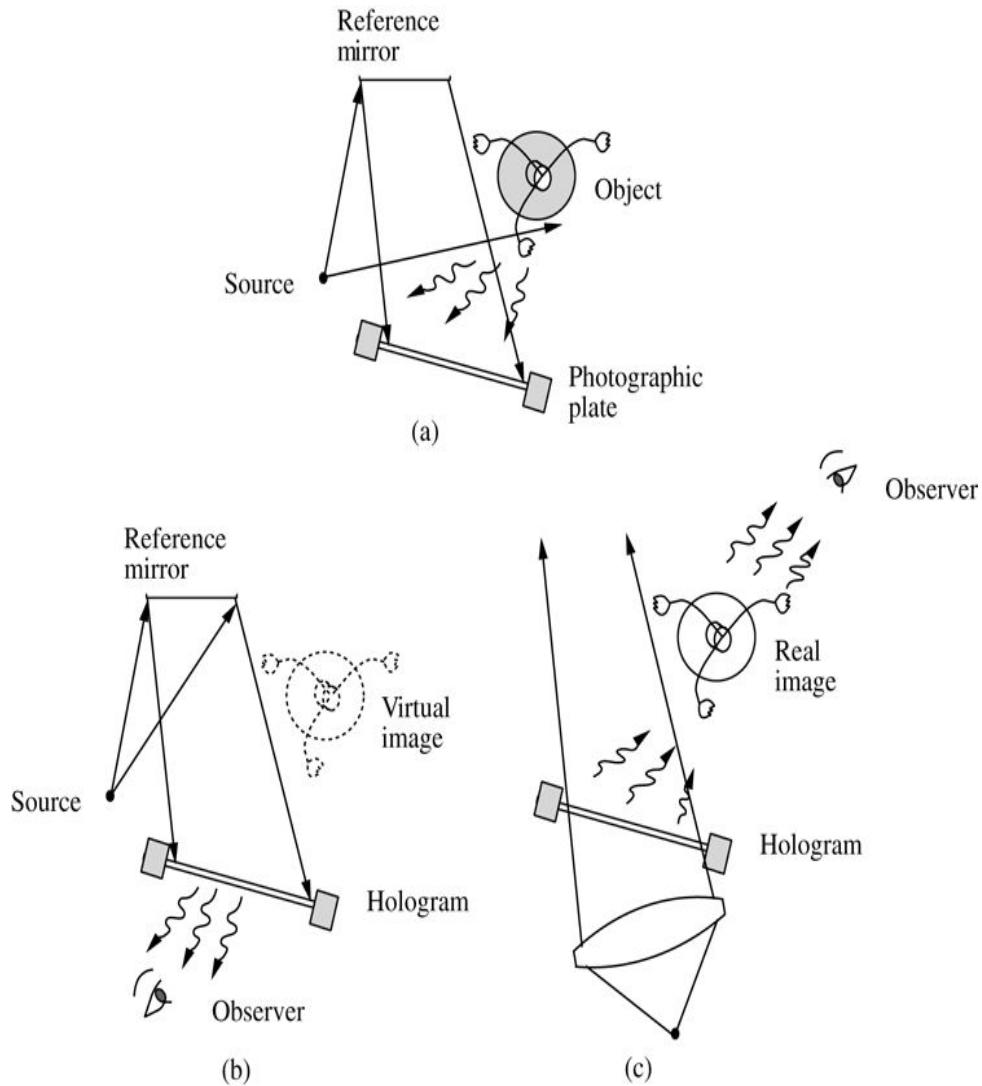


Figure 11.9
Goodman, Introduction to Fourier Optics, 4e, © 2017 W.H. Freeman and Company

Figure 11.9 Holographic imaging of a three-dimensional scene. (a) Recording the hologram; (b) reconstructing the virtual image; (c) reconstructing the real image.

Illustration a shows a dark spot labeled source and an arrow from source points toward the left end of the reference mirror which is a small horizontal line at the top. Another slanting line from source points toward the right. A circular structure labeled object is shown at the right with a hole at the center. Three thread structures with a flower like structure at the end of each structure emerge from the center and point in different directions. A slanting thin horizontal bar labeled photographic plate is placed below the source and a rectangle shaped structure is shown at both ends. Two lines from the upper and lower ends of reference mirror point toward the both ends of photographic plate. Three zigzag arrows from object move toward the photographic plate.

Illustration b shows a dark spot labeled source and an arrow from source point toward left end of reference mirror which is a small horizontal line at the top. Another slanting line from source points toward the right end of reference mirror. A slanting thin horizontal bar labeled hologram is placed below the source and a rectangle shaped structure is shown at both ends of the bar. Two lines from the two ends of hologram point toward the upper and lower ends of photographic plate. Three zigzag arrows from hologram move toward the human eye labeled observer shown at the bottom. On the right is a dotted circular structure labeled virtual image with a hole at the center. Three dotted thread-like structures with a flower like structure at the end of each structure emerge from the center and point in different directions.

Illustration c shows a shaded dot at the bottom. Above it is a horizontally placed lens followed by a slanting thin horizontal bar labeled hologram with a rectangle shaped structure at both ends. A line from source points toward the lens at a point before its left end and an arrow from the point moves upward passing through hologram. A line from source point toward the lens at a point before its right end and an arrow from the point moves upward passing through hologram. A circular structure labeled real image is shown at the right with a hole at the center. Three thread structures with a flower like structure at the end of each structure emerge from the center and point in different directions. Three zigzag arrows from hologram move toward real image and three zigzag arrows from real image move toward the human eye labeled observer at the top.

To reconstruct a three-dimensional image of the scene, two different geometries are recommended, one for viewing the virtual image and the other for viewing the real image. As indicated in [Fig. 11.9\(b\)](#), to view the virtual image we illuminate the hologram with an exact duplicate of the original reference wave, in which case the virtual image appears fixed in three-dimensional space behind the photographic plate at exactly the same location where the object was originally located. Since the wavefronts originally incident on the plate have been duplicated during the reconstruction process, the image retains all three-dimensional properties of the object. In particular, it is possible to “look behind” objects in the foreground simply by changing one’s viewing position or perspective.

The real image is best viewed when we illuminate the hologram in a different manner. Let the reconstruction wave be a wave that duplicates the reference wave in all respects except one, namely it is traveling backwards towards the original location of the reference source as if time had been reversed during the recording process. This wave can be referred to as an “anti-reference” wave, and can be thought of as being obtained by reversing the direction of the local \vec{k} vector of the reference wave at each point on the hologram. The result is a reconstruction wave with a complex distribution of field that is the complex conjugate of the original reference wave, i.e. $A^*(x,y) \vec{A}^*(x, y)$. Under such illumination, the real image forms in space between the photographic plate and the observer, as shown in [Fig. 11.9\(c\)](#). For three-dimensional objects, the real image has certain properties that make it less useful than the virtual image in many applications. First, points on the object that were closest to the photographic plate (and therefore closest to an observer of the original scene) appear in the real image closest to the photographic plate again, which in this case is *farthest from the observer* (cf. [Fig. 11.9\(c\)](#)). Thus to an observer of the real image, the parallax relations are not those associated with the original object, and the image appears (in a certain peculiar sense that must be personally observed to be fully appreciated) to be “inside out.” Images of this type are said to be *pseudoscopic*, while images with normal parallax relations (like the virtual image) are said to be *orthoscopic*.

As a second disadvantage of the real image, if photographic film is inserted directly into that image in an attempt to record it directly, the experimenter soon discovers that (for holograms of

reasonable size) the depth of focus is generally so small that a recognizable recording can not be obtained. This problem can be alleviated by illuminating only a small portion of the hologram, in which case the depth of focus is increased and a usable two-dimensional image can be recorded. If the illuminating spot on the hologram is moved, then the apparent perspective of the two-dimensional image changes. Thus every small region of a large hologram is capable of producing a real image of the original object with a different perspective!

[Figure 11.10](#) shows a photograph of a portion of a hologram of a diffusely reflecting three-dimensional scene. Note that there is nothing recognizable in the structure recorded on the hologram. In fact, most of the observable structure is irrelevant to the reconstruction in the sense that it arises from imperfections in the optical apparatus (e.g. from dust specks on mirrors and lenses). The structure that generates the reconstructed images is far too fine to be resolved in this photograph.



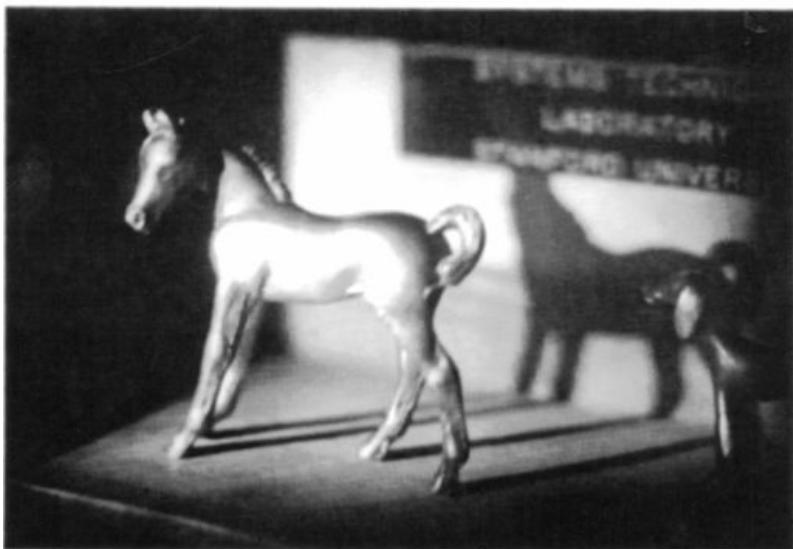
Figure 11.10
Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 11.10 Photograph of a portion of a hologram of a diffuse three-dimensional scene.

To illustrate the truly three-dimensional nature of the reconstructed images, we refer to [Fig. 11.11](#), which shows two photographs of the virtual image. In [11.11\(a\)](#), the camera is focused on the background of the virtual image; the sign in the background is sharply in focus, while the figurines in the foreground are out of focus. Note also that the tail of the horse obscures the head of the shadow of the horse. The camera is now refocused on the foreground and moved to change perspective, with the result shown in [Fig. 11.11\(b\)](#). The foreground is now in focus and the background out of focus. The tail of the horse no longer obscures the head of the shadow of the horse, a consequence of the change of perspective. Thus the camera has succeeded in “looking behind” the tail by means of a simple lateral movement.



(a)



(b)

Figure 11.11

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.11 Photographs showing the three-dimensional character of the virtual image reconstructed from a hologram.

Image a shows a toy of a horse and the head portion of the horse is not covered in the frame. Image a also shows a toy of a dog and a wall with a board at the top. The shadow of the horse including the head portion and the shadow of the dog is shown on the wall. The text on the board reads “System techniques laboratory Standford University.” Image b shows a toy of a horse and a toy of a dog of which only the front portion is covered in the frame and a wall is shown with a board at the top portion. The text on the board is unreadable. The shadow of the horse is shown which is not clear as compared to Image a.

11.4.5 Practical Problems in Holography

There are several problems that any practitioner of holography faces and must overcome in order to successfully make a hologram. To become better acquainted with the practice of holography, the reader may wish to consult [307].

Historically, the extremely short coherence lengths of optical sources available before the advent of the laser seriously constrained the types of holograms that could be recorded. Today the availability of high-quality laser sources has vastly alleviated this problem. However, the experimenter must still take some precautions, for the coherence of lasers is not perfect. For example, it is good practice to measure the distances the reference beam and the object beam travel from source to photographic plate and to equalize the lengths of these paths as closely as possible.

The process of recording a hologram is an exercise in interferometry. As with any interferometric experiment, if clean and sharp interference fringes are to be recorded, it is essential that all path-length differences for interfering light be kept stable to within a fraction of an optical wavelength during the duration of the exposure period. The higher the power available from the laser source, the shorter the required exposure time and the less severe the stability requirements become. The exposure time required depends on a multitude of factors, including the transmissivity or reflectivity of the object, the distances and geometry involved, and the particular film or plate used to record the hologram. Pulsed lasers with pulse durations as short as a few nanoseconds have been used in some instances, and CW exposures as long as several hours have been used in some cases.

Some of the most stringent experimental requirements are associated with the recording of holograms of three-dimensional scenes. Photographic emulsions with extremely high resolution are required in such cases (see [Section 11.8](#) for a more complete discussion of this point). It is invariably true that high-resolution emulsions are extremely insensitive.

An additional problem of some significance is the limited dynamic range of photographic recording materials. The amplitude transmittance versus exposure curve is linear over only a limited range of exposure. It is desirable to choose an average exposure that falls at the midpoint of this linear region. However, when the object is, for example, a transparency with a rather coarse structure, there may exist significant areas on the hologram with exposures falling well outside the linear region. As a consequence of this nonlinearity, degradation of the reconstructed images can be expected (see [Section 11.10.2](#) for further discussion). The dynamic range problem can be largely overcome by a technique first demonstrated by [Leith and Upatnieks \[224\]](#). The object is illuminated through a diffuser, which spreads the light passed by any one point on the object to cover the entire hologram. Thus a bright spot on the object will no longer generate a strong Fresnel diffraction pattern on part of the hologram, but rather contributes a more uniform distribution of light. Attendant with the advantageous reduction of dynamic range of the exposing light pattern is another advantage: since each object point contributes to every point on the hologram, an observer looking at a reconstructed image through only a portion of the hologram will always see the entire image. As might be expected, the virtual image appears to be backlit with diffuse illumination.

11.5 Image Locations and Magnification

To this point we have considered primarily collimated reference and reconstruction waves of the same wavelength. In practice these waves are more commonly spherical waves diverging from or converging toward particular points in space, and sometimes have different wavelengths.

Therefore this section is devoted to an analysis of the holographic process in this more general case. We begin by determining image locations, and then utilize these results to find the axial and transverse magnifications characteristic of the imaging process. The section then concludes with an example.

11.5.1 Image Locations

Referring to [Fig. 11.12\(a\)](#), we suppose that the reference wave is generated by a point source located at coordinates (x_r, y_r, z_r) . Since the mapping of object amplitudes into image amplitudes is linear, provided the reference offset angle is sufficiently large to separate the twin images from each other and from other unwanted terms of transmitted light, it suffices to consider a single object point source located at coordinates (x_o, y_o, z_o) . Note from the figure that, for our choice of the location of the center of the coordinate system, both z_r and z_o are negative numbers for point sources lying to the left of the hologram recording plane (i.e. for diverging spherical waves), and positive numbers for points lying to the right of that plane (i.e.\ for converging spherical waves).

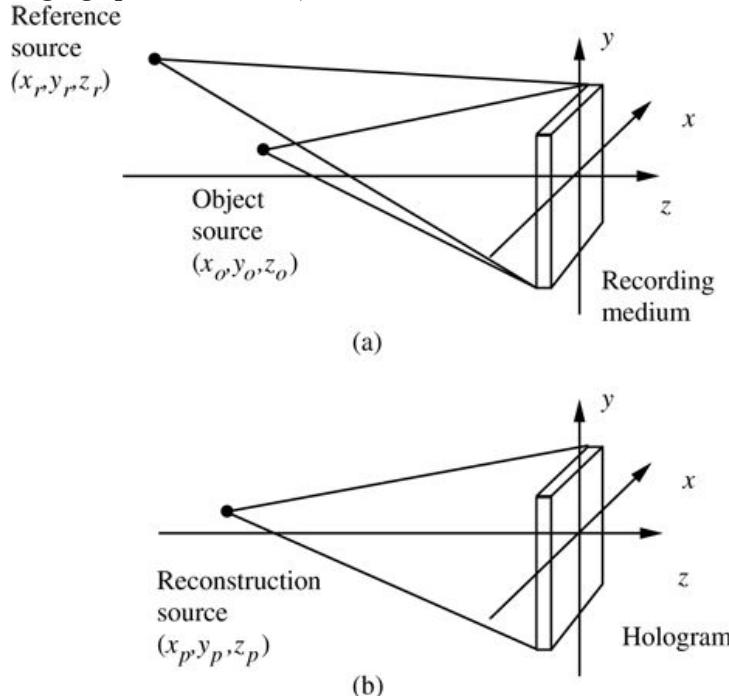


Figure 11.12

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.12 Generalized (a) recording and (b) reconstruction geometries.

Illustration a shows the horizontal axis labeled z, vertical axis labeled y and a third axis labeled x. A rectangular structure labeled recording medium is shown with vertical axis at the center. A dark spot labeled Reference source (X_r, Y_r, z_r) is on the left extreme. A line from the reference source points toward the top right corner of the recording medium and another line from the reference source points toward the bottom left corner of the recording medium. To the right of the reference source just above the horizontal axis is a dark spot labeled Object source (x_0, y_0, z_0) and a line from the object source points toward the top right corner of the recording medium and another line from the object source points toward the bottom left corner of the recording medium.

Illustration b shows the horizontal axis labeled z, vertical axis labeled y and a third axis labeled x. A rectangular structure labeled hologram is shown with vertical axis at the center. A dark spot labeled Reconstruction source (X_p, Y_p, z_p) is on the left extreme slightly above the horizontal line. A line from the reconstruction source points toward the top right corner of the hologram and another line from the reconstruction source points toward the bottom left corner of the hologram.

During the reconstruction step, illustrated in [Fig. 11.12\(b\)](#), the hologram is assumed to be illuminated by a spherical wave originating from a point source at coordinates (x_p, y_p, z_p) (x_p, y_p, z_p) . Again z_p is negative for a diverging wave and positive for a converging wave.

To achieve the maximum possible generality, we allow for the possibility that the recording and reconstruction processes may involve radiation with different wavelengths. Such is the case, for example, in microwave holography, in which the hologram is recorded with microwaves and reconstructed using visible light. The recording wavelength will be represented by λ_1 , and the reconstruction wavelength by λ_2 .

Our analysis will determine the paraxial approximations to the (twin) image locations for an object point source at the given coordinates. An extended coherent object may then be regarded as a collection of many mutually coherent point sources.

Using quadratic-phase approximations to the spherical waves involved,³ the total field incident on the recording plane may be written

$$U(x, y) = A \exp -j\pi\lambda_1 z_r (x - x_r)^2 + (y - y_r)^2 + a \exp -j\pi\lambda_1 z_o (x - x_o)^2 + (y - y_o)^2,$$

$$U(x, y) = A \exp \left\{ -j\frac{\pi}{\lambda_1 z_r} [(x - x_r)^2 + (y - y_r)^2] \right\}$$

$$+ a \exp \left\{ -j\frac{\pi}{\lambda_1 z_o} [(x - x_o)^2 + (y - y_o)^2] \right\},$$

(11-31)

where A and a are complex constants representing the amplitudes and relative phases of the two spherical waves. The corresponding intensity distribution in the pattern of interference between the two waves is

$$\mathcal{I}(x, y) = |A|^2 + |a|^2$$

$$\mathcal{I}(x, y) = |A|^2 + |a|^2$$

(11-32)

$$\begin{aligned}
& + Aa^* \exp -j\pi\lambda_1 z_r (x-x_r)^2 + (y-y_r)^2 + j\pi\lambda_1 z_o (x-x_o)^2 + (y-y_o)^2 + A^* a \exp j\pi\lambda_1 z_r (x-x_r)^2 + (y-y_r)^2 \\
& + Aa^* \exp \left\{ -j\frac{\pi}{\lambda_1 z_r} [(x-x_r)^2 + (y-y_r)^2] + j\frac{\pi}{\lambda_1 z_o} [(x-x_o)^2 + (y-y_o)^2] \right\} \\
& + A^* a \exp \left\{ j\frac{\pi}{\lambda_1 z_r} [(x-x_r)^2 + (y-y_r)^2] - j\frac{\pi}{\lambda_1 z_o} [(x-x_o)^2 + (y-y_o)^2] \right\}.
\end{aligned}$$

If the amplitude transmittance of the developed transparency is proportional to exposure, then the two important terms of transmittance are

$$t_3 = \beta' Aa^* \exp -j\pi\lambda_1 z_r (x-x_r)^2 + (y-y_r)^2 + j\pi\lambda_1 z_o (x-x_o)^2 + (y-y_o)^2 + t_4 = \beta' A^* a \exp j\pi\lambda_1 z_r (x-x_r)^2 + (y-y_r)^2 + j\pi\lambda_1 z_o (x-x_o)^2 + (y-y_o)^2.$$

$$\begin{aligned}
t_3 &= \beta' Aa^* \exp \left\{ -j\frac{\pi}{\lambda_1 z_r} [(x-x_r)^2 + (y-y_r)^2] + j\frac{\pi}{\lambda_1 z_o} [(x-x_o)^2 + (y-y_o)^2] \right\} \\
t_4 &= \beta' A^* a \exp \left\{ j\frac{\pi}{\lambda_1 z_r} [(x-x_r)^2 + (y-y_r)^2] - j\frac{\pi}{\lambda_1 z_o} [(x-x_o)^2 + (y-y_o)^2] \right\}.
\end{aligned}$$

(11-33)

The hologram is illuminated with a spherical wave, which in the paraxial approximation is described by

$$U_p(x, y) = B \exp -j\pi\lambda_2 z_p (x-x_p)^2 + (y-y_p)^2.$$

$$U_p(x, y) = B \exp \left\{ -j\frac{\pi}{\lambda_2 z_p} [(x-x_p)^2 + (y-y_p)^2] \right\}.$$

(11-34)

The two wavefronts of interest behind the transparency are found by multiplying (11-33) and (11-34), yielding

$$U_3(x, y) = t_3 B \exp -j\pi\lambda_2 z_p (x-x_p)^2 + (y-y_p)^2 + U_4(x, y) = t_4 B \exp -j\pi\lambda_2 z_p (x-x_p)^2 + (y-y_p)^2.$$

$$\begin{aligned}
U_3(x, y) &= t_3 B \exp \left\{ -j\frac{\pi}{\lambda_2 z_p} [(x-x_p)^2 + (y-y_p)^2] \right\} \\
U_4(x, y) &= t_4 B \exp \left\{ -j\frac{\pi}{\lambda_2 z_p} [(x-x_p)^2 + (y-y_p)^2] \right\}.
\end{aligned}$$

(11-35)

To identify the nature of these transmitted waves, we must examine their (x, y) dependence. Since only linear and quadratic terms in x and y are present, the two expressions U_3 and U_4 may be regarded as quadratic-phase approximations to spherical waves leaving the hologram. The presence of linear terms simply indicates that the waves are converging

towards or diverging from points that do not lie on the z axis. It remains to determine the exact locations of these real or virtual points of convergence.

Since the waves emerging from the hologram are given by a product of quadratic-phase exponentials, they must be representable as quadratic-phase exponentials themselves. Thus we can identify the coordinates (x_i, y_i, z_i) of the images if we compare the expanded equations (11-35) with a quadratic-phase exponential of the form

$$U_i(x, y) = K \exp[-j\pi\lambda_2 z_i(x - x_i)^2 + (y - y_i)^2].$$

$$U_i(x, y) = K \exp \left\{ -j\frac{\pi}{\lambda_2 z_i} [(x - x_i)^2 + (y - y_i)^2] \right\}. \quad (11-36)$$

From the coefficients of the quadratic terms in x and y we conclude that the axial distance z_i of the image points is

$$z_i = \frac{1}{z_p} \pm \frac{\lambda_2}{\lambda_1 z_r} \mp \frac{\lambda_2}{\lambda_1 z_o}$$

$$z_i = \left(\frac{1}{z_p} \pm \frac{\lambda_2}{\lambda_1 z_r} \mp \frac{\lambda_2}{\lambda_1 z_o} \right)^{-1} \quad (11-37)$$

where the upper set of signs applies for one image wave and the lower set of signs for the other. When z_i is negative, the image is virtual and lies to the left of the hologram, while when z_i is positive, the image is real and lies to the right of the hologram.

The x and y coordinates of the image points are found by equating the linear terms in x and y in (11-35) and (11-36), with the result

$$x_i = \mp \lambda_2 z_i \lambda_1 z_o x_o \pm \lambda_2 z_i \lambda_1 z_r x_r + z_i p x_p = \mp \lambda_2 z_i \lambda_1 z_o y_o \pm \lambda_2 z_i \lambda_1 z_r y_r + z_i p y_p.$$

$$\begin{aligned} x_i &= \mp \frac{\lambda_2 z_i}{\lambda_1 z_o} x_o \pm \frac{\lambda_2 z_i}{\lambda_1 z_r} x_r + \frac{z_i}{z_p} x_p \\ y_i &= \mp \frac{\lambda_2 z_i}{\lambda_1 z_o} y_o \pm \frac{\lambda_2 z_i}{\lambda_1 z_r} y_r + \frac{z_i}{z_p} y_p. \end{aligned} \quad (11-38)$$

Equations (11-37) and (11-38) provide the fundamental relations that allow us to predict the locations of images of point sources created by the holographic process. Depending on the geometry, it is possible for one image to be real and the other virtual, or for both to be real or both virtual (see Prob. 11-2).

11.5.2 Axial and Transverse Magnifications

The axial and transverse magnifications of the holographic process can now be found from the equations derived above for image locations. The transverse magnification is easily seen to be

given by

$$Mt = \partial x_i / \partial x_o = \partial y_i / \partial y_o = \lambda_2 z_i / \lambda_1 z_o = 1 - z_o z_r \mp \lambda_1 z_o / \lambda_2 z_p - 1.$$

$$M_t = \left| \frac{\partial x_i}{\partial x_o} \right| = \left| \frac{\partial y_i}{\partial y_o} \right| = \left| \frac{\lambda_2 z_i}{\lambda_1 z_o} \right| = \left| 1 - \frac{z_o}{z_r} \mp \frac{\lambda_1 z_o}{\lambda_2 z_p} \right|^{-1}. \quad (11-39)$$

Similarly, the axial magnification is found to be

$$Ma = \partial z_i / \partial z_o = \partial z_o / \partial z_p = \lambda_2 / \lambda_1 z_r \mp \lambda_2 / \lambda_1 z_o - 1 = \lambda_1 \lambda_2 Mt^2.$$

$$M_a = \left| \frac{\partial z_i}{\partial z_o} \right| = \left| \frac{\partial}{\partial z_o} \left(\frac{1}{z_p} \pm \frac{\lambda_2}{\lambda_1 z_r} \mp \frac{\lambda_2}{\lambda_1 z_o} \right)^{-1} \right| = \frac{\lambda_1}{\lambda_2} M_t^2. \quad (11-40)$$

Note that in general the axial and transverse magnifications will not be identical. This can be very important when we consider the imaging of three-dimensional objects, as we shall do shortly, for the difference between these magnifications will create a three-dimensional distortion of the image. There does exist one additional parameter that can be used to combat such distortions, namely, it is possible to *scale* the hologram itself between the recording and reconstruction process. For example, if the hologram were formed with microwaves or acoustic waves, it would be possible to plot out the hologram at any scale size we choose, and record a transparency of the hologram with magnification or demagnification. If m is the magnification ($m > 1$) or demagnification ($m < 1$) to which the hologram is subjected, then we can show that the transverse and axial magnifications take the form

$$Mt = m(1 - z_o z_r) \mp m^2 \lambda_1 z_o / \lambda_2 z_p - 1$$

$$M_t = m \left| 1 - \frac{z_o}{z_r} \mp m^2 \frac{\lambda_1 z_o}{\lambda_2 z_p} \right|^{-1}$$

$$(11-41)$$

$$Ma = \lambda_1 \lambda_2 Mt^2.$$

$$M_a = \frac{\lambda_1}{\lambda_2} M_t^2.$$

$$(11-42)$$

11.5.3 An Example

Consider an example in which we record a hologram at wavelength $\lambda_1 = 10$ cm in the microwave region of the spectrum, and reconstruct images in the visible region of the spectrum

with $\lambda_2 = 5 \times 10^{-5}$ cm. [Figure 11.13](#) illustrates how the experiment might be performed. A microwave source illuminates an object with variable microwave transmittance, perhaps a three-dimensional structure which partially absorbs microwaves. A mutually coherent microwave source provides a reference wave which interferes with the radiation diffracted by the object. Some distance away a scanning rig with a microwave horn antenna measures the microwave intensity impinging on a larger aperture. To be specific, suppose that the size of the scanned aperture is $10\text{m} \times 10\text{m}$. Attached to the scanning microwave horn antenna is a light bulb, which is driven with a current that is proportional to the incident microwave power at each point in the scanned aperture. A camera records a time exposure of the brightness pattern of the light bulb as it scans across the microwave aperture, and this recorded photograph generates an optical transparency that is inserted into an optical system and illuminated at the visible wavelength quoted above.

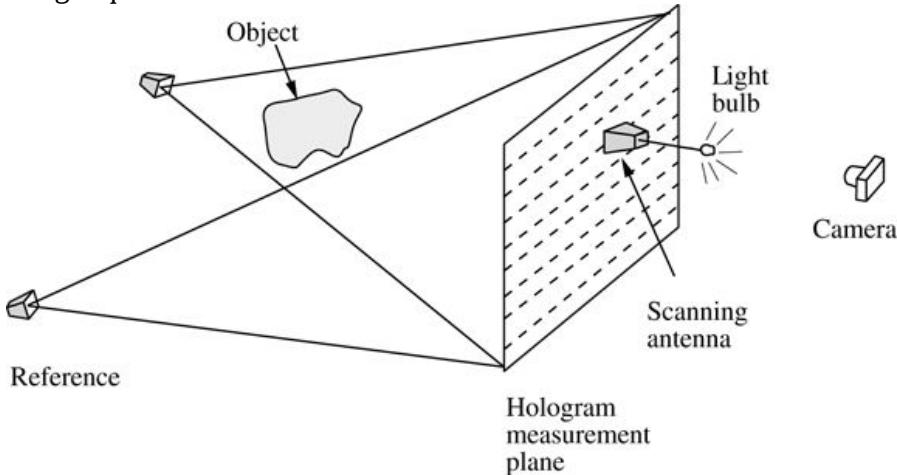


Figure 11.13

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.13 Recording a microwave hologram. The sources that provide the object and reference illuminations are derived from the same microwave signal generator to assure coherence.

The illustration shows a small rectangular box labeled reference on extreme left and another rectangular box above reference. Two lines from the upper rectangular box point toward the upper and lower ends of hologram measurement plane, respectively. Two lines from reference point toward the upper and lower ends of hologram measurement plane, respectively. To the right of reference is a rectangular box labeled hologram measurement plane with dashes arranged horizontally in 8 rows. Between the hologram measurement plane and the reference is an irregular shaped structure labeled object. Inside the hologram measurement plane is a rectangular box labeled scanning antenna from which a line extends to the right and ends with a bulb shaped structure labeled light bulb and 6 small horizontal lines emerge from the bulb and point in different directions. To the right of hologram measurement plane is the camera which is a small rectangular structure with a lens shaped structure facing the left.

If we imagined a physically impossible case of a photographic transparency that is as large as the total scanned microwave aperture, and we suppose that the microwave reference wave supplied in the recording process is a plane wave ($z_r = \infty$) and the optical reconstruction

wave is a plane wave ($z_p = \infty$), then application of (11-39) and 11-40 yield the following transverse and axial magnifications:

$$Mt=1Ma=2\times 105.$$

$$\begin{aligned} M_t &= 1 \\ M_a &= 2\times 10^5. \end{aligned}$$

As can be seen from these numbers, there is an enormous amount of distortion of the image, with the transverse magnification being more than five orders of magnitude smaller than the axial magnification.

Now suppose that we modify this experiment such that the photograph is optically reduced to be only 50μ m on a side, which corresponds to a demagnification of $m=\lambda_2/\lambda_1=5\times 10^{-6}$
 $m = \lambda_2 / \lambda_1 = 5\times 10^{-6}$. Again the reference wave and the reconstruction wave are assumed to be plane waves. In this case we find that the transverse and axial magnifications are given by

$$Mt=5\times 10^{-6}Ma=5\times 10^{-6}.$$

$$\begin{aligned} M_t &= 5\times 10^{-6} \\ M_a &= 5\times 10^{-6}. \end{aligned}$$

Thus we see that the two magnifications have been made equal to each other by means of the scaling of the hologram by the wavelength ratio, thereby removing the three-dimensional distortion. Unfortunately in the process the image has been made so small (5×10^{-6} times smaller than the original object) that we may need a microscope to examine it, in which case the microscope will introduce distortions similar to what we have removed from the holographic process.

The above example is somewhat contrived, but it does illustrate some of the problems that can be encountered when the recording and reconstruction wavelengths are significantly different. Such is often the case for acoustic holography and microwave holography. For holography at very short wavelengths such as the ultraviolet and X-ray regions of the spectrum, the problem is reversed, and the hologram must be scaled upwards in size in order to avoid distortions.

11.6 Some Different Types of Holograms

Attention is now turned to a brief guided tour through several different kinds of holograms. There are many different aspects with respect to which holograms may differ, and this has led to a rather confused classification system, in which a given hologram may in fact be properly classified in two or more different classes at the same time. There is nothing fundamentally wrong with this, as long as we understand what the different classes mean. In what follows we do not include the categorization “thin” versus “thick” as a classification, only because these differences will be discussed in detail in later sections.

11.6.1 Fresnel, Fraunhofer, Image, and Fourier Holograms

Our first dimension of classification is one that distinguishes between the diffraction or imaging conditions that exist between the object and the photographic plate where the hologram is recorded. Thus we say that a hologram is of the *Fresnel* type if the recording plane lies within the region of Fresnel diffraction of the illuminated object, whereas it is of the *Fraunhofer* type if the transformation from object to hologram plane is best described by the Fraunhofer diffraction equation.

In some cases a hologram is recorded in what must be called an image plane, and such a hologram would be referred to as an *image* hologram. This geometry is most frequently used when the object is three-dimensional but perhaps not extremely deep in the third dimension. The middle of the object can then be brought to focus in the plane of the photographic plate, and the resulting image obtained from the hologram will appear to float in space at the hologram, with parts of the object extending forwards and backwards from the hologram.

A category that applies primarily to transparency objects is the *Fourier* hologram, for which the recording plane resides in a plane that will yield the Fourier transform of the object amplitude transmittance. Thus with a normally illuminated object transparency in the front focal plane of a lens and the recording plane in the rear focal plane of the lens, the relation between fields in the two planes will be that of a Fourier transform. For such a hologram, the light from each point on the object interferes with the reference beam (assumed planar) to create a sinusoidal fringe with a vector spatial frequency that is unique to that object point. The transformation from object points into sinusoidal fringes of unique spatial frequencies is thus characteristic of the Fourier transform hologram. To view the reconstructed images, we can place the hologram in front of a positive lens, illuminate it with a normally incident plane wave, and look for images in the rear focal plane. Note that both of the twin images come to focus in the same plane for such a hologram, as can be verified by applying (11-37).

Finally, for transparency objects one sometimes sees mention of a hologram that is called a *lensless Fourier transform* hologram. The name is a misnomer, for the geometry usually requires a lens, but not a Fourier transforming lens as is used in the ordinary Fourier transform geometry. Rather, as shown in Fig. 11.14, the reference wave is brought to focus in the plane of the object transparency, and then diverges to the recording plane without passing through any optical elements. Likewise, the wave transmitted by the object propagates to the recording plane without the intervention of any optical elements. The interference pattern is then recorded. The distance between the object and the hologram recording plane is immaterial. The reason for associating the

words *Fourier transform* with such a hologram, when in fact no Fourier transform actually takes place, can be understood by considering the interference pattern generated by light from a single point on the object. Both the object wave and the reference wave are diverging spherical waves with the same curvature, and as a consequence when they interfere, the pattern of intensity is (within the paraxial approximation) a sinusoidal fringe of a vector spatial frequency that is unique to that object point. This is the same property that holds for a true Fourier transform hologram, and hence the mention of *Fourier* in the name for this type of hologram. The difference between this type of hologram and the true Fourier transform hologram lies in the spatial phases that are associated with the various sinusoidal fringes, which in this case are not the phases of the Fourier transform of the object. The twin images can be observed if the fields transmitted by the hologram are Fourier transformed by a positive lens. Again, if the hologram is illuminated with a plane reconstruction wave, both images appear in the focal plane of the transforming lens.

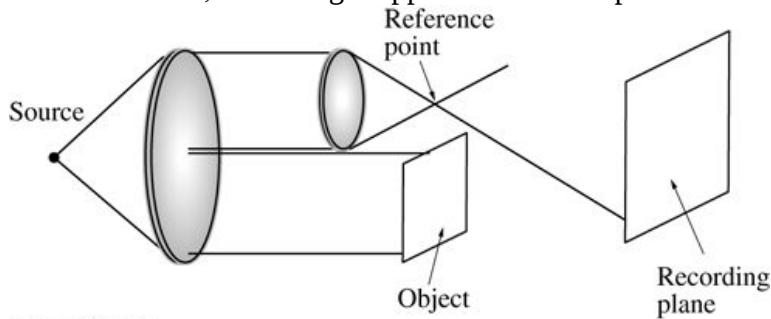


Figure 11.14

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.14 Recording a lensless Fourier transform hologram.

The illustration shows a dark spot labeled source on the extreme left. To its right is a lens. Two lines from source point toward the upper and lower ends of the lens. To the right of the lens is another lens which is smaller in size and is followed by a square shaped structure labeled object. A line from the upper end of the first lens points toward the upper end of the second lens. A line from the center of the first lens points toward the lower end of the second lens. A line from the lower end of the first lens points toward the lower end of object. To the right of object, is a big square shaped structure labeled recording plane. A line from the top end of the second lens points toward the bottom end of recording plane. A line from the bottom end of second lens points toward the upper right direction. The meeting point of the two lines is labeled reference point.

The encoding of object points into sinusoidal fringes of constant frequency should be contrasted with the *Fresnel* hologram, in which each object point is encoded into a portion of a frequency-chirped sinusoidal fringe (a sinusoidal zone plate) with an entire range of spatial frequency components present. The Fourier transform and lensless Fourier transform holograms make the most efficient use of the space bandwidth product of the hologram.

11.6.2 Transmission and Reflection Holograms

The majority of the holograms discussed so far have been of the *transmission* type. That is, we view the images in light that has been transmitted through the hologram. Such holograms are comparatively tolerant to the wavelength used during the reconstruction process (although the amount of tolerance depends on the thickness of the emulsion), in the sense that a bright image can be obtained without exactly duplicating the wavelength used during exposure. However this also

leads to chromatic blur when a transmission hologram is illuminated with white light, so some filtering of the source is generally required.

Another important class of holograms is that of *reflection* holograms, for which we view the images in light that is reflected from the hologram. The most widely used type of reflection hologram is that invented by Y. Denisyuk in 1962 [89]. The method for recording such a hologram is illustrated in [Fig. 11.15\(a\)](#). In this case there is only one illumination beam, which supplies both the object illumination and the reference beam simultaneously. As shown in the figure, the object is illuminated *through the holographic plate*. The incident beam first falls upon the holographic emulsion, where it serves as a reference wave. It then passes through the photographic plate and illuminates the object, which usually is three-dimensional. Light is scattered backwards from the object, towards the recording plane, and it passes through the emulsion traveling in a direction that is approximately opposite to that of the original incident beam. Within the emulsion the two beams interfere to produce a standing interference pattern with extremely fine fringes. As will be seen in [Section 11.7](#), the period of the sinusoidal fringe formed when two plane waves traveling at angle 2Θ with respect to each other interfere is given by

$$\Lambda = 2\pi|K| = \lambda / 2\sin\Theta.$$

$$\Lambda = \frac{2\pi}{|K|} = \frac{\lambda}{2\sin\Theta}.$$

(11-43)

When $2\Theta=180^\circ$, as is the case for the reflection hologram, the fringe period is half of an optical wavelength in the emulsion.⁴ As will be seen in [Section 11.7](#), the fringes are oriented such that they bisect the angle between directions of travel of the reference and the object waves, and are therefore approximately *parallel to the surface of the emulsion* for a reflection hologram.

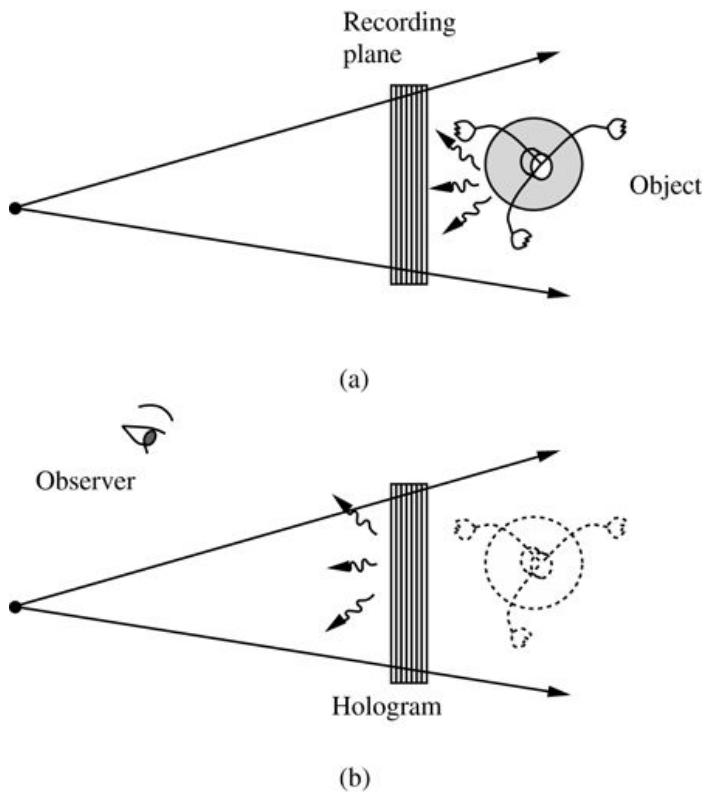


Figure 11.15
Goodman, Introduction to Fourier Optics, 4e,
 © 2017 W. H. Freeman and Company

Figure 11.15 (a) Recording a reflection hologram, and (b) reconstructing an image in reflected light.

Illustration a shows a dark spot in the extreme left followed by a vertical rectangular bar labeled recording plane with vertical lines inside. Next to it is a circular shaped structure labeled object with a hole at the center. Three thread-like structures with a flower like structure at the end of each thread is shown and the threads point in different directions. Three Zig zag arrows from object point toward the recording plane. Two arrows from the dark spot pointing rightward pass through the two points, one near the upper end and other one near the lower end of recording plane.

Illustration b shows a dark spot in the extreme left followed by a vertical rectangular bar labeled hologram with vertical lines inside. Next to it is a dotted circular shaped structure with a hole at the center. Three dotted thread-like structures with a flower like structure at the end of each thread is shown and the threads point in different directions. Three Zig zag arrows from hologram point in three different directions. Two arrows from the dark spot pointing rightward pass through two points, one near the upper end and other one near the lower end of recording plane. A human eye labeled observer is shown on the upper left.

[Figure 11.15\(b\)](#) shows how the virtual image would be viewed for a reflection hologram. The hologram is illuminated by a duplication of the original reference wave, and a duplicate of the object wave is generated, which in this case is a reflected wave. The observer looks into the reflected wave and sees the virtual image in the original location of the object, behind the hologram. [Figure 11.16](#) shows a photograph of the virtual image reconstructed from a reflection hologram that is being illuminated by white light.



Figure 11.16

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.16 Photograph of a virtual image reconstructed from a reflection hologram.

This type of hologram can be illuminated with white light, for the hologram is highly wavelength selective, and the wavelength that satisfies the Bragg condition will automatically be reflected, while others will not. In this regard it should be noted that photographic emulsions usually suffer some shrinkage during the chemical processing and drying steps, and therefore the color of the light reflected from this type of hologram will usually be different than that used during recording. For example, a hologram recorded with red light may reflect green light. Such effects can be compensated by intentionally swelling the emulsion by means of proper chemical treatment.

11.6.3 Holographic Stereograms

At a relatively early stage in the development of holography, several ideas emerged for using the holographic process to capture a multitude of images that were recorded by conventional photography and to create the illusion of three dimensions through the stereo effect. The function of holography in these schemes is to allow the observer to see different images, taken from different perspectives, in each eye, thereby creating the stereo effect. The fact that the process begins with ordinary photography, and does not require that the original scene be illuminated by a laser, is a distinct advantage. A laser is required in the hologram-recording process. References include [244], [88], and [296].

One method for recording such a hologram is illustrated in Fig. 11.17 [88]. A series of black and white photographs are taken of the subject from a sequence of horizontal positions, each with its own unique perspective. Each frame of the sequence is then projected with light from a laser onto a translucent screen. A reference beam is introduced and a hologram is recorded through a movable slit. As the photographic frame is advanced, the slit is moved, with the result that a

multitude of holograms are recorded side by side, each hologram capable of reconstructing an image of the original object taken from a different horizontal perspective. If the resulting hologram is illuminated in its entirety by a duplicate of the reference wave, the observer will look through a different holographic stripe with each eye, and therefore each eye will see the subject from a different perspective, creating a three-dimensional image through the stereo effect.

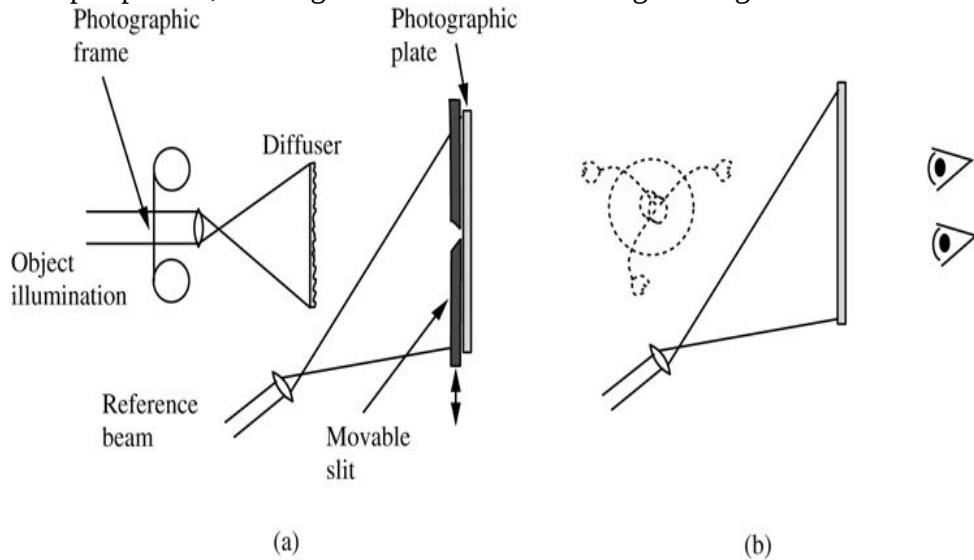


Figure 11.17

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 11.17 Recording a holographic stereogram (top view). (a) Recording the holograms, and (b) viewing the image.

Illustration a shows a small horizontal bar labeled photographic plate whose right end is in the shape of a lens. Two circular shaped structures are shown in the upper and lower sides of the bar which are connected by a line. The text on the left extreme reads “Object illumination.” To the right of the photographic plate is a curved vertical line labeled diffuser. A line from the top end of photographic plate points toward the lower end of diffuser and a line from the lower end of photographic plate points toward the upper end of diffuser. At the bottom, another slanting bar labeled reference beam is shown with a lens shaped structure at its upper end. To its right is a dark vertical plate labeled movable slit with a V shaped opening at the center. Adjoining the movable slit is another thin plate labeled photographic plate. A line from the upper end of the reference beam points toward a point slightly above the lower end of movable slit and a line from the lower end of the reference beam points toward a point slightly below the upper end of movable slit.

Illustration b shows a slanting bar at the bottom left with a lens shaped structure at its upper end. To its right is a vertical plate. A line from the upper end of the slanting bar points toward a point slightly above the lower end of the vertical bar and a line from the lower end of the slanting bar points toward a point slightly below the upper end of the vertical bar. Above the slanting bar is a dotted circular shaped structure with a hole at the center. Three dotted thread-like structures with a flower like structure at the end of each thread is shown and the threads point in different directions.

An alternative approach [296] uses angular multiplexing in thick holograms to superimpose a multitude of overlapping holograms on a photographic plate. Each eye views the photographic plate from a slightly different angle, and as a consequence the Bragg effect leads to the reconstruction of a different image seen by each eye, and a resulting three-dimensional image is created.

11.6.4 Rainbow Holograms

An important advance in the field of display holography was the invention of the *rainbow hologram* by [S. Benton \[23\]](#). This invention provides a method for utilizing white light as the illumination when viewing the hologram, and does so by minimizing the blur introduced by color dispersion in a transmission hologram, at the price of giving up parallax information in one dimension. The ability to view holographic images in white light was a vital step on the road to making holography suitable for display applications.

The method entails a two-step process, in which an initial hologram is made, and then a second hologram is made using the first hologram as part of the process. The first step in the process is to record a hologram of a three-dimensional scene in the usual way, in particular using monochromatic or nearly monochromatic light, as is illustrated in [Fig. 11.18\(a\)](#). The light from the reference source R_1 and light scattered by the object O interfere to form holographic recording H_1 . This recording is processed in the usual way, and a hologram results. We now illuminate this first hologram with a monochromatic “anti-reference” wave, i.e. a wave that duplicates the original reference wave, except that the direction of travel is reversed, as illustrated in [Fig. 11.18\(b\)](#). A real image of the original object is produced by hologram H_1 , and the location of that real image coincides with the original location of the object.

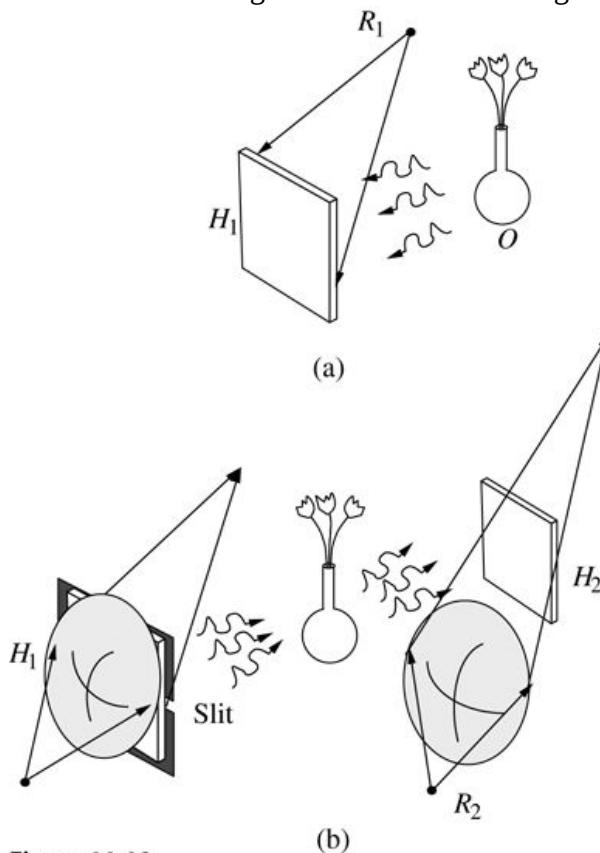


Figure 11.18

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.18 The rainbow hologram. (a) The first recording step, and (b) the second recording step.

Illustration a shows a square shaped structure with left side of the square labeled H1. To its upper right is a dark spot labeled R1. An arrow from R1 points to the upper left corner of the square and another arrow from R1 points to the right side of square at a point slightly lower than the midpoint. To the right of the square is a flower vase from which three curved arrows point toward the square. Illustration b shows a dark spot at the bottom left. Slightly above it is an oval shaped structure with an ‘X’ mark on it. Two arrows from the dark spot point toward the left and right sides of the oval. Adjoining the oval shaped structure is a square shaped structure labeled slit with its left side labeled H1. To its upper right is a dark spot labeled R1. An arrow from R1 points to the upper left corner of the square and another arrow from R1 points to the right side of square at a point slightly lower than the midpoint. A flower vase is shown at the center. Three curved arrows from the slit point toward the flower vase. To the right of the flower vase is a square shaped structure whose right side is labeled H2. Below H2 is an oval shaped structure with an ‘X’ mark on it. A dark spot labeled R2 is shown below the oval and two arrows from the dark spot point toward the left and right ends of the oval. Two arrows from the left and right ends of the oval point towards a dark spot above the square. Three curved arrows from the flower vase points toward the square shaped structure.

Now a new element is introduced in the reconstruction geometry of [Fig. 11.18\(b\)](#), namely a narrow horizontal slit immediately following hologram H1 H_1 . The light passing through this slit again reconstructs a real image of the original object, but this time vertical parallax is eliminated; the image that is formed is the one that would have been seen from the particular vertical location of the slit. Having created this real image, a second hologram H2 H_2 is recorded, this time a hologram of the image produced by the first hologram, with a new reference wave being used, in particular a reference wave that is a converging spherical wave. Again the light used in the recording process is monochromatic, and the pattern of interference is between the reference wave from R2 R_2 and the light that has passed through the focus of the real image and traveled on to the recording plane, as shown in [Fig. 11.18\(b\)](#). H2 H_2 is the final hologram created by this process.

The hologram obtained by the method described above is now illuminated with a diverging spherical wave, which is in fact the “anti-reference” wave for the converging reference wave from R2 R_2 , as shown in [Fig. 11.19\(a\)](#).

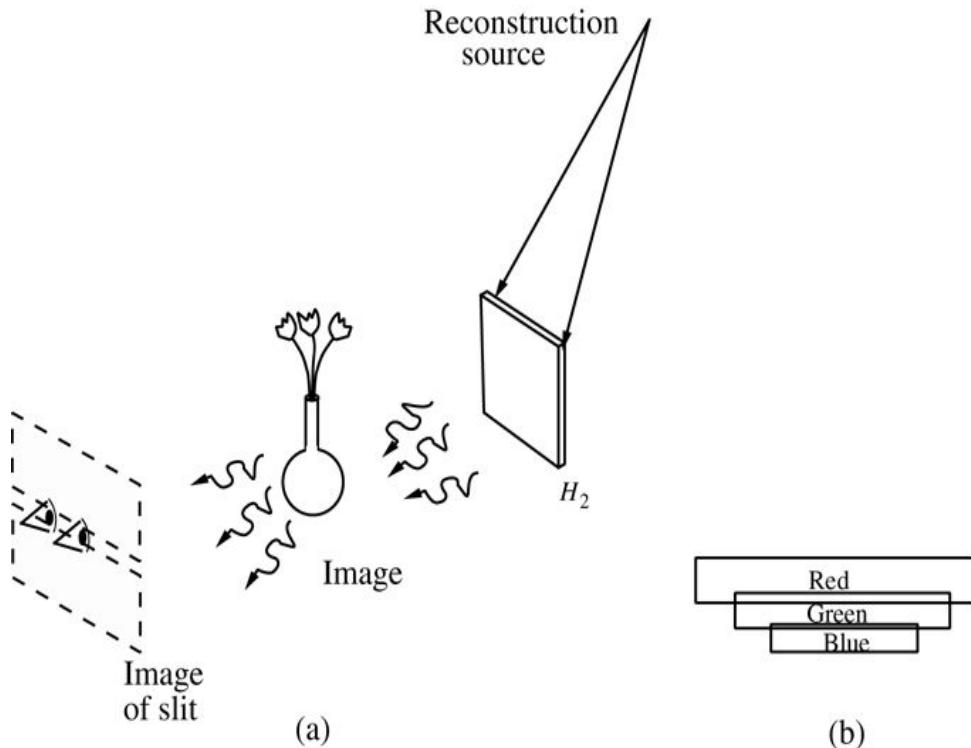


Figure 11.19
Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 11.19 Reconstruction of the image from a rainbow hologram; (a) Reconstruction geometry, (b) slit sizes at different wavelengths.

Illustration a shows a dotted slanting parallelogram labeled image of slit at the extreme left with two horizontal vertical lines running in the center. Two dotted triangular shaped structures with curved bases are shown with their bases facing the right and with a shaded dot in both triangles. A flower vase labeled image is at the center. Three curved arrows from the vase point toward the image of slit. To the right of the flower vase is a square labeled H_2 . Two arrows from a point above the square point toward the top left and right corners of the square. Three curved arrows from the square point toward the image.

Illustration b shows a rectangular bar labeled blue at the base. Above it is another rectangular bar labeled green which is lengthier than the one at the base. A third rectangular bar above green is labeled red which is bigger than the bottom two.

The hologram forms a real image⁵ of the original object, but beyond that image, closer to the viewer, there is also formed an image of the slit that was introduced when hologram H_2 was recorded. Now if the reconstruction source in this last step emits white light, then the dispersion of the hologram will in fact generate a blur of images of both the object and the slit. In particular, each different narrow band of colors of the reconstruction source will create an image of the slit at a different vertical location (and with a different magnification), with red light having been diffracted vertically the most and blue light the least. An observer located in the plane of the images of the slit will in effect look through a slit appropriate for only a narrow color band and will intercept no light outside of this color band. Thus the image will appear free from color blur, and will have a color that depends on exactly where the observer's eyes are located in the vertical dimension. Tall observers will see an image with a different color (and a somewhat different

magnification) than short observers. Thus the dispersive properties built into hologram H_2 have been used to advantage to eliminate color blur and to allow the image to be viewed with a white light source. As shown in [Fig. 11.19\(b\)](#), the slit position varies with color in both vertical position and magnification.

11.6.5 Multiplex Holograms

Another major type of hologram that has been widely used for display purposes is the *multiplex hologram* invented by [Lloyd Cross \[80\]](#). Good descriptions of the multiplex hologram process can be found in [\[307\]](#).

The process begins with a series of still-frame photographs, typically made with a motion picture camera operated a single frame at a time. [Figure 11.20\(a\)](#) shows the process. A subject is placed on a slowly rotating platform and still-frame photographs are taken, typically at a rate of three frames for every degree of rotation of the subject. Thus for a hologram that is to offer a 120° viewing angle, a total of 360 images are recorded. During the rotation process, the subject may undergo some motion or movement of its own, a motion that will eventually be evident in the image viewing process.

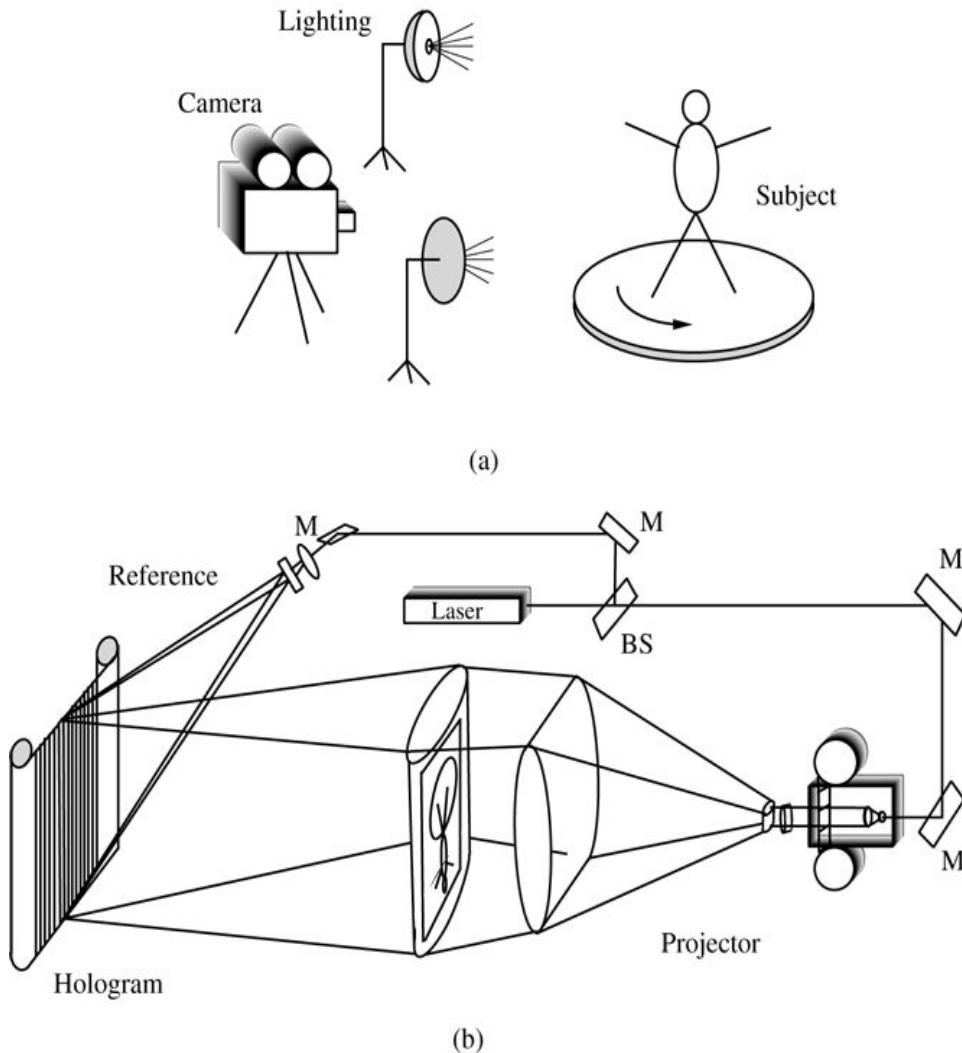


Figure 11.20
Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 11.20 Constructing a multiplex hologram. (a) Recording the still-frame sequence, and (b) recording the multiplex hologram. M indicates a mirror, BS a beam splitter. The reference wave arrives at the recording plane from above.

Illustration a shows a drawing of a camera on the extreme left. To the right of it are two flash lights which are labeled lightening. To the right of the flash lights is a rotating disc above which is a drawing of a human which is labeled subject.

Illustration b shows a rectangular structure with curved corners labeled hologram with vertical lines and small oval shaped structure at the top ends. Right to hologram is a vertical rectangular structure with a diagram of a human inside and is followed by a cylindrical shaped structure. Two lines from the top center point toward the upper left and right ends respectively. Two lines from the bottom center point toward the upper left and right ends, respectively. Two lines from the upper left and upper right ends of the vertical rectangular structure point toward the upper left and right end of the cylindrical bar. Similarly, two lines from the lower right and lower left ends of the vertical rectangular structure point toward the lower left and right ends of the cylindrical bar. Right to the cylindrical bar is a rectangular box labeled projector with two circular shaped structures at the top and bottom and has a thin tube at the center that extends to the left. Two lines from the

upper left and right ends and two lines from the lower left and lower right ends of the cylindrical bar point toward the tube shaped structure of the projector. A line from the right end of the tube shaped structure extends to the right and bends upward and again bends leftward and points toward a rectangular shaped structure labeled laser. At the lower and upper corners of the line is a small slanting rectangular bar labeled M. Two lines from the upper center of hologram points toward a short rectangular bar labeled reference at the top and get reflected back to the bottom center of hologram. Two lines from the rectangular bar points toward a lens shaped structure and a line from the lens shaped structure bends rightward and again bends in downward direction and small rectangular shaped structure labeled M is shown at the corners of the line. The line meets the line that starts from projector and a slanting small horizontal bar is shown at the meeting point of the two lines and is labeled BS.

The sequence of photographs obtained as above is now fed through a special projector, as shown in [Fig. 11.20\(b\)](#). Using light from a laser, the images are projected onto a large cylindrical lens, which focuses the light down to a narrow vertical stripe on the large film strip that is to record the hologram. At the same time a portion of the laser light is brought above the projection system and focused to a vertical stripe that coincides with the projected image and provides a reference beam that is offset from the object beam by an angle in the vertical dimension. Thus a narrow vertical stripe hologram is recorded for a given still-frame photo, with a carrier frequency running vertically. The film is now advanced to the next still-frame photo, and the holographic film is moved so that a new vertical stripe on the film will be exposed, usually partially overlapping the adjacent stripe. Through a sequence of such exposures, the 360 holograms are recorded. Note that each still-frame photo, and therefore each holographic stripe, contains image information taken from a different perspective in the original photographic process.

To view a three-dimensional image with the hologram after processing, the film is bent into a cylindrical shape and illuminated from within the cylinder with a white-light source, properly positioned in the vertical direction to account for the reference angle used during recording (see [Fig. 11.21](#)). An illumination source with a clear bulb and a vertical filament is required in order to avoid image blur. The observer looks into the hologram and sees an apparently three-dimensional image within the cylinder. The white light is dispersed in the vertical dimension by the holographic gratings, with red light suffering more downward deflection than blue light. An observer looking into the hologram will automatically perform two types of selection. First the vertical position of the observer's head will place him or her in a certain narrow region of the color spectrum, so color filtering is performed simply by geometry. Second, the two eyes of the observer will look through different regions of the multiplex hologram, and therefore will be looking predominantly through two different holographic stripes, each of which yields an image of the original object taken from a different perspective. As a consequence the stereo effect creates the illusion that the object is three-dimensional. As the observer moves horizontally, the image appears to be stationary in space and the perspective changes accordingly. If the subject goes through some form of motion while the original platform is being rotated, then corresponding motion of the three-dimensional image will be seen as the viewer moves around the hologram, or alternatively as the hologram is rotated. Note that, as with the rainbow hologram, vertical parallax is not present in the image.

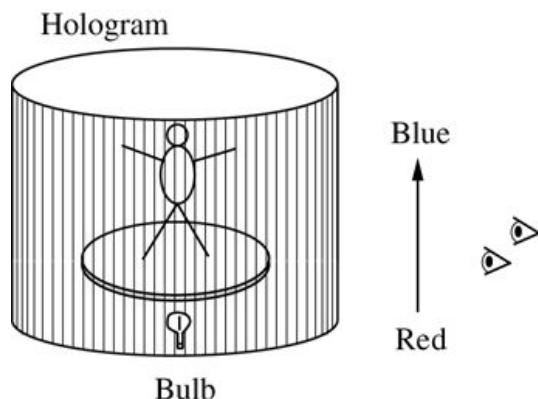


Figure 11.21

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.21 Viewing the image with a multiplex hologram.

An illustration shows a cylindrical structure labeled hologram with vertical lines inside. Inside hologram is a disc above which a human diagram is shown. Below the disc is a bulb. To the right of hologram is a vertical arrow pointing upward. The text at the bottom of arrow reads “Red.” The text at the top of the arrow reads “Blue.” To its right are two human eyes.

11.6.6 Embossed Holograms

Embossing has become a highly refined and advanced technique for replicating compact disks and DVDs, which have structures of the same order of size as an optical wavelength. The same techniques can be applied to the replication of holograms, with substantial cost savings as compared with optical methods of duplication. The ability to produce holograms inexpensively has led to their use, for example, in security cards, credit cards, magazines, books, and in some cases on monetary bills. We outline here the various steps that are involved in creating an embossed hologram.

The first step in the process is to record a hologram of the subject of interest, on photoresist. With a proper choice of photoresist, the resolution is fully adequate to the task at hand. Usually a rather powerful argon-ion laser is used in the recording step. The exposed photoresist is then developed, leading to a relief pattern that constitutes the photoresist master hologram.

A metal master hologram is now made from the photoresist hologram by means of an electroforming process. A silver spray is applied to the photoresist surface, making it conducting. The master is then immersed in a plating tank, together with a bar of pure nickel, and current is passed through the tank with the result that a thin layer of nickel is plated on top of the photoresist master. The layer of nickel, which forms the metal master, is then separated from the photoresist. It is now possible to use the metal master in a second electroforming process, in which a second-generation metal submaster can be made from the original. The process can be repeated to make many metal submasters, which will serve as stampers in the reproduction process.

With the metal submasters in hand it is now possible to initiate the embossing process. There are several different methods for embossing, including flat-bed embossing, roll embossing, and hot stamping. In all cases the metal submaster is heated to an elevated temperature, and used to stamp the hologram pattern, usually into a polyester material. Often the embossed pattern is metallized to create a reflection hologram.

Without doubt, of all the holograms in existence today, the largest number are of the embossed type, for only with embossing can the cost of reproducing holograms be brought down to the levels needed for extremely high-volume applications.

11.7 Thick Holograms

Just as for acousto-optic spatial light modulators (see [Section 9.5](#)), holograms behave differently depending on the relation between the period of the finest fringe they contain and the thickness of the recording medium. It is therefore common to categorize holograms as *thick* or *thin*, depending on this relation. Like the acoustic waves in an acousto-optic SLM, a hologram is a grating. Unlike the acousto-optic case, the grating is stationary rather than moving, and it may also be highly absorbing, partially absorbing, or nonabsorbing, depending on the conditions of exposure and the photographic processing to which it has been subjected.

If we consider a hologram consisting of a single sinusoidal grating with grating planes normal to the surface of the emulsion, it behaves as a thick or thin grating depending on the value of the Q parameter of (9-27), which is repeated here,

$$Q = 2\pi\lambda_0 d n \Lambda^2,$$

$$Q = \frac{2\pi\lambda_0 d}{n\Lambda^2},$$

(11-44)

where λ_0 is the vacuum wavelength of the light used during reconstruction, n is the refractive index of the emulsion after processing, Λ is the period of the sinusoidal grating, and d is the emulsion thickness. Again, for $Q > 2\pi$ the grating is considered “thick,” while for $Q < 2\pi$ the grating is “thin.”

The most common photographic plates used in holography have thicknesses of the order of $15 \mu m$, while the fringes formed in holograms may be only a few wavelengths, or in some cases as small as half a wavelength, depending on the angle between the reference wave and the object wave. Therefore a hologram of an object with any significant angular subtense at the hologram plane will contain at least some fringes that exhibit the properties of a thick grating. Hence Bragg diffraction effects must be considered in most cases.

In this section we consider in more detail the properties of the gratings recorded by the holographic process, and the requirements for achieving high diffraction efficiency from such gratings. Finally we determine the diffraction efficiencies of thick holograms and compare them with those of thin holograms. An excellent and far more detailed treatment of this subject will be found in [324].

11.7.1 Recording a Volume Holographic Grating

Consider the very simple case of a plane reference wave and a plane object wave incident on an emulsion of non-negligible thickness. These two simple waves may be regarded as generating a simple holographic grating.

With reference to [Fig. 11.22](#), it is assumed for simplicity that the two wave normals (represented by arrows and pointing in the directions of the two \vec{k} vectors), are each inclined at angle Θ to the surface normal. Wavefronts, or successive lines of zero phase, are shown dotted; the wavefronts of any one wave are spaced by a normal distance of one wavelength. Along the lines (points in this two-dimensional figure) within the emulsion where the wavefronts of the two waves intersect, the two amplitudes add in phase, yielding high exposure. As time progresses, the wavefronts move in the direction of their respective wave normals, and the lines of constructive interference move through the emulsion, tracing out *quasi-planes* of high exposure. Simple geometry shows that these planes bisect the angle 2Θ between the two wave normals and occur periodically throughout the emulsion.

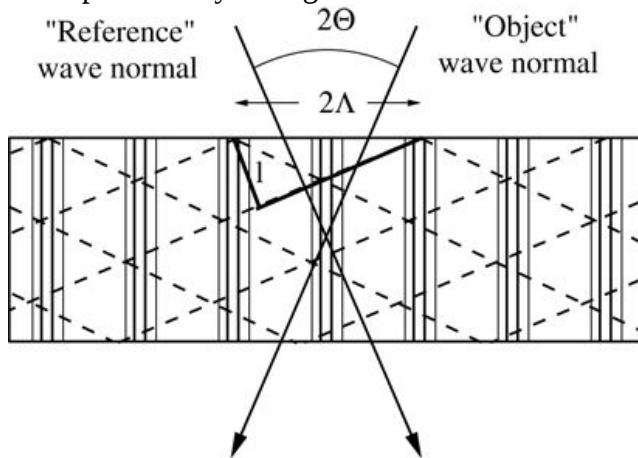


Figure 11.22
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.22 Recording an elementary hologram with a thick emulsion.

The illustration shows a rectangular bar with seven batches of vertical lines. Each batch consists of four lines. A slanting arrow labeled “Reference” wave normal starts slightly left from the midpoint of upper surface of the rectangular bar and point toward the bottom. A slanting arrow labeled “Object” wave normal starts slightly right from the midpoint of the upper surface of the rectangular bar and point toward the bottom. The arrows cross each other at the center. The angle between the two arrows is 20 degrees. Inside the rectangular bar an inverted obtuse triangle is shown with its base resting on the upper surface. The length of the short side of triangle is 1. The length of the base of the triangle is 2 fringe period.

Describing the three-dimensional interference process mathematically, the complex amplitudes of the two waves can be represented by

$$U_r(\vec{r}) = A \exp(jk_r \cdot \vec{r}), \quad U_o(\vec{r}) = a \exp(jk_o \cdot \vec{r}),$$

$$U_r(\vec{r}) = A \exp\left(j \vec{k}_r \cdot \vec{r}\right)$$

$$U_o(\vec{r}) = a \exp\left(j \vec{k}_o \cdot \vec{r}\right),$$

(11-45)

where \vec{k}_r and \vec{k}_o are the wave vectors of the reference and object waves, respectively, and \vec{r} is a position vector with components (x, y, z) . The intensity distribution that results from superimposing these waves is given by

$$I(\vec{r}) = |A|^2 + |a|^2 + 2|A||a|\cos[\vec{k}_r \cdot \vec{k}_o + \phi],$$

$$(11-46)$$

where ϕ is the phase difference between the phasors A and a .

At this point it is convenient to define a grating vector $K \rightarrow \vec{K}$ as the difference of the two wave vectors,

$$K \rightarrow = k_r - k_o.$$

$$\vec{K} = \vec{k}_r - \vec{k}_o.$$

$$(11-47)$$

The vector $K \rightarrow \vec{K}$ has a magnitude that is $2\pi/\Lambda$, where Λ is the fringe period, and points in the direction of the difference between k_r and k_o . A pictorial representation of $K \rightarrow \vec{K}$ is given by the wave vector diagram shown in Fig. 11.23. From this figure we can deduce that the period Λ of the grating is given by

$$\Lambda = 2\pi|K \rightarrow| = \lambda 2\sin\Theta,$$

$$\Lambda = \frac{2\pi}{|\vec{K}|} = \frac{\lambda}{2\sin\Theta},$$

$$(11-48)$$

as asserted in an earlier section.

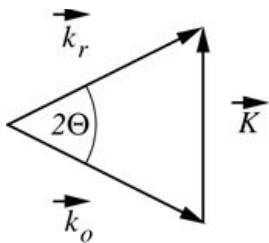


Figure 11.23

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.23 Wave vector diagram illustrating the length and direction of the grating vector.

The illustration shows three arrows arranged in the shape of a triangle with its base toward the right. The base arrow is labeled K vector and the upper side arrow of the triangle is labeled vector k_r and the lower side arrow of the triangle is labeled vector k_o . The angle between the two sides is approximately equal to 20 degrees.

If the photographic plate is developed, silver atoms will appear concentrated along the quasi-planes of high exposure, which we will call silver “platelets.” The distance between these platelets is the period Λ specified above.

11.7.2 Reconstructing Wavefronts from a Volume Grating

Suppose that we attempt to reconstruct the original object plane wave by illuminating the volume grating with a reconstruction plane wave. The question naturally arises as to what angle of illumination should be used to obtain a reconstructed object wave of maximum intensity. To answer this question, we may regard each platelet of high silver concentration as a partially reflecting mirror, which diverts part of the incident wave according to the usual laws of reflection and transmits part of the wave. If the plane-wave illumination is incident on the silver platelets at angle α , as shown in Fig. 11.24, then the reflected wave will travel in the direction satisfying the law of reflection. However, such reflections occur at all the platelets, and if the various reflected plane waves are to add *in phase*, then it is essential that the various path lengths traveled by waves reflected from adjacent platelets differ by precisely one optical wavelength.⁶ With reference to the figure, simple geometry shows that this requirement will be satisfied only if the angle of incidence satisfies the *Bragg condition*,

$$\sin\alpha = \pm \frac{\lambda}{2\Lambda}.$$

$$\sin\alpha = \pm \frac{\lambda}{2\Lambda}.$$

(11-49)

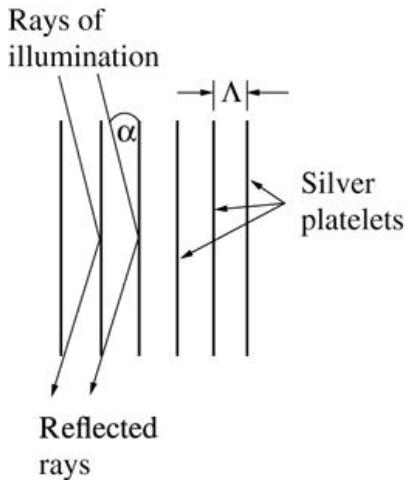


Figure 11.24

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.24 Reconstruction geometry.

The illustration shows 7 vertical lines. Two slanting lines labeled rays of illumination start from top and point toward the center of the second and third vertical lines from the left and then get reflected and point toward the bottom end of first and second lines, respectively and end with arrows. The lines are labeled reflected rays. The angle between the second ray of illumination and third vertical line is alpha. The first three vertical lines from the right are labeled silver platelets. The distance between the first and second silver platelet is labeled lambda.

Comparison of (11-48) and (11-49) shows that maximum intensity of the diffracted wave will be obtained only if

$$\alpha = \pm\Theta \pm (\pi - \Theta).$$

$$\alpha = \begin{cases} \pm\Theta \\ \pm(\pi - \Theta) \end{cases}.$$

(11-50)

This result is a very important one, for it indicates the condition necessary to obtain a reconstructed plane wave of maximum intensity. Actually, this equation defines a cone of reconstruction angles that will yield the desired results. It is only necessary that the wave vector

\vec{k}_p of the reconstruction wave be inclined at angle Θ to the planes of the silver platelets. [Figure 11.25](#) shows the allowable cones of incident and diffracted wave vectors. As the reconstruction (or “playback”) wave vector \vec{k}_p moves around circle shown, the wave vector of the diffracted light \vec{k}_i moves with it such that the \vec{k} -vector diagram always closes. Note that it is possible to interchange the roles of \vec{k}_p and \vec{k}_i in this figure and still satisfy the Bragg condition.⁷ The fact that an entire cone of incident \vec{k} -vectors will diffract strongly from a given volume grating is referred to as “Bragg degeneracy.”

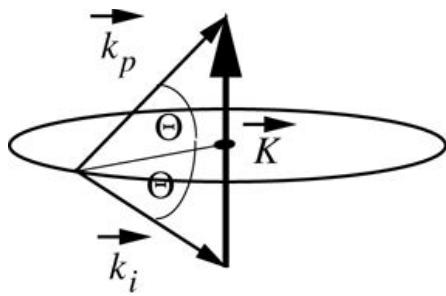


Figure 11.25

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.25 Cone of incident wave vectors that satisfies the Bragg condition.

The illustration shows an oval shaped structure. A triangle shaped vector diagram is drawn over the oval shaped structure with the meeting point of the sides lying at a point on the left bottom of the oval. The base of the triangle is a thick shaded line toward the right and labeled vector K. The upper side of the triangle is labeled vector k_p and the lower side of the triangle is labeled vector k_i . A horizontal line from the center of base of the triangle points toward the meeting point of two sides. The angle between the two sides and the center line is marked theta each.

11.7.3 Fringe Orientations for More Complex Recording Geometries

The discussion above has focused on the simplest case of a hologram formed by the interference of two plane waves that have equal but opposite angles with respect to the normal to the surface of the recording medium. This case is less restrictive than it might appear, for it is possible to consider two arbitrary wavefronts to be locally planar and their interference to be similar to the interference of two plane waves in any local region, albeit with a different tilt angle with respect to the recording medium than has been assumed here. In all such cases the general principle governing the orientation of the fringe pattern is the same as for the simple case examined: *the fringes formed in the recording medium are always oriented locally to bisect the angle between the two interfering waves within the medium.*⁸

Application of the principle stated above allows one to accurately predict the fringe structures expected in any given case. [Figure 11.26](#) shows several cases of interest, including plane waves interfering to produce slant fringes, plane waves and spherical waves interfering, and waves interfering from opposite sides of the recording medium, a case that leads to a *reflection hologram*.

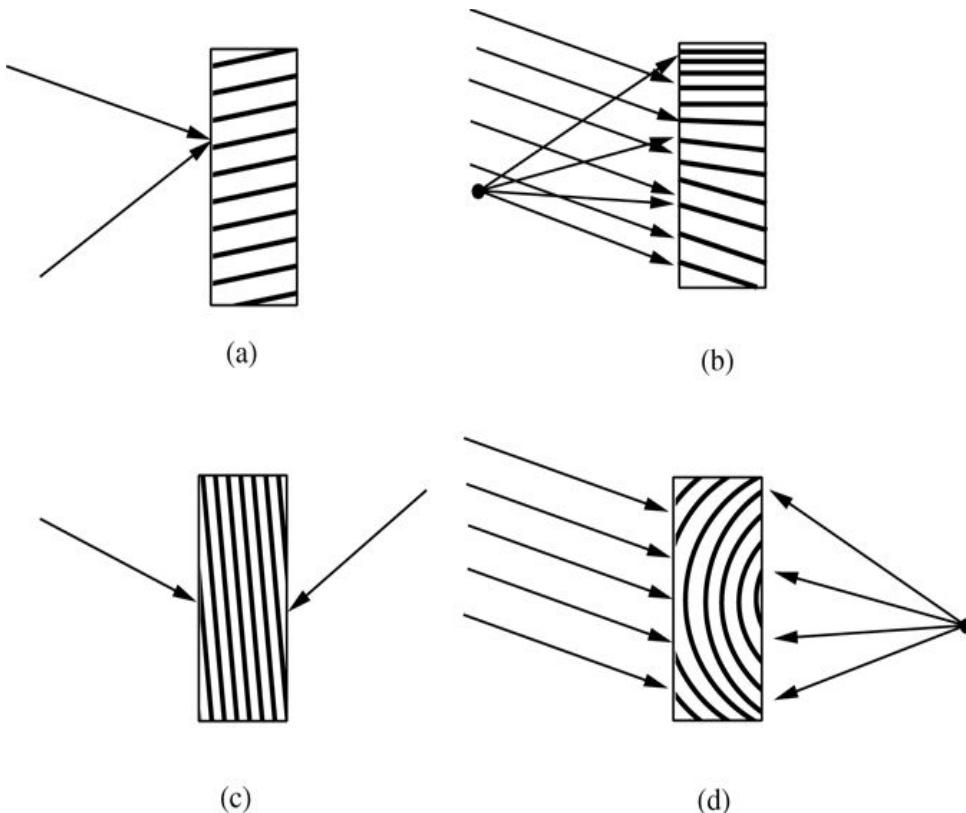


Figure 11.26

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 11.26 Orientation of interference fringes within a recording medium. (a) Two plane waves forming slant fringes, (b) a plane wave and a spherical wave, (c) two plane waves impinging from opposite sides of the emulsion, and (d) a plane wave and a spherical wave impinging from opposite sides of the recording medium.

Illustration a

Illustration a shows a vertical rectangular bar with slanting horizontal lines inside. One slanting arrow from upper left and another one from bottom left point toward the left side of the rectangular bar at a point slightly above the center.

Illustration b shows a vertical rectangular bar with horizontal lines. The horizontal lines at the upper section are close to one another whereas at the bottom section slanting horizontal lines are at a considerable distance from one another. A dark spot is shown on the extreme left and four arrows, one below the other from the dark spot point toward the left side of the rectangular bar and another 6 slanting arrows, one below the other, point toward the left side of the vertical rectangular bar.

Illustration c shows a vertical rectangular bar with slanting vertical lines inside. A slanting arrow from extreme left points toward the left side of the vertical rectangular bar at a point slightly below the center and another arrow from extreme right points toward the right side of the rectangular bar at a point slightly below the center.

Illustration d shows a vertical rectangular bar with concentric semi circles facing the left inside. A dark spot is shown on the extreme right and four arrows, one below the other, from the dark spot point toward the right side of the rectangular bar and another 6 slanting arrows from extreme right, one below the other, point toward the left side of the vertical rectangular bar.

Another general case worth considering is that of two equiphase point sources, perhaps at different distances from the recording medium, generating interference fringes. The fringe peaks form along surfaces for which the difference of distances from the two point sources is an integer multiple of an optical wavelength. Such surfaces are hyperboloids, and any slice through the surface shows hyperboloidal lines of fringe peaks, as shown in Fig. 11.27. Note that if our distance from the two sources is much greater than their separation, and if we examine the fringes over a region that is small compared with the distance from the sources, the fringes will appear to be of an approximately constant spatial frequency determined by the angular separation of the sources, viewed from the recording plane. Notice also that the fringe spacing is smallest when the spherical waves are approaching one another from opposite directions. When the angle between reference and object reaches 180° , (11-48) implies that the fringe spacing is $\lambda_o/2n$, where n is the refractive index of the recording medium.

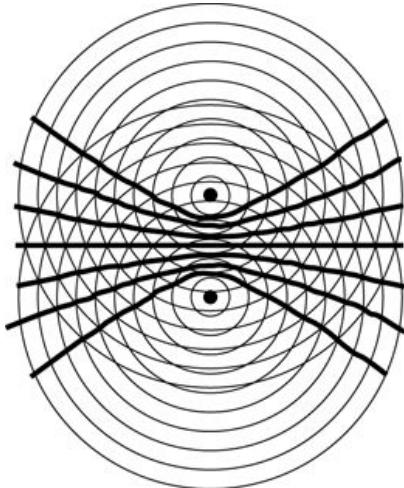


Figure 11.27

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.27 Slice through the hyperboloids of fringe maxima for the case of two point sources. The dark lines represent interference fringes, while the lighter lines are the wavefronts.

The illustration shows a dark spot and surrounding it 10 concentric circles are shown. At the lower portion, at the center of the fifth circle from the bottom, another dark spot is shown and surrounding the spot, 10 concentric circles are shown such that the top portion of the second set of circles overlap the lower portion of first set of circles. At the overlapped section, 7 thick shaded lines are shown. The center line is straight and three lines above the center line bend at the center in the shape of V and three lines below the center line also bend toward the center in the shape of inverted V.

11.7.4 Gratings of Finite Size

The theoretical treatments of volume gratings are, for simplicity, often based on the assumption that the grating is infinite in extent. Such is never the case in practice, of course, so it is important to understand the consequences of finite grating size. Such gratings are confined spatially to the finite volume occupied by the recording medium, and usually that volume has a highly asymmetric

shape.⁹ For example, photographic emulsions are usually very much thinner than their lateral extent.

We now present an analysis which is at best a rough approximation to the full description of the effects of finite grating size. The approach is an approximation primarily because it neglects the effects of absorption on the readout beam, but it does provide some physical intuition regarding some of the properties of thick holograms.

For this purpose we use three-dimensional Fourier analysis to express a finite-size grating as a superposition of a multitude of infinite-size gratings, each having a different \mathbf{K} vector.

Suppose that $g(\vec{r})$ represents the local refractive index of a volume phase grating, or the local absorption coefficient of a volume amplitude grating. It is convenient to represent g^g with a three-dimensional Fourier integral,

$$g(\vec{r}) = \iiint_{-\infty}^{\infty} G(\mathbf{K}) e^{j\mathbf{K} \cdot \vec{r}} d^3 K,$$

$$g(\vec{r}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} G(\vec{K}) e^{j\vec{K} \cdot \vec{r}} d^3 \vec{K},$$

(11-51)

where $G(\mathbf{K})$ describes the amplitude and phase of \mathbf{K} -vector components contained in g^g , and $d^3 K = dK_X dK_Y dK_Z$.

In the special case of a sinusoidal fringe of constant grating vector $\mathbf{K} \rightarrow g$, the form of g^g is

$$g(\vec{r}) = 1 + m \cos(\mathbf{K}_g \cdot \vec{r} + \phi_o),$$

$$g(\vec{r}) = [1 + m \cos(\mathbf{K}_g \cdot \vec{r} + \phi_o)] \operatorname{rect}\frac{x}{X} \operatorname{rect}\frac{y}{Y} \operatorname{rect}\frac{z}{Z},$$

(11-52)

where ϕ_o is an unimportant spatial phase of the grating, m is the modulation of the grating, and the recording medium has been assumed to have dimensions X, Y, Z in the three rectangular coordinate directions.

The grating-vector spectrum of the above spatially bounded fringe is easily found to be

$$\begin{aligned} G(\mathbf{K}) &= \delta(\mathbf{K}) + 12\delta(\mathbf{K} - \\ &\quad \mathbf{K}_g) + 12\delta(\mathbf{K} + \mathbf{K}_g) * XYZ \operatorname{sinc}\frac{XK_X}{2\pi} \operatorname{sinc}\frac{YK_Y}{2\pi} \operatorname{sinc}\frac{ZK_Z}{2\pi}. \end{aligned}$$

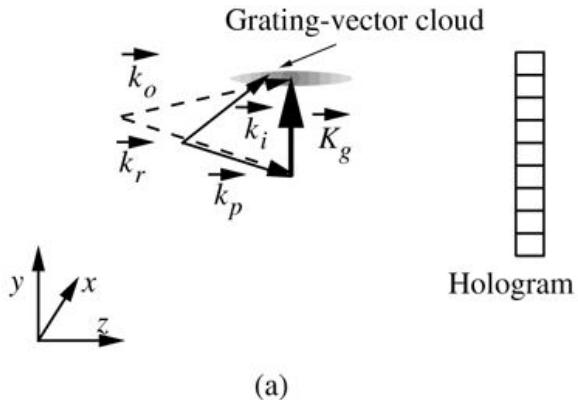
$$G(\vec{K}) = [\delta(\vec{K}) + \frac{1}{2}\delta(\vec{K} - \vec{K}_g) + \frac{1}{2}\delta(\vec{K} + \vec{K}_g)] * XYZ \operatorname{sinc}\frac{XK_X}{2\pi} \operatorname{sinc}\frac{YK_Y}{2\pi} \operatorname{sinc}\frac{ZK_Z}{2\pi}.$$

(11-53)

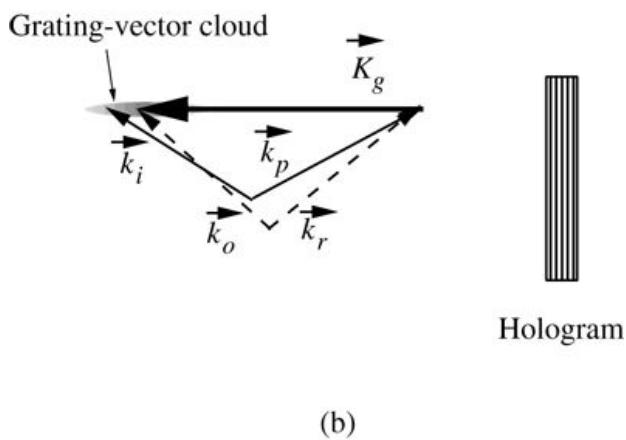
The result of this three-dimensional convolution is a blurring of the grating-vector tip into a continuum of grating vectors surrounding the ideal location of the tip of the grating vector for an infinite grating. This blurring operation then leads to the possibility that the \vec{k} -vector triangle required by the Bragg effect can be closed in many different ways, perhaps at some cost in terms of the strength of the diffracted wave. If the \vec{k} -vector triangle closes within the central portion of the primary lobe of the three-dimensional sinc function above, then the diffraction efficiency should still be near its maximum possible value.

[Figure 11.28](#) shows the effects of the grating-vector cloud on \vec{k} -vector closure in two different cases. In all cases, the angle of illumination of the grating is assumed to be identical with the angle of the reference wave used during the recording process. In [11.28\(a\)](#), the grating has been recorded by plane waves incident from the same side of the recording medium, which is assumed much thinner in the z direction than in the other directions. Since the grating-vector cloud is extended in the direction normal to the recording surface, this geometry is quite tolerant to

changes of the wavelength of the reconstruction beam (i.e. the length of $\vec{k} \rightarrow \vec{p}$) relative to that used during recording but less tolerant to changes of the direction of the reconstruction beam. In [11.28\(b\)](#), the object and reference waves have come from opposite sides of the emulsion, producing a grating that operates by reflection rather than transmission. In this case the grating-vector blur extends predominantly in a direction along the grating-vector direction. This orientation leads to tolerance with respect to the angle of illumination and less wavelength tolerance than in the previous case. The degree of tolerance to angular or wavelength changes depends, in both cases, on the thickness of the grating as well as on the period of the fringes, but it is generally true that transmission gratings are more tolerant to wavelength changes than are reflection gratings, and reflection gratings are more tolerant to angle changes than are transmission gratings.



(a)



(b)

Figure 11.28

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.28 Grating-vector clouds and their effect on closing the \vec{k} -vector triangle. The dotted vectors correspond to \vec{k} -vectors when the grating is recorded, and the solid vectors correspond to the \vec{k} -vectors when reconstruction takes place. Changes of the lengths of the \vec{k} -vectors correspond to reconstruction at a different wavelength than was used for recording. In part (a), a change of the length of $\vec{k} \rightarrow p$ does not prevent closure of the \vec{k} -vector diagram. In part (b), a change of the angle of $\vec{k} \rightarrow p$ does not prevent closure.

Illustration a shows a graphical representation at the extreme left. The vertical axis is labeled y and horizontal axis is labeled z . A slanting line starts from 0 and points rightward and is labeled x . To the right of the graph is a vector diagram which is in the shape of a triangle with the base toward the right. The upper and lower sides of the triangle are labeled vector k_i and vector k_p , respectively. Another dotted triangle with the same base and the sides lengthier than the previous triangle is shown. The upper and lower sides are labeled vector k_0 and vector k_r , respectively. A shaded oval portion is at the upper end of the base and is labeled grating-vector cloud. To the right, is a thin vertical rectangular bar with horizontal partitions and is labeled hologram.

Illustration b shows a vector diagram in the shape of inverted triangle. The base of the triangle is thick and is labeled vector K_g . The left and right sides are labeled vector k_i and vector k_p ,

respectively. The left side doesn't touch the base and moves slightly outward. Another dotted triangle is shown with the same base vector K_g . The height of the dotted triangle is more when compared to the other triangle. The left and right sides of the triangle are labeled vector k_0 and vector k_r , respectively. A shaded portion is shown near the left corner and is labeled grating-vector cloud. To the right, is a thin vertical rectangular bar with vertical lines inside and is labeled hologram.

A more exact understanding of the tolerance of volume gratings to the angles and wavelengths of illumination requires a more sophisticated analysis. An example of such an analysis is the coupled mode theory that follows.

11.7.5 Diffraction Efficiency—Coupled Mode Theory

It is extremely important to know the diffraction efficiencies that can be theoretically expected from thick holograms of various types. To find these efficiencies, as well as the tolerance of various types of gratings to the particular angle and wavelength used for reconstruction, it is necessary to embark on a significant analysis. Many methods for calculating these quantities have been found. Most entail some form of approximation, and some are more accurate than others. For an in-depth discussion of a variety of different methods, see [324]. However, the most widely used method is the coupled mode theory, pioneered by [Kogelnik \[202\] \[203\]](#) in holography. This is the approach that we shall use here, although we shall follow [\[160\], Chapter 4](#), most closely. See also [\[128\]](#) for another useful reference.

The general geometry is illustrated in [Fig. 11.29](#). In this general case, the grating within the emulsion is tilted at angle ψ with respect to the normal to the surface of the recording medium and has grating period $\Lambda = 2\pi/K$. The reconstruction wave is incident at angle θ to that same normal.¹⁰

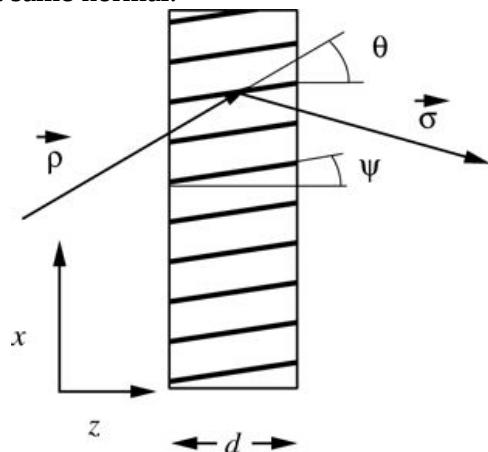


Figure 11.29
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.29 Geometry for analysis of a thick hologram.

The illustration shows a vertical rectangular bar of width d with 10 slanting horizontal lines inside. A slanting line labeled vector ρ passes through the vertical rectangular bar intersecting the third slanting horizontal line at its midpoint. A small straight line is drawn on the right side from the point where the third slanting line ends on the right side. The angle between the horizontal line and slanting line that passes through the vertical rectangular bar is θ . A line facing downward is

drawn from the intersection point of slanting line that passes through the vertical horizontal bar and the third horizontal line inside the bar and is labeled vector sigma. A straight horizontal line is drawn from the left end of the fifth slanting line inside the vertical rectangular bar and the angle between the horizontal line and the slanting line is psi. To the extreme left is a graph, with horizontal axis labeled z and vertical axis labeled x.

The Analysis

The analysis begins with the scalar wave equation,

$$\nabla^2 U + k^2 U = 0,$$

$$\nabla^2 U + k^2 U = 0,$$

(11-54)

valid in a source free region for monochromatic light. The wave number in the most general case is complex-valued, $k = (2\pi n / \lambda_0) + j\alpha$, where α is the absorption constant and λ_0 is the vacuum wavelength. The refractive index n and the absorption constant α within the grating are assumed to vary in sinusoidal fashions according to

$$n = n_0 + n_1 \cos K \cdot r \rightarrow \alpha = \alpha_0 + \alpha_1 \cos K \cdot r \rightarrow ,$$

$$\begin{aligned} n &= n_0 + n_1 \cos \vec{K} \cdot \vec{r} \\ \alpha &= \alpha_0 + \alpha_1 \cos \vec{K} \cdot \vec{r} , \end{aligned}$$

(11-55)

where $r \rightarrow \sim (x, y, z)$ and $K \rightarrow \vec{K}$ is the grating vector. The hologram is assumed to lie with its faces parallel to the (x, y) plane and to be of thickness d in the z dimension.

A number of assumptions are needed for simplification of the problem of solving the wave equation. First it is assumed that the hologram is thick enough that only two waves need be

considered within the grating. One is the reconstruction or playback wave $U_p(r \rightarrow)$, which is gradually depleted by diffraction and absorption, and the other is the first-order Bragg-matched grating order $U_i(r \rightarrow)$. We assume that the total field within the grating is composed of a sum of these two waves, and we accordingly write that field as

$$U(r \rightarrow) = U_p(r \rightarrow) + U_i(r \rightarrow) = R(z) \exp(j\vec{\rho} \cdot \vec{r}) + S(z) \exp(j\vec{\sigma} \cdot \vec{r}),$$

$$\begin{aligned} U(\vec{r}) &= U_p(\vec{r}) + U_i(\vec{r}) \\ &= R(z) \exp(j\vec{\rho} \cdot \vec{r}) + S(z) \exp(j\vec{\sigma} \cdot \vec{r}), \end{aligned}$$

(11-56)

where the symbols $\rho \rightarrow \vec{\rho}$ and $\sigma \rightarrow \vec{\sigma}$ are conventionally used in place of what would be $k \rightarrow p$, \vec{k}_p and $k \rightarrow i \vec{k}_i$, respectively, in our previous notation. We assume that the wave vector $\rho \rightarrow \vec{\rho}$ of R is that of the playback wave in the absence of coupling, and that the wave vector $\sigma \rightarrow \vec{\sigma}$ of the diffracted wave is given by

$$\sigma \rightarrow = \rho \rightarrow - K \rightarrow .$$

$$\vec{\sigma} = \vec{\rho} - \vec{K} .$$

(11-57)

In addition, it is assumed that absorption in a distance of one wavelength is small and that the variations of the refractive index are small compared to its mean,

$$n_0 k_o \gg \alpha_0 n_0 k_o \gg \alpha_1 n_0 \gg n_1 ,$$

$$\begin{aligned} n_0 k_o &\gg \alpha_0 \\ n_0 k_o &\gg \alpha_1 \\ n_0 &\gg n_1 , \end{aligned}$$

(11-58)

where k_o is the vacuum wave number, $k_o = 2\pi/\lambda_o$.

It is now possible to expand and simplify k^2 for use in the wave equation as follows¹¹:

$$k^2 = k_o^2 + n_1 \cos K \cdot r + j\alpha_0 + \alpha_1 \cos K \cdot r \approx B^2 + 2jB\alpha_0 + 4\kappa B \cos K \cdot r ,$$

$$\begin{aligned} k^2 &= \left[k_o \left(n_0 + n_1 \cos \vec{K} \cdot \vec{r} \right) + j \left(\alpha_0 + \alpha_1 \cos \vec{K} \cdot \vec{r} \right) \right]^2 \\ &\approx B^2 + 2jB\alpha_0 + 4\kappa B \cos \vec{K} \cdot \vec{r} , \end{aligned}$$

(11-59)

where liberal use of the approximations (11-58) has been made, $B = k_o n_0$, and κ is the *coupling constant*, given by

$$\kappa = 12k_o n_1 + j\alpha_1 .$$

$$\kappa = \frac{1}{2}(k_o n_1 + j\alpha_1) .$$

(11-60)

The next step is to substitute the assumed solution (11-56) and the expression for k^2 above into the wave equation (11-54). During the substitution, $R(z)$ and $S(z)$ are

assumed to be slowly varying functions of z so that their second derivatives can be dropped, the term $\cos K \rightarrow \cdot r \rightarrow \cos \vec{K} \cdot \vec{r}$ is expanded into its two complex-exponential components, and $\sigma \rightarrow \vec{\sigma}$ is replaced according to (11-57). Terms with wave vectors $\sigma \rightarrow -K \rightarrow = \rho \rightarrow -2K \rightarrow \vec{\sigma} - \vec{K} = \vec{\rho} - 2\vec{K}$ and $\rho \rightarrow +K \rightarrow = \sigma \rightarrow +2K \rightarrow \vec{\rho} + \vec{K} = \vec{\sigma} + 2\vec{K}$ are dropped, since they correspond to propagation directions that are far from satisfying the Bragg condition. Finally, equating the sum of all terms multiplying $\exp[j\rho \rightarrow \cdot r \rightarrow]$ to zero and similarly for the sum of all terms multiplying $\exp[j\sigma \rightarrow \cdot r \rightarrow]$, we find that $R(z)$ and $S(z)$ must individually satisfy the following equations in order for the wave equation to be satisfied:

$$c_R dR dz + \alpha_0 R = j\kappa S c_S S dS dz + \alpha_0 - j\zeta S = j\kappa R,$$

$$c_R \frac{dR}{dz} + \alpha_0 R = j\kappa S$$

$$c_S \frac{dS}{dz} + (\alpha_0 - j\zeta) S = j\kappa R,$$

(11-61)

where ζ is called the “detuning parameter,” given by

$$\zeta = B^2 - |\vec{\sigma}|^2 / (2B),$$

$$\zeta = \frac{B^2 - |\vec{\sigma}|^2}{2B},$$

(11-62)

and the quantities c_R and c_S are given by

$$c_R = \rho Z B = \cos \theta, \quad c_S = \sigma Z B = \cos(\theta - 2\psi),$$

$$c_R = \frac{\rho_Z}{B} = \cos \theta$$

$$c_S = \frac{\sigma_Z}{B} = \cos(\theta - 2\psi),$$

(11-63)

where θ and ψ are defined in Fig. 11.29.

The quantity ζ is a measure of the “Bragg mismatch” of the reconstructed wave, and deserves further discussion. Equation (11-57) is a statement of the Bragg matching condition. Using this equation, we see

$$B^2 - |\sigma \rightarrow |^2 = B^2 - (\rho \rightarrow -K \rightarrow) \cdot (\rho \rightarrow -K \rightarrow) = B^2 - |\rho \rightarrow|^2 + 2\rho \rightarrow \cdot K \rightarrow - K^2 = 2\rho K \cos(\psi + \pi/2 - \theta) - K^2,$$

$$\begin{aligned} B^2 - |\vec{\sigma}|^2 &= B^2 - \left(\vec{\rho} \cdot \vec{K} \right) \cdot \left(\vec{\rho} \cdot \vec{K} \right) \\ &= B^2 - \left| \vec{\rho} \right|^2 + 2 \vec{\rho} \cdot \vec{K} - K^2 \\ &= 2\rho K \cos(\psi + \pi/2 - \theta) - K^2 \\ &= 2\rho K \sin(\theta - \psi) - K^2, \end{aligned}$$

(11-64)

where $K = |K \rightarrow| = |\vec{K}|$ and $\rho = |\rho \rightarrow| = B = k_o n_o$. Thus

$$\zeta = B^2 - |\sigma \rightarrow|^2 / 2B = K \left[\sin(\theta - \psi) - \frac{K}{2k} \right].$$

$$\zeta = \frac{B^2 - |\vec{\sigma}|^2}{2B} = K \left[\sin(\theta - \psi) - \frac{K}{2k} \right].$$

(11-65)

Note that the quantity in brackets will be zero when the Bragg condition is satisfied. Consider now a departure from the Bragg matched conditions caused by a combination of a small mismatch in the illumination angle $\theta' = \theta_B - \Delta\theta$ and a small mismatch in the wavelength $\lambda' = \lambda - \Delta\lambda$. Substitution into (11-65) yields the following expression for the detuning parameter in terms of the angular and wavelength mismatches:

$$\zeta = K \Delta\theta \cos(\theta_B - \psi) - \Delta\lambda / 2\Lambda.$$

$$\zeta = K \left[\Delta\theta \cos(\theta_B - \psi) - \frac{\Delta\lambda}{2\Lambda} \right].$$

(11-66)

It can now be clearly seen that mismatch due to wavelength error grows as the grating period Λ shrinks, and therefore wavelength selectivity will be maximum for counterpropagating object and reference beams, which produce a reflection hologram. Selectivity to angular mismatch can be shown (see [Prob. 11-10](#)) to be maximum when the reference and object beams are separated by an angle of 90° . With the help of (11-66), we can estimate the value of the detuning parameter for any combination of angular or wavelength mismatch.

Returning to the coupled wave equations, note that the equation for S contains a driving or forcing term on the right that depends on the incident wave R . It is this term that leads to a transfer of energy from the incident wave to the diffracted wave. If the coupling constant κ is zero, no such coupling will occur. The detuning parameter ζ , if sufficiently large, will swamp the driving term in R , leading to a spoiling of the coupling phenomena due to phase mismatch.

through the coupling region. In addition, the equation for the amplitude of the incident wave contains a driving term that depends on the diffracted wave, leading to coupling from that wave back into the incident wave.

For all specific solutions discussed in the following, we assume that the grating is unslanted. For a transmission grating, this implies that $\psi=0^\circ$ while for a reflection grating, $\psi=90^\circ$.

Solution for a Thick Phase Transmission Grating

For a pure phase grating we have $\alpha_0=\alpha_1=0$. For a transmission geometry, the boundary conditions to be applied to the differential equations (11-61) are $R(0)=1$ and $S(0)=0$. The solution for the diffracted wave S at the exit of the grating ($z=d$) then takes the form

$$S(d)=je^{j\chi}\sin\Phi\sqrt{1+\chi^2/\Phi^2},$$

$$S(d)=je^{j\chi}\frac{\sin(\Phi\sqrt{1+\chi^2}/\Phi^2)}{\sqrt{1+\chi^2}/\Phi^2},$$

(11-67)

where¹²

$$\Phi=\pi n_1 d \lambda \cos\theta \chi = \zeta d \lambda \cos\theta = K d \lambda \cos\theta \Delta\theta \cos(\theta - \psi) - \Delta\lambda 2\Lambda.$$

$$\begin{aligned} \Phi &= \frac{\pi n_1 d}{\lambda \cos\theta} \\ \chi &= \frac{\zeta d}{2 \cos\theta} = \frac{K d}{2 \cos\theta} \left[\Delta\theta \cos(\theta - \psi) - \frac{\Delta\lambda}{2\Lambda} \right]. \end{aligned}$$

(11-68)

The diffraction efficiency of the grating is given by

$$\eta=|S(d)|^2|R(0)|^2=\sin^2\Phi\sqrt{1+\chi^2}/\Phi^2.$$

$$\eta=\frac{|S(d)|^2}{|R(0)|^2}=\frac{\sin^2(\Phi\sqrt{1+\chi^2}/\Phi^2)}{1+\chi^2/\Phi^2}.$$

(11-69)

When the grating is illuminated at the Bragg angle with the same wavelength used during recording, the parameter χ is identically zero, yielding for the diffraction efficiency

$$\eta_B=\sin^2\Phi.$$

$$\eta_B=\sin^2\Phi.$$

(11-70)

The diffraction efficiency is seen to increase initially with increasing thickness of the grating, reach a maximum of 100%, fall to zero, rise to 100%, etc., periodically. Since the grating is lossless, the power in the undiffracted wave oscillates similarly but with minima and maxima interchanged. The first maximum of 100% for the diffraction efficiency is reached when $\Phi=\pi/2$
 $\Phi = \pi/2$, or when

$$d\cos\theta = \lambda/2n_1.$$

$$\frac{d}{\cos\theta} = \frac{\lambda}{2n_1}.$$

(11-71)

[Figure 11.30](#) shows the oscillations of the diffracted power and the undiffracted power as a function of the parameter Φ .

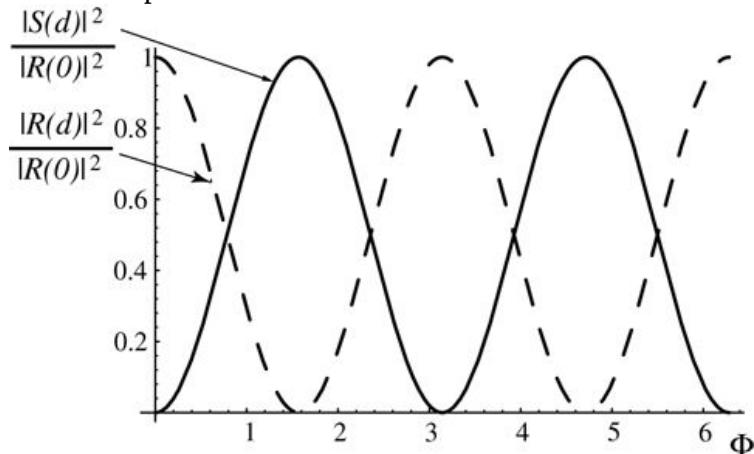


Figure 11.30

Goodman, *Introduction to Fourier Optics*, 4e,
 © 2017 W. H. Freeman and Company

Figure 11.30 Normalized intensities of the diffracted and undiffracted waves as a function of Φ for the Bragg matched case.

The graph has its horizontal axis labeled Φ with markings starting from 1 to 6 with equal increments of 1. The marking on the vertical axis starts from 0.2 and ends at 1 with equal increments of 0.2. The curve labeled square of determinant of $S(d)$ divided by square of determinant of $R(0)$ is a sine wave that starts at 0 and gradually increases to 1 on vertical axis corresponding to 1.5 on horizontal axis and decreases to touch the horizontal axis at 3.1 and increases again to 1 on vertical axis corresponding to 4.5 on horizontal axis and again decreases and touches the horizontal axis at 6.2. The curve labeled square of determinant of $R(d)$ divided by square of determinant of $R(0)$ is a sine wave represented as dashed line which starts at 1 and gradually decreases and touches horizontal axis at 1.5 and increases to 1 on vertical axis corresponding to 3 on horizontal axis and again decreases and touches horizontal axis at 4.4 and increases again to reach 1 on vertical axis corresponding to 6.1 on horizontal axis. The values of the graph mentioned above are approximate.

When the grating is illuminated off of the Bragg angle or with a different wavelength than was used during recording, the parameter χ is nonzero. [Figure 11.31](#) shows a three-dimensional plot illustrating efficiency as a function of the both Φ and χ . It can be seen that for any fixed value of Φ , an increase in χ leads to a loss of diffraction efficiency, although oscillations of a diminishing magnitude occur for some values of Φ . This figure is useful when either Φ or χ is fixed and we wish to understand the effect of changing the other parameter. Note, however, that both parameters are proportional to the grating thickness d , so if the behavior as a function of thickness is of interest, a slice through the surface at some angle with respect to the Φ axis is required.

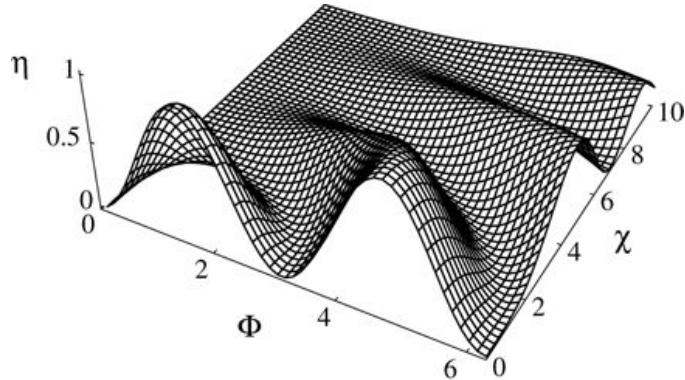


Figure 11.31

Goodman, *Introduction to Fourier Optics*, 4e,

© 2017 W. H. Freeman and Company

Figure 11.31 Diffraction efficiency of a thick phase transmission grating when Bragg mismatch is present.

The illustration shows a three dimensional plot with two axes in the shape of the alphabet “V” with the left axis labeled phi and the right axis labeled chi and a straight line begins at the left end of V which is the third axis. The third axis is labeled eta. A square shaped grid with two sides resting on the V shaped structure is shown. The side of the square corresponding to psi rises to .8 on the eta axis which corresponds to 1 on the axis labeled phi and 6 on the chi axis. It touches the phi axis at 3 and again increases to 0.5 on eta axis corresponding to 5 on phi axis and 5 on chi axis and decreases to touch 0 on psi axis. The side of square corresponding to chi axis shows a gradual increase starting at 2 on chi axis and reaches the peak at 6 on chi axis which corresponds to 0.8 on eta axis and 2 on phi axis and decreases and touches chi axis at 7 and again increases to reach 0.7 on eta axis corresponding to 1.9 on chi axis. The values mentioned above are approximate.

Solution for a Thick Amplitude Transmission Grating

For an unslanted amplitude grating, the index modulation n_1 is zero and α_1 is nonzero. The appropriate boundary conditions are the same as for the phase transmission grating, $R(0)=1$ $R(0) = 1$ and $S(0)=0$. The solution for the diffracted amplitude at the grating output is now given by

$$S(d) = -\exp - \alpha_0 d \cos \theta e^{j\chi} \sinh \Phi a_1 - \chi^2 / \Phi a_2 + \chi^2 / \Phi a_2,$$

$$S(d) = - \exp\left(-\frac{\alpha_0 d}{\cos\theta}\right) e^{j\chi} \frac{\sinh\left(\Phi_a \sqrt{1 - \chi^2} / \Phi_a^2\right)}{\sqrt{1 - \chi^2} / \Phi_a^2},$$

(11-72)

where \sinh is a hyperbolic sine function,

$$\Phi_a = \alpha_1 d / 2\cos\theta,$$

$$\Phi_a = \frac{\alpha_1 d}{2\cos\theta},$$

(11-73)

and again

$$\chi = \zeta d / 2\cos\theta.$$

$$\chi = \frac{\zeta d}{2\cos\theta}.$$

(11-74)

For Bragg matched conditions, $\chi=0$ and the diffraction efficiency is given by

$$\eta_B = \exp\left(-2\alpha_0 d \cos\theta \sinh 2\alpha_1 d / 2\cos\theta\right).$$

$$\eta_B = \exp\left(-\frac{2\alpha_0 d}{\cos\theta}\right) \sinh^2\left(\frac{\alpha_1 d}{2\cos\theta}\right).$$

(11-75)

This solution is a product of two functions, the first of which simply represents the attenuation of light due to the average absorption coefficient α_0 as it propagates through the distance $d/\cos\theta$ in the hologram. The second term represents the rising effect of diffraction as the wave propagates through increasing thickness. The absorption can never be negative in a passive medium, and therefore the modulation of absorption can never exceed the average absorption, $\alpha_1 \leq \alpha_0$. Because of this constraint, the two terms balance one another in such a way that there is an optimum thickness where diffraction efficiency is maximized.

Diffraction efficiency will be maximized if the attenuation modulation is taken to have its

largest possible value, $\alpha_1 = \alpha_0$. Defining $\Phi_a' = \alpha_0 d / 2\cos\theta$, this maximum diffraction efficiency can be expressed as

$$\eta_B = \exp\left(-4\Phi_a'\right) \sinh^2\left(\Phi_a'\right)$$

(11-76)

under Bragg matched conditions, a plot of which is shown in [Fig. 11.32](#). This expression takes on a maximum value of 0.037 or 3.7% for $\Phi_a' = 0.55$.

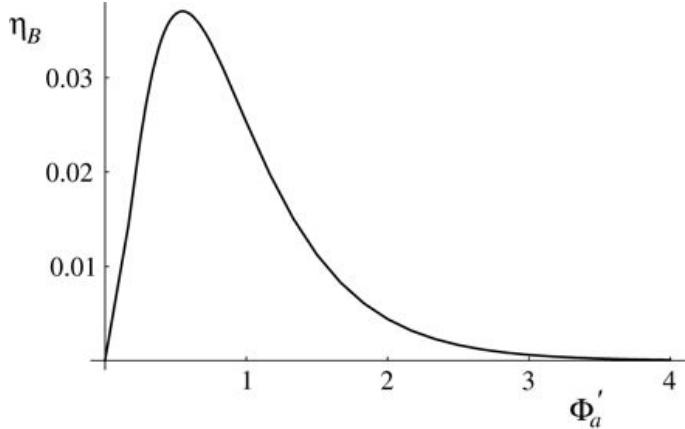


Figure 11.32

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.32 Maximum possible Bragg matched diffraction efficiency versus Φ_a' for a thick amplitude transmission grating. The graphical representation shows the horizontal axis labeled Φ_a' with marking from 1 to 4 with equal increments of 1 and vertical axis labeled η_B with marking from 0.01 to 0.03 with equal increments of 0.01. The curve starts at 0 and increases steeply to 0.04 on the vertical axis corresponding to 0.5 on the horizontal axis and shows a gradual increase and touches the horizontal axis at 3.5 and moves along the axis till 4 on the horizontal axis. The values mentioned above are approximate.

[Figure 11.33](#) shows a three-dimensional plot of the maximum possible diffraction efficiency (again, $\alpha_1 = \alpha_0$) with the quantity Φ_a' running from the left and the quantity χ running into the right, thus illustrating the effects of Bragg mismatch on the diffraction efficiency. Note that when Φ_a' is in the vicinity of the value that yields maximum diffraction efficiency, values of χ of the order of 2.5 will drive the diffraction efficiency to near zero.

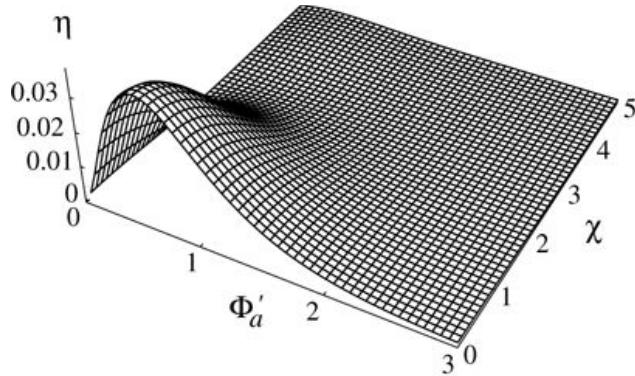


Figure 11.33

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.33 Diffraction efficiency of a thick amplitude transmission grating with Bragg mismatch.

The illustration shows a three dimensional plot with two axis in the shape of the alphabet “V” with the left axis labeled phi apostrophe a and the right axis labeled chi and a straight line begins at the left end of V which is the third axis. The third axis is labeled eta. The eta axis shows marking from 0 to 0.03 with equal increments of 0.01 and phi apostrophe a shows marking from 0 to 3 at equal intervals of 1 and chi axis shows marking from 0 to 5 with equal intervals of 1. A square shaped grid with two sides resting on the V shaped structure is shown. The side corresponding to phi apostrophe a raises to .03 on the eta axis corresponding to 0.5 on the axis labeled phi apostrophe a and 5 on the chi axis and then decreases gradually up to the point 3 on the phi apostrophe a. The side that corresponds to chi rests on the chi axis and doesn’t show any deviations.

Solution for a Thick Phase Reflection Grating

For a reflection grating, the grating planes run nearly parallel with the face of the recording medium. In what follows we assume for simplicity that the grating is unslanted, i.e. that the grating planes are exactly parallel to the surface ($\psi=90^\circ \psi = 90^\circ$). The boundary conditions change, now being $R(0)=1$ $R(0) = 1$ and $S(d)=0$ $S(d) = 0$ (i.e. the diffracted wave is now growing in the “backwards” direction. Again for a pure phase grating, $\alpha_0=\alpha_1=0$ $\alpha_0 = \alpha_1 = 0$. The solution of the coupled mode equations for the amplitude of the diffracted wave is now

$$S(0) = -j\chi\Phi + 1 - \chi^2\Phi^2 \coth \Phi - \chi^2\Phi^2 - 1,$$

$$S(0) = -j\left[-j\frac{\chi}{\Phi} + \sqrt{1 - \frac{\chi^2}{\Phi^2}} \coth \left(\Phi \sqrt{1 - \frac{\chi^2}{\Phi^2}} \right) \right]^{-1},$$

(11-77)

where Φ and χ are again given by (11-68) and \coth is a hyperbolic cotangent. The diffraction efficiency then becomes¹³

$$\eta = 1 + 1 - \chi^2\Phi^2 \sinh^2 \Phi - \chi^2\Phi^2 - 1.$$

$$\eta = \left[1 + \frac{1 - \frac{\chi^2}{\Phi^2}}{\sinh^2 \left(\Phi \sqrt{1 - \frac{\chi^2}{\Phi^2}} \right)} \right]^{-1}.$$

(11-78)

Under Bragg matched conditions, $\chi=0$, and the diffraction efficiency can be expressed as

$$\eta_B = \tanh 2\Phi,$$

$$\eta_B = \tanh^2 \Phi,$$

(11-79)

where \tanh is a hyperbolic tangent. [Figure 11.34](#) shows a plot of this diffraction efficiency versus the parameter Φ . As can be seen, the diffraction efficiency asymptotically approaches 100% as the parameter Φ increases.

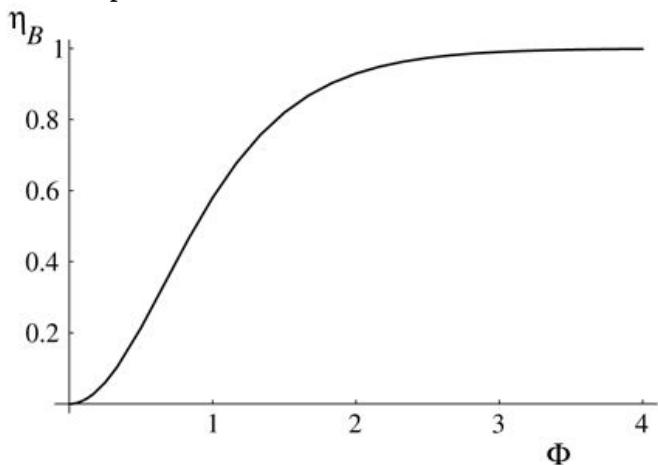


Figure 11.34

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.34 Diffraction efficiency of a thick Bragg matched phase reflection grating.

The graphical representation shows the horizontal axis labeled Φ with marking from 1 to 4 with equal increments of 1 and vertical axis labeled η_B with marking from 0.2 to 1 with equal increments of 0.2. The curve starts at 0 and increases gradually to 0.95 on the vertical axis corresponding to 2 on the horizontal axis and then stabilizes and moves parallel to the horizontal axis. The values of the graph mentioned above are approximate.

The behavior of the diffraction efficiency with Bragg mismatch is illustrated in the three-dimensional plot of [Fig. 11.35](#). In this figure we have interchanged the directions of Φ and χ from those of the other cases shown previously in order to better reveal the structure of the surface.

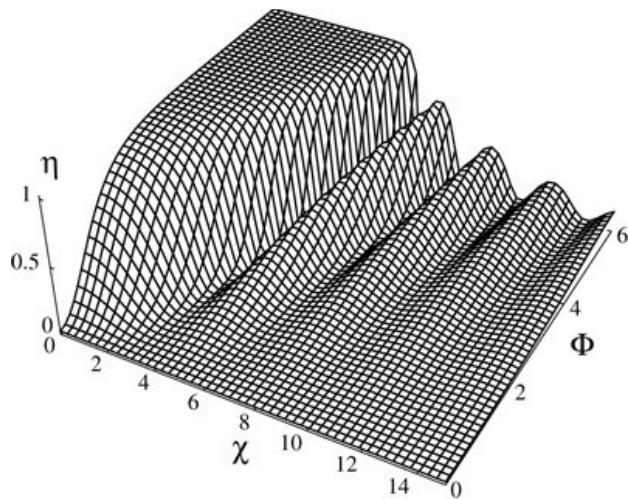


Figure 11.35

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.35 Diffraction efficiency of a thick phase reflection grating when Bragg mismatch is present.

The illustration shows a three dimensional plot with two axes in the shape of the alphabet “V” with the left axis labeled chi and the right axis labeled psi and a straight line begins at the left end of V which is the third axis. The third axis is labeled eta. The eta axis shows marking from 0 to 1 with equal increments of 0.5 and phi axis shows marking from 0 to 6 at equal intervals of 2 and chi axis shows marking from 0 to 14 with equal intervals of 2. A square shaped grid with two sides resting on the V shaped structure is shown. The side that corresponds to chi starts from a point above 1 on eta axis and moves parallel to chi axis up to the point 2 on chi axis and then shows a decrease up to the point 3 on chi axis and increases again to the point 1 on eta axis which corresponds to 4 on chi axis and then decreases at 5 on chi axis and increases to point 0.5 on eta axis which corresponds to 6 on chi axis and then decreases at 9 on chi axis and increases again to point 0.25 on eta axis which corresponds to 10 and the side corresponding to psi axis starts at 0 and increases gradually to 1 on the eta axis corresponding to 2 on psi axis and then moves parallel to psi axis up to the point 6. The values of the graph mentioned above are approximate.

Solution for a Thick Amplitude Reflection Grating

The final case of interest is that of a thick amplitude reflection grating. The boundary conditions are the same as for the previous case, but now $n_1=0$ and $n_2 \neq 0$ and diffraction is caused by variations α_1 of the absorption coefficient.

Solution of the coupled-wave equations now yields the following expression for the diffracted amplitude:

$$S(0) = -j\chi a \Phi a + 1 - \chi a^2 \Phi a^2 \coth \Phi a - \chi a^2 \Phi a^2 - 1$$

$$S(0) = -j \left[-j \frac{\chi a}{\Phi a} + \sqrt{1 - \frac{\chi^2 a^2}{\Phi^2 a^2}} \coth \left(\Phi a \sqrt{1 - \frac{\chi^2 a^2}{\Phi^2 a^2}} \right) \right]^{-1}$$

(11-80)

where Φ_a is again given by (11-73) and

$$\chi_a = \alpha_0 d \cos \theta + j \zeta d / 2 \cos \theta.$$

$$\chi_a = \frac{\alpha_0 d}{\cos \theta} + \frac{j \zeta d}{2 \cos \theta}.$$

(11-81)

Under Bragg matched conditions, ζ goes to zero. Again maximum diffraction efficiency will be achieved if the variations of absorption have their largest possible value, $\alpha_1 = \alpha_0$. Under these conditions, $\chi_a / \Phi_a = 2$, and

$$\eta_B = 2 + 3 \coth(3\Phi_a) - 2$$

$$\eta_B = [2 + \sqrt{3} \coth(\sqrt{3}\Phi_a)]^{-2}$$

(11-82)

which is shown plotted versus Φ_a in Fig. 11.36. The diffraction efficiency is seen to asymptotically approach its maximum value of 0.072, or 7.2%.

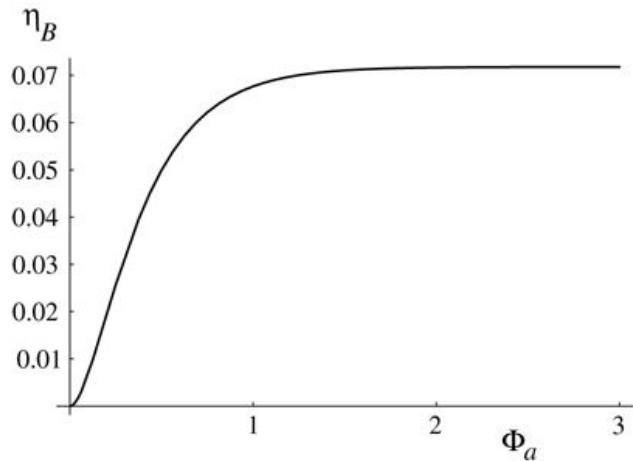


Figure 11.36

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.36 Bragg matched diffraction efficiency of a thick amplitude reflection grating.

The graph has its horizontal axis labeled Φ_a with markings from 1 to 3 with equal increments of 1 and vertical axis labeled η_B with markings from 0.01 to 0.07 with equal increments of 0.01. The curve starts at 0 and increases gradually to the point 0.07 on the vertical axis corresponding to 1 on the horizontal axis and then stabilizes and moves parallel to the horizontal axis. The values of the graph mentioned above are approximate.

Under Bragg mismatched conditions, again with the largest possible modulation of absorption, the following expression holds for χ_a :

$$\chi_a = 2\Phi_a + j\chi,$$

$$\chi_a = 2\Phi_a + j\chi,$$

(11-83)

where χ is as given in (11-68) earlier. Thus the expression for the Bragg mismatched diffraction efficiency can be written

$$\eta = 2 + j\chi\Phi_a + 2 + j\chi\Phi_a^2 - 1 \coth(\Phi_a^2 + j\chi\Phi_a^2 - 1)^{-2}.$$

$$\eta = \left| 2 + j\frac{\chi}{\Phi_a} + \sqrt{\left(2 + j\frac{\chi}{\Phi_a}\right)^2 - 1} \coth\left(\Phi_a \sqrt{\left(2 + j\frac{\chi}{\Phi_a}\right)^2 - 1}\right) \right|^2.$$

(11-84)

[Figure 11.37](#) illustrates the dependence of diffraction efficiency on Φ_a and χ , again with the display rotated to make its structure most clear. The broadening tolerance to Bragg mismatch as the parameter Φ_a increases is a result of the increasing absorption of the grating, and a resulting decrease of its effective thickness.

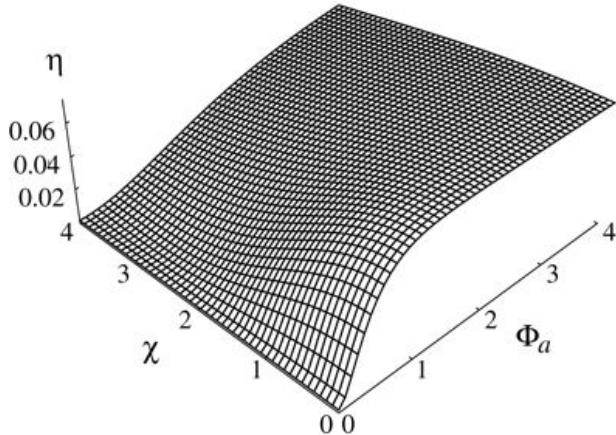


Figure 11.37

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.37 Diffraction efficiency of a thick amplitude reflection hologram when Bragg mismatch is present.

The illustration shows a three dimensional plot with two axes in the shape of the alphabet “V” with the left axis labeled chi and the right axis labeled psi a and a straight line begins at the left end of V which is the third axis. The third axis is labeled eta. A square shaped grid with two sides resting on the V-shaped structure is shown. The side of the square corresponding to chi rests on the chi axis without showing any deviations whereas the side that corresponds to psi a shows a gradual increase and reaches the point 0.07 on eta axis which corresponds to the point 1 on horizontal axis and then stabilizes and moves parallel to the psi a axis up to the point 4.

Summary of Maximum Possible Diffraction Efficiencies

In [Table 11.1](#) the various maximum diffraction efficiencies possible with thick gratings of various kinds are summarized. For comparison purposes, recall that for a thin sinusoidal amplitude grating the maximum possible diffraction efficiency is 6.25% and for a thin sinusoidal phase grating the maximum is 33.8%.

Table 11.1: Maximum possible diffraction efficiencies of volume sinusoidal gratings.

Phase transmission	Amplitude transmission	Phase reflection	Amplitude reflection
100%	3.7%	100%	7.2%

11.8 Recording Materials

Holograms have been recorded in a vast number of different materials during the history of holography. In this section we will offer a brief review of some of the more important recording materials. Unfortunately space limitations prevent a complete coverage here. For further information, the reader may wish to consult any of the textbooks on holography already cited. A particularly relevant reference is [323].

11.8.1 Silver Halide Emulsions

The most widely used recording materials for holograms are certainly those based on silver halide photographic technology. A detailed review of such materials, with particular reference to holography, can be found in [29]. It should be noted at the start that a major distinguishing characteristic of holographic materials is that they must be capable of recording extremely fine structures, as compared with the structures encountered in ordinary photography. The spatial frequency response of holographic recording materials often exceeds 3000 cycles/mm, whereas in conventional photography, a spatial frequency response of 300 cycles/mm is considered high. A corollary to this fact is that high resolution is always accompanied by low sensitivity. High resolution is achieved by constructing emulsions with very small grain sizes, but a certain number of photons must be absorbed by each grain to make it developable. It follows that the energy densities needed to expose high-resolution materials are much greater than those required for low-resolution materials.

In the past, the major manufacturers of silver halide materials suitable for holography were Kodak, Agfa-Gaevert, and Ilford. These manufacturers no longer supply silver halide materials suitable for holography, although materials similar to Ilford's are available from the company Harman Holo. Other manufacturers that supply silver halide materials for holography include ThorLabs, Integraph, Ultimate Holography, and Laser Reflections, just to name a few. These providers can supply materials with sensitivity in the red (HeNe laser wavelength) or in the green (argon laser wavelength). They have resolutions that exceed 3000 lines/mm and various sensitivities. Holographic materials are also available from Slavich in Russia, marketed by the company Geola in Lithuania.

11.8.2 Photopolymer Films

Photopolymer films provide a recording medium with two major virtues: (1) the holograms obtained are primarily phase holograms, and (2) the films can be coated with considerable thickness (as thick as 8 mm). The thick phase holograms that result can have excellent efficiency.

The modulation mechanism for these holograms is a combination of thickness variation and internal refractive index change. The recording material is a photopolymerizable monomer, i.e. a monomer that experiences polymerization or cross-linking under exposure to light. After the initial polymerization of the portions of the monomer exposed to light, diffusion of the remaining monomer takes place away from areas of high concentration (low exposure). A final uniform exposure polymerizes the remaining monomer, but due to the previous diffusion, the distribution of polymer is now nonuniform, leading to the phase modulation properties of the hologram. Changes of refractive index of 0.2% to 0.5% are possible.

Work on recording volume holograms in photopolymers began in the late 1960s at the Hughes Research Laboratories [72]. Further important work included that of [Booth](#), [\[32\]](#), [\[33\]](#), [Colburn and Haines](#), [\[74\]](#), and many others. See [\[324\]](#), pp. 293–298, for a more detailed consideration of the history of this material in volume holography. An excellent review article is also found in [\[155\]](#).

Photopolymer materials are either self-developing or dry-developing, for example by exposure to UV light. Resolutions are excellent but sensitivities are low, typically a few mJ/ cm² cm² being required for exposure. Manufacturers of photopolymer materials include Dupont, Baer and Polaroid.

11.8.3 Dichromated Gelatin

Dichromated gelatin films are widely used to record extremely efficient volume phase holograms, particularly of the reflection type. Diffraction efficiencies in excess of 90% are readily achieved repeatably.

A gelatin film containing a small amount of a dichromate, such as (NH₄)₂Cr₂O₇ (NH₄)₂Cr₂O₇, is found to harden under exposure to light. The process is a form of molecular cross-linking, similar to that observed in polymer films. Since dichromated gelatin plates are not available commercially, users must make their own photosensitive plates from gelatin films, typically coated on a glass plate. The methods used for preparing such plates and developing them are quite complex and must be performed with great care. A description of these methods can be found, for example, in [\[160\]](#), [\[307\]](#), and [\[324\]](#).

Particularly important publications on this material from an historical point of view include [\[313\]](#), [\[228\]](#), [\[60\]](#), [\[252\]](#), and others. Again a more detailed discussion of the history can be found in [\[324\]](#), pp. 278–286.

A number of theories have been proposed to explain the physical mechanism that takes place in the dichromated gelatin plates. Currently the best-accepted theory [\[56\]](#) is that a large number of very tiny vacuoles, with sub-wavelength dimensions, form in unhardened areas of the film. The density of vacuoles changes the local refractive index, allowing smooth and continuous variations of phase shift.

Recording using dichromated gelatin films is carried out typically at 488 nm or 514.5 nm wavelengths in the blue and green, respectively. Emulsion thickness may be of the order of 15 μ^μ m, and exposures required are of the order of 50 to 100 mJ/ cm² cm², a very high exposure indeed.

11.8.4 Photorefractive Materials

A number of crystals, including lithium niobate (LiNbO₃), barium titanate (BaTiO₃ BaTiO₃), bismuth silicon oxide (BSO), bismuth germanium oxide (BGO), potassium tantalum niobate (KTN), and strontium barium nitrate (SBN), exhibit a combination of sensitivity to light and an electro-optic effect. This combined effect has come to be known as the *photorefractive effect*, and the materials that exhibit it are known as *photorefractives* or photorefractive materials. For an excellent background on these materials and their applications in optics, see [\[153\]](#) and [\[154\]](#).

Early work on photorefractive holograms took place at the Bell Laboratories [61] and the RCA Laboratories [6], [5]. Considerable advances in theoretical understanding were developed in this early work. The reader should consult the more general references cited above for a more complete historical picture.

The mechanisms by which incident optical interference patterns are stored as changes of the local refractive index of these materials are extremely complex and in some cases not completely understood. Index change is known to occur as a result of charge transport and the electro-optic effect. The charge transport results from photoexcitation of trapped carriers, transport of these carriers, and retrapping at new locations. In some materials (e.g. SBN) the transport mechanism is diffusion, while in others (e.g. LiNbO₃), it may be the photovoltaic effect under some circumstances and diffusion under others. After charge transport and re-trapping, internal electric fields will result from the charge redistribution. These internal electric fields cause, through the electro-optic effect, local changes of the refractive index experienced by polarized light.

[Figure 11.38](#) illustrates an incident sinusoidal intensity pattern and the resulting distributions of charge, electric field, and refractive index. The charge carriers, which in this case carry positive charge, migrate to the nulls of the intensity pattern, establishing a charge distribution that is 180° 180° out of phase with the incident intensity distribution. An electric field is proportional to the spatial derivative of charge, and hence the electric field distribution is 90° 90° out of phase with the charge distribution and (in the opposite sense) with the intensity distribution. Assuming the linear electro-optic effect, the refractive index change is proportional to the electric field, and a volume index grating, spatially phase-shifted by 90° 90° from the exposure pattern, results.

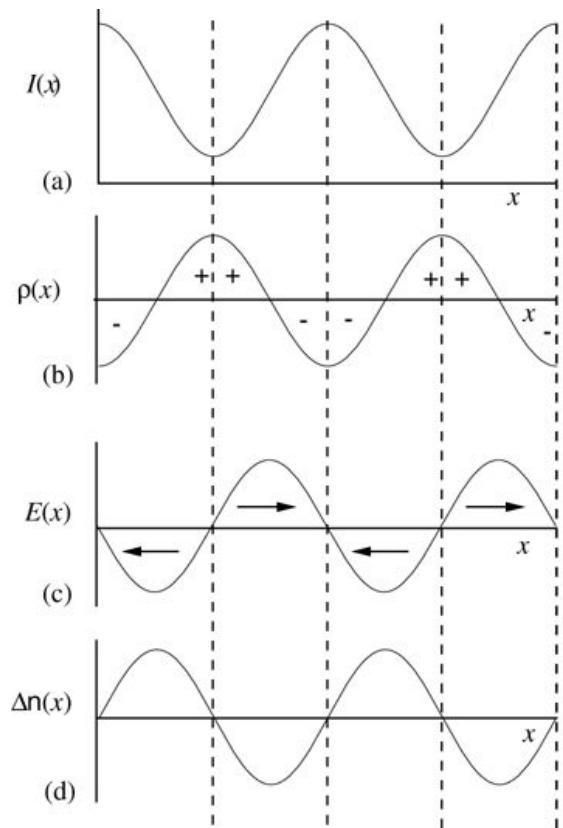


Figure 11.38

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.38 Relations between (a) an incident sinusoidal intensity pattern and the resulting distributions of (b) charge, (c) electric field, and (d) refractive index change in a photorefractive material.

Graph a shows the horizontal axis labeled x and vertical axis labeled $I(x)$. The curve is a sine wave that starts at a point slightly lower than the upper end of vertical axis and decreases to reach negative peak value and increases to reach positive peak value and follows the same motion showing a wave-like pattern. Four dashed vertical lines pass through the midpoint of each crest and trough which extend downward up to graph d.

Graph b shows the horizontal axis labeled x and vertical axis labeled $\rho(x)$. The curve is a sine wave which starts at a point above the lower end and then increases to reach the peak value and decreases again to reach negative peak value and repeats the same motion resulting in a wave-like pattern. The dashed line passes through the midpoint of each crest and trough with one minus sign in the portion below the horizontal axis and two plus signs, one on either side of the first dashed line above the horizontal axis, two minus signs, one on either side of the second dashed line below the horizontal axis, two plus signs, one on either side of the third dashed line above the horizontal axis and one minus sign below the horizontal axis before the last dashed line.

Graph c

Graph c shows the horizontal axis labeled x and vertical axis labeled $E(x)$. The curve is a sine wave which starts at 0 and gradually decreases and moves downward below the horizontal axis reaching a negative peak value and again increases gradually above the horizontal value reaching the positive peak value and follows the same wave-like pattern such that the dashed lines pass through the meeting points of the curve on the horizontal axis. An arrow facing the right is shown

in the crest portions of the wave that are above the horizontal axis and an arrow facing left is shown in trough portions of the wave that are below the horizontal axis.

Graph D

Graph d shows the horizontal axis labeled x and vertical axis delta n(x). The curve is a sine wave which starts at 0 and gradually increases and reaches the positive peak value and decreases gradually and moves downward below the horizontal axis reaching a negative peak value and follows the same wave-like pattern such that the dashed lines pass through the meeting points of the curve on the horizontal axis.

The 90° phase shift between the exposure pattern and the pattern of refractive index change plays an important role in the transfer of energy between the two interfering beams during the exposure process. The two interfering beams create, in any incremental distance Δz normal to the grating fringes, a weak phase grating with amplitude transmittance of the form

$$tA(x,y) = \exp[j2\pi n(x,y)\Delta z\lambda_0]$$

$$t_A(x, y) = \exp \left[j2\pi \frac{n(x, y) \Delta z}{\lambda_o} \right].$$

(11-85)

Since the grating is weak, the argument of the exponential is small, and

$$tA(x,y) = \exp[j2\pi\Delta n\Delta z\lambda_0\sin 2\pi x/\Lambda] \approx 1 + j2\pi\Delta n\Delta z\lambda_0\sin 2\pi x/\Lambda,$$

$$t_A(x, y) = \exp \left[j2\pi \frac{\Delta n \Delta z}{\lambda_o} \sin 2\pi x \right] \approx 1 + j2\pi \frac{\Delta n \Delta z}{\lambda_o} \sin 2\pi x \Big| \Lambda, \\ (11-86)$$

where Δn is the peak refractive index change in the grating, and Λ is the grating period.

Note in particular the 90° phase difference between the zero order (represented by unity) and the combination of the two first orders (represented by the sinusoid) in the last expression. For one of the first-order diffracted components, the spatial shift of the index grating with respect to the incident intensity pattern compensates for the similar phase shift in the above equation, with the result that strong coupling and energy transfer can occur between the two incident beams. In this fashion a strong incident beam can couple to a weak incident beam such that the component diffracted in the direction of the weak beam is amplified by energy transfer from the strong beam.

It is often found desirable to apply an external voltage across the photorefractive crystal in a direction orthogonal to the grating planes to induce a drift component of charge transfer. Such a voltage is found to strengthen the diffraction efficiency of the crystal for low spatial frequency components of the interference pattern, whereas without the applied field the diffraction efficiency may be poorer for low frequencies than for high frequencies.

Many photorefractive crystals are extremely slow when compared with photographic emulsions, at least for exposures with typical CW lasers. In fact their response time depends on the rate at which energy is delivered to them, and therefore a recording can be made in a very short time (e.g. a few nsec) with a powerful pulsed laser.

The chief difficulty found with the use of photorefractive materials as a medium for holography is the fact that the reconstruction beam will partially or totally erase the stored

hologram as the image is read out. While in some simple cases it is possible to read out the image with a different wavelength than was used for recording, in particular a wavelength to which the crystal is not sensitive, this is not possible in general due to the impossibility of properly Bragg matching the entire set of gratings that were recorded for a complex object when there is a wavelength change. Various methods for “fixing” the recorded holograms have been investigated.

Photorefractive crystals have found applications in interferometry, adaptive optics, holographic memories, and optical signal and image processing. They have formed the basis for certain types of spatial light modulators. For a review of many of these applications, see [\[154\]](#).

11.9 Computer-Generated Holograms

The vast majority of holograms are made using interference of coherent light, as described in previous sections. However, a significant amount of study has been given to methods for creating holograms by means of calculations on a digital computer, which are then transferred to a transparency by means of a plotting or printing device. The advantage gained by such a process is that one can create images of objects that in fact never existed in the real physical world. We thus become limited in the creation of images (two-dimensional or three-dimensional) only by our ability to describe that image mathematically, our ability to compute the hologram numerically in a reasonable amount of time, and our ability to transfer the results of that computation to a suitable transparent medium, such as photographic film or plate.

The process of creating a computer-generated hologram can be broken down into three separate parts. First is the computational part, which involves calculation of the fields that the object would produce in the hologram plane if it existed. It is these fields, or an approximation to them, that we wish the hologram to generate. This portion of the problem has itself two distinct parts: (1) a decision as to how many sampling points we should use for the object and the hologram (we can calculate only a discrete set of samples of the desired field starting from a discrete representation of the object); and (2) the carrying out of the correct discrete Fresnel or Fourier transform on the object fields, which is usually accomplished with a fast Fourier transform algorithm. These aspects are covered in [Chapter 5](#).

The second part of the process is the choice of a suitable representation of the complex fields in the hologram plane. The result of the calculation mentioned above is usually a discrete set of samples of a complex field, each sample point having both a magnitude and a phase. In general we cannot create structures that directly control both the amplitude and the phase of the amplitude transmittance in arbitrary ways, so some form of encoding of these quantities into a form suitable for representation on a transparency must be chosen.

The third part of the problem is the transfer of the encoded representation of the fields to a transparency. This plotting or printing operation is constrained by the properties of available computer output devices, whether they be pen plotters, laser printers, or electron-beam lithography machines. In fact, the choice of an encoding step is often influenced by the properties of the plotting device that will be used, so the second and third parts of the problem are not entirely independent. Most plotting devices are capable of writing small rectangles at various locations on the output plane. In some cases those rectangles can be written with gray scale, while in others they are restricted to binary values, i.e. transparent or opaque.

Many different methods for creating computer-generated holograms have been discovered, but we are limited by space constraints here to discuss only a few of the most important kinds. For more complete discussions, see [\[219\]](#) and [\[379\]](#). It should be noted that computer-generated holograms are almost invariably *thin* holograms, due to the constraints imposed by the various methods that are used to write such holograms onto transparencies.

11.9.1 The Sampling and Computation Problems

The process of holography, whether analog or digital, invariably involves the creation of a complex field in the hologram plane, a field that we wish to regenerate during the wavefront

reconstruction process. For computer-generated holograms, we must calculate that field using a digital computer; of necessity the field must be sampled, and complex values computed at each sample point. How many samples of the field must we compute?

This question has been answered in [Chapter 5](#). The answer depends on the Fresnel number of the geometry, which in turn determines how much zero padding must be used in the object space (the factor $Q^{\frac{1}{2}}$ in [chapter 5](#)). Any of the several approaches discussed there, especially the Fresnel transform approach or the Fresnel transfer function approach, may be used to perform the computation. In either case, heavy use is made of the discrete Fourier transform for the computation. The reader is referred to that chapter for details. We now assume that the computation has been performed and the result is adequately sampled.

11.9.2 The Representational Problem

Having obtained the complex field in the hologram plane, the remaining important step is to adopt a representation of that field that can be encoded in a hologram. Just as with holograms recorded by analog optical means, it is not practical in general to attempt to control both the amplitude and the phase of the amplitude transmittance of a transparency (an exception is the so-called ROACH, to be mentioned below). Therefore some method for encoding complex amplitude into either amplitude or phase is required. We discuss various such methods in what follows. The reader should keep in mind that once a suitable hologram has been plotted or printed by any of the means discussed, it is then necessary to photo-reduce the plot and produce a transparency that can be illuminated with coherent light.

Detour-Phase Holograms

The oldest and perhaps the best known method for creating holograms from computed complex fields is the so-called “detour-phase” method, invented by [Brown and Lohmann \[44\], \[45\]](#) and [Lohmann and Paris \[233\]](#). This method accepts the constraints imposed by most plotting devices, namely that it is easiest to plot binary patterns (ink or no ink) and that convenient basic building blocks are black rectangles that can be centered at any of a quantized set of locations and can be controlled in size at least in certain quantized increments.

Suppose that the final hologram transparency will be illuminated by an off-axis plane wave, and the image will be obtained with a positive lens of focal length f by looking on the optical axis in the focal plane behind the lens. Let the illuminating wave be inclined with respect to the x axis for simplicity, so that its complex field distribution incident on the hologram plane is

$$U_p(x, y) = \exp[-j2\pi\alpha x], \quad (11-87)$$

where α is equal to $\sin\theta/\lambda$, θ being the angle between the k vector of the incident beam and the normal to the hologram. Then for each value of x on the hologram plane, the optical phase of the illuminating beam has a different value.

Let the hologram plane be divided into $N_X \times N_Y$ separate cells, with the width of a cell in the x direction being equal to one full period of the incident phase function, i.e. the width

is $\alpha^{-1} \alpha^{-1}$. The width in the y direction need not necessarily be the same but for simplicity might be chosen so. Each cell defined in this way will encode one of the Fourier coefficients that was calculated with the fast Fourier transform.

Suppose that one particular Fourier coefficient is given by

$$a_{pq} = U_h(p\Delta x, q\Delta y) = |a_{pq}| \exp(j\phi_{pq}).$$

$$a_{pq} = U_h(p\Delta x, q\Delta y) = |a_{pq}| \exp(j\phi_{pq}).$$

(11-88)

Then within that cell we will plot a black rectangle with an area proportional to $|a_{pq}|$ and with a position in the x direction such that at the center of the rectangle, the incident phase from the reconstruction beam is precisely ϕ_{pq} . Remembering that a black rectangle will be changed to a transparent rectangle after the plot is photographed, we have created a transmitted wave component from this cell that has the amplitude of the desired Fourier component and a phase equal to that of the desired Fourier component. Phase shift has been achieved by moving the center of the plotted rectangle, a method known as *detour phase*, and illustrated in Fig. 11.39. Note that our goal is to synthesize an image field of the form

$$U_f(u, v) = \sum_{p=0}^{N_X-1} \sum_{q=0}^{N_Y-1} |a_{pq}| e^{j\phi_{pq}} \exp\left[j\frac{2\pi}{\lambda f}(u p\Delta x + v q\Delta y)\right],$$

$$U_f(u, v) = \sum_{p=0}^{N_X-1} \sum_{q=0}^{N_Y-1} \left| a_{pq} \right| e^{j\phi_{pq}} \exp\left[j\frac{2\pi}{\lambda f}(u p\Delta x + v q\Delta y)\right].$$

(11-89)

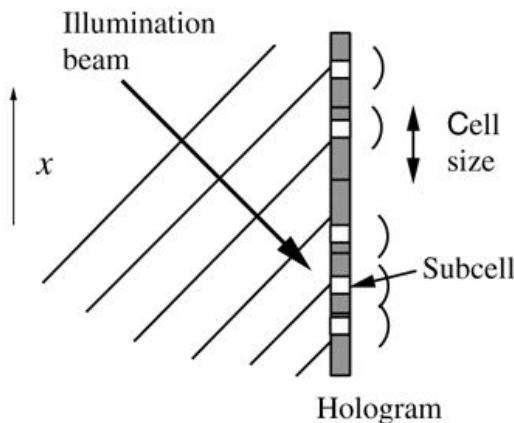


Figure 11.39

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.39 The detour-phase concept. The subcells are moved within a cell to control the phase of the transmitted light. Zero-phase lines of the reconstruction wavefront are shown.

The illustration shows a vertical bar labeled hologram with five partitions called cells. Each partition has a small horizontal patch at the centre which is labeled subcell. The length of each cell

is labeled cell size. Five slanting lines begin from the left bottom, with four pointing toward the hologram, one below the other and the line at the top ends at a point slightly before the hologram. A vertical arrow labeled x facing upward is on the extreme left. An arrow from top points toward the fourth subcell and is labeled Illumination beam.

which expresses the image field as the sum of its Fourier components, all with proper amplitudes and phases.

To understand the approximations inherent in the detour-phase approach, we undertake a short analysis. Consider first the diffraction pattern created in the rear focal plane of the transforming lens when a single transparent rectangle of widths (w_X, w_Y) exists in the hologram plane, that rectangle being describable by the function

$$tA(x,y) = \text{rect}(x - x_0) \text{rect}(y - y_0),$$

$$t_A(x, y) = \text{rect}\left(\frac{x - x_0}{w_X}\right) \text{rect}\left(\frac{y - y_0}{w_Y}\right),$$

(11-90)

where (x_0, y_0) is the center of the rectangle. When this rectangle is illuminated by the reconstruction wave of (11-87), the transmitted field is

$$U_t(x,y) = e^{-j2\pi\alpha x} \text{rect}(x - x_0) \text{rect}(y - y_0),$$

$$U_t(x, y) = e^{-j2\pi\alpha x} \text{rect}\left(\frac{x - x_0}{w_X}\right) \text{rect}\left(\frac{y - y_0}{w_Y}\right),$$

and the optical Fourier transform of this field is given by

$$U_f(u,v) = w_X w_Y \lambda f \text{sinc}(w_X(u + \lambda f \alpha)) \text{sinc}(w_Y(v + \lambda f \alpha)) \exp(j2\pi\lambda f[(u + \lambda f \alpha)x_0 + vy_0]),$$

$$U_f(u, v) = \frac{w_X w_Y}{\lambda f} \text{sinc}\left[\frac{w_X(u + \lambda f \alpha)}{\lambda f}\right] \text{sinc}\left[\frac{w_Y v}{\lambda f}\right] \exp\left\{j\frac{2\pi}{\lambda f}[(u + \lambda f \alpha)x_0 + vy_0]\right\},$$

(11-91)

where we have made use of the similarity and shift theorems of Fourier analysis.

If the width w_X of this rectangle is limited in the x direction so that it is a small fraction of the period of the reconstruction beam,

$$w_X \ll \alpha^{-1},$$

$$w_X \ll \alpha^{-1},$$

then the shift of the first sinc function can be neglected. In addition, if the region of interest in the image plane (size $L_U \times L_V$) is much smaller than the width of the sinc functions, then those functions can be replaced by unity within that region. The resulting approximation to the contribution of this single rectangle can then be written

$$U_f(u,v) = w_X w_Y \lambda f e^{-j2\pi\alpha x_0} \exp(j2\pi\lambda f(ux_0 + vy_0)).$$

$$U_f(u, v) = \frac{w_X w_Y}{\lambda f} e^{-j2\pi\alpha x_0} \exp \left[j \frac{2\pi}{\lambda f} (ux_0 + vy_0) \right].$$

(11-92)

Now consider the result of introducing many such rectangles, one for each cell defined in the hologram plane. The cells are indexed by (p, q) , since each cell represents one Fourier coefficient of the image. For the moment we assume that all rectangles are located precisely in the center of their respective cells, but in general each may have a different set of widths (w_X, w_Y) (w_X, w_Y) , subject to the constraint on $w_X w_Y$ introduced above. Thus the center of the (p, q) th cell is located at

$$\begin{aligned} (x_0)_{pq} &= p\Delta x \\ (y_0)_{pq} &= q\Delta y. \end{aligned}$$

(11-93)

The total reconstructed field in the image plane becomes

$$U_f(u, v) = \sum_{p=0}^{N_X-1} \sum_{q=0}^{N_Y-1} (w_X)_{pq} (w_Y)_{pq} e^{-j2\pi p} \exp(j2\pi\lambda f(u p\Delta x + v q\Delta y)),$$

$$U_f(u, v) = \sum_{p=0}^{N_X-1} \sum_{q=0}^{N_Y-1} (w_X)_{pq} (w_Y)_{pq} e^{-j2\pi p} \exp \left[j \frac{2\pi}{\lambda f} (u p\Delta x + v q\Delta y) \right],$$

(11-94)

where the period α^{-1} of the reconstruction wave must equal Δx , the width of one cell. Thus when the subcells are all centered in their respective cells, the phase of the first exponential term is seen to be an integer multiple of 2π , and that term can be replaced by unity. The terms represented by the second exponential are the Fourier basis functions that we are attempting to add to synthesize the final image. The amplitude of the (p, q) th Fourier component is $w_X w_Y / \lambda f$ $w_X w_Y / \lambda f$ and the phases of all components are identical. While we can properly control the amplitudes of the Fourier components by controlling the y widths $(w_Y)_{pq}$ (which are not constrained by the limitation imposed on $w_X w_Y$ in our earlier approximation), we have not yet controlled the phases of these components properly.

Phase control is introduced by moving the centers of the subcells in the x direction within each cell. Suppose that the center of the (p, q) th cell is now located at

$$(x_0)_{pq} = p\Delta x + (\delta x)_{pq}$$

$$\begin{aligned} (x_0)_{pq} &= p\Delta x + (\delta x)_{pq} \\ (y_0)_{pq} &= q\Delta y. \end{aligned}$$

(11-95)

With this change, the expression of (11-94) becomes

$$U_f(u, v) = \sum_{p=0}^{N_X-1} \sum_{q=0}^{N_Y-1} (w_X)_{pq} (w_Y)_{pq} e^{-j2\pi \frac{(\delta x)_{pq}}{\Delta x}} \exp \left[j \frac{2\pi}{\lambda f} (u p \Delta x + v q \Delta y) \right]$$

(11-96)

where an exponential term with a phase that is an integer multiple of 2π has been replaced by unity. One further approximation is needed. We assume that the width of the image region of interest, which extends in the u direction over $(L_U/2, -L_U/2)$ is sufficiently small that

$$L_U(\delta x)_{pq} 2\lambda f \ll 1,$$

$$\frac{L_U(\delta x)_{pq}}{2\lambda f} \ll 1,$$

in which case the exponential term $\exp[-j2\pi\lambda f u(\delta x)_{pq}] \approx 1^{\exp[-j\frac{2\pi}{\lambda f} u(\delta x)_{pq}]} \approx 1$, leaving the following expression for the image field:

$$U_f(u, v) = \sum_{p=0}^{N_X-1} \sum_{q=0}^{N_Y-1} (w_X)_{pq} (w_Y)_{pq} e^{-j2\pi \frac{(\delta x)_{pq}}{\Delta x}} \exp \left[j \frac{2\pi}{\lambda f} (u p \Delta x + v q \Delta y) \right].$$

(11-97)

This field does have the phases of the Fourier components properly controlled, provided

$$\exp[-j2\pi(\delta x)_{pq}\Delta x] = \exp(j\phi_{pq}).$$

$$\exp \left(-j2\pi \frac{(\delta x)_{pq}}{\Delta x} \right) = \exp(j\phi_{pq}).$$

Given the phase ϕ_{pq} of the (p, q) th Fourier component, the subcell in the (p, q) th cell should be centered at $(\delta x)_{pq}$ satisfying

$$-(\delta x)_{pq}\Delta x = \phi_{pq}2\pi.$$

$$-\frac{(\delta x)_{pq}}{\Delta x} = \frac{\phi_{pq}}{2\pi}.$$

(11-98)

In addition we choose the width $(w_Y)_{pq}$ of the (p,q) th subcell to be proportional to the desired magnitude of the (p,q) th Fourier component,

$$(w_Y)_{pq} \propto |a_{pq}|.$$

$$(w_Y)_{pq} \propto |a_{pq}|.$$

(11-99)

$(w_X)_{pq}$ is held constant to satisfy the previous approximation regarding the overlap of the sinc functions. Thus we have created a field in the image plane that is, to within a proportionality constant, equal to the desired field represented by [Eq. \(11-89\)](#). [Figure 11.40](#) illustrates a single cell in the detour-phase hologram.

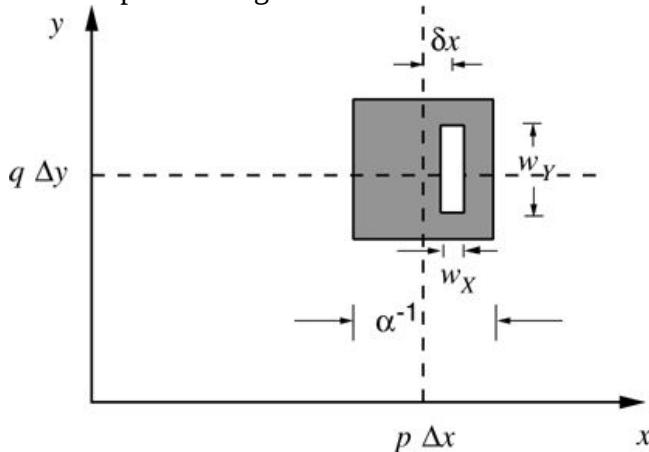


Figure 11.40

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.40 A single cell in a detour-phase hologram.

The graph shows the horizontal axis x and vertical axis labeled y . A horizontal dashed line is shown that starts at the point $q \Delta y$ on the vertical axis which is slightly above the midpoint. A vertical dashed line starts at the point $p \Delta x$ on the horizontal axis which is slightly to the right of the midpoint. A shaded square is shown with the meeting point of the dashed horizontal and vertical lines as its centre. Inside the square is a vertical rectangular bar which is to the right of square's centre point. The length of the patch is labeled w_Y and the width of the patch is labeled w_X . The distance between the vertical dashed line and the center of the vertical bar is labeled δx . The bottom side of the square is labeled α^{-1} .

Once the desired reconstructed field is generated by the hologram, an image will appear in the rear focal plane of a positive lens placed behind the hologram. In fact, as in the case of optically recorded holograms, this type of computer-generated hologram utilizes a carrier frequency α and generates twin images. The second image can be made to appear on the optical axis of the transforming lens if the incident illumination wave is taken to be the conjugate of the previous reconstruction wave, i.e. if its angle with respect to the normal to the hologram is the negative of that in the previous case. Alternatively, the incident wave can be normal to the

hologram, in which case both twin images appear with opposite displacements off axis in the rear focal plane.

[Figure 11.41](#) shows (a) a binary detour-phase hologram and (b) an image reconstructed from such a hologram.

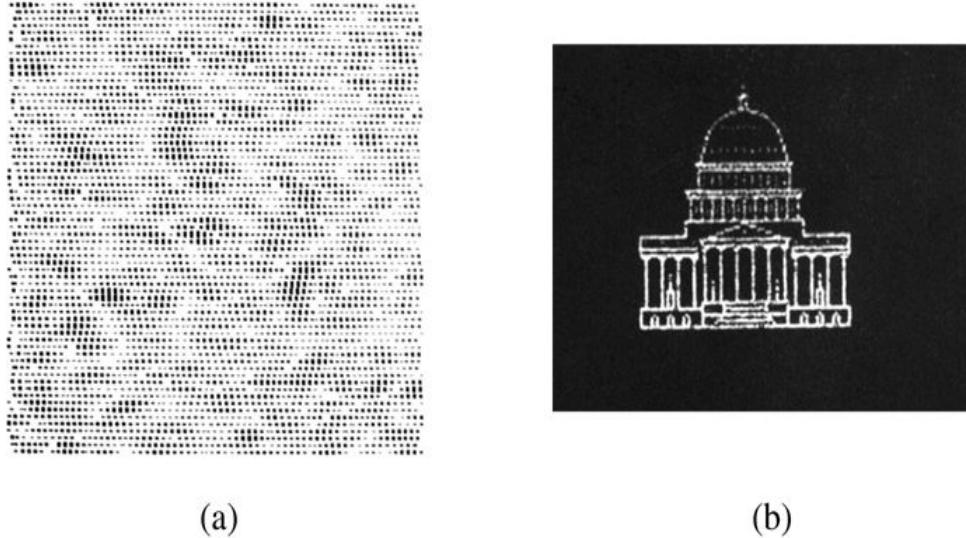


Figure 11.41
Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 11.41 (a) Binary detour-phase hologram; (b) image reconstructed from such a hologram. Courtesy of International Business Machines Corporation, © (1969) International Business Machines Corporation.

Illustration a shows a square patch inside which sequence of dots are arranged horizontally wherein some dots are darker than the others. Illustration b shows a square patch of darkness and at the centre a bright outline of a building is shown.

Note that in practice it is necessary to quantize both the amplitude and the phase of a binary detour-phase hologram, a process that leads to noise in the reconstructed image. The effects of phase quantization are particularly important and interesting [142], [85], [86].

Alternative methods of representation using the detour-phase concept exist. For example, [Lee \[217\]](#) utilized four fixed subcells per cell, each with a gray-level transmittance, the first representing the real and positive part of the Fourier coefficient (angle 0° 0°), the second the imaginary and positive part (angle 90° 90°), the third the real and negative part (angle 180° 180°), and the fourth the imaginary and negative part (angle 270° 270°). Since the real and imaginary parts are either positive or negative but not both, two of the subcells in every cell are normally opaque. [Burckhardt \[48\]](#) recognized that any point in the complex plane can be reached with only three gray-level phasors, one at 0° 0° , one at 120° 120° , and the third at 240° 240° .

The Kinoform and the ROACH

An entirely different method for computer-generated hologram representation is known as the [*kinoform*](#) [226]. In this case, an assumption is made that the *phases* of the Fourier coefficients carry the majority of information about an object, and the amplitude information can be entirely eliminated. While this assumption might at first glance appear surprising, it turns out to be quite

accurate if the object is a diffuse one, i.e. if the object points all are assigned random and independent phases.

Considering a Fourier geometry again, the hologram is divided up into $N_X \times N_Y$ cells, each representing one Fourier coefficient of the object. The amplitudes $|a_{pq}|$ of all Fourier coefficients are assigned value unity, and it is only the phases ϕ_{pq} that we attempt to encode in the hologram. The encoding is done by linearly mapping the phase range $(0, 2\pi)$ into a continuum of gray levels displayed by an output device such as a photographic plotter. The gray-level transparency obtained from this process is subjected to photographic bleaching. Thus each gray level is mapped into a corresponding phase shift introduced by the transparency, and if the bleaching process is well enough controlled to assure that the complete phase range $(0, 2\pi)$ is exactly and properly realized by the transparency, an excellent image can be obtained in the Fourier plane of the kinoform. In this case there is only a single image and it appears on the optical axis. The diffraction efficiency of the kinoform is very high because it is a pure phase transparency. Errors in “phase matching” the $(0, 2\pi)$ interval result in a single bright spot on the optical axis, generally in the midst of the desired image.

[Figure 11.42](#) shows a photograph of the gray-level recording that leads to a kinoform after bleaching, and the image obtained from the same kinoform.

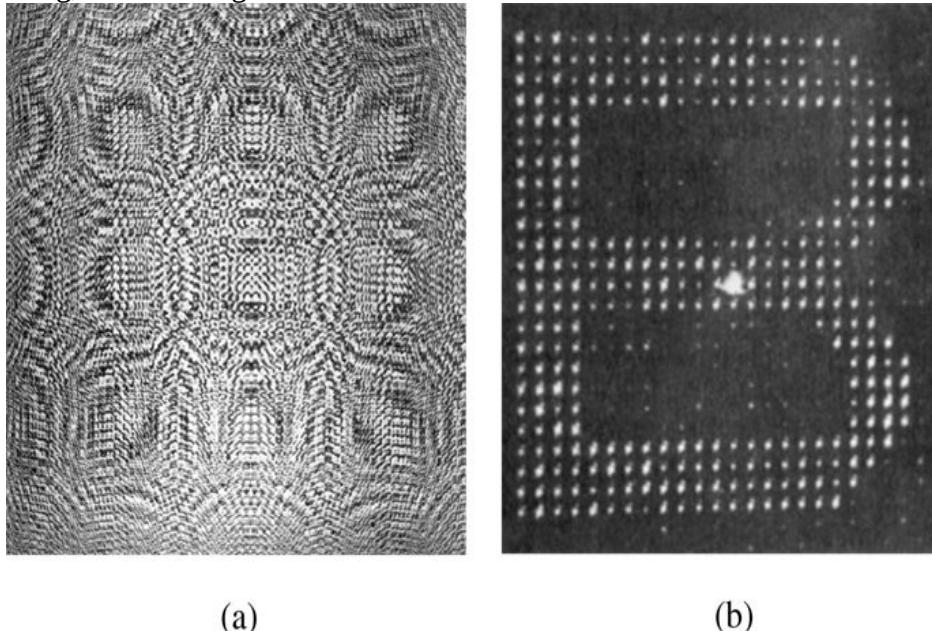


Figure 11.42
Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 11.42 (a) The gray level display that leads to a kinoform, and (b) the image obtained from that kinoform. Courtesy of International Business Machines Corporation, © (1969) International Business Machines Corporation.
Illustration a shows a square patch inside which dots are arranged in random pattern. Illustration B shows a dark shaded square patch with four sequences of bright dots arranged in the shape of alphabet “B.”

A related approach known as the “referenceless on-axis complex hologram” (ROACH) utilizes color film to control both the amplitude and the phase of the Fourier coefficients simultaneously [62]. Suppose we wish to create a computer-generated Fourier hologram which will reconstruct an image in red light. Let the magnitudes $|a_{pq}|$ of the Fourier coefficients first be displayed as gray levels on a black-and-white CRT display. This display is photographed through a red-transmitting filter onto a frame of reversal color film. The red-absorbing layer of the three-layer film records this exposure. Then the desired array of Fourier phases is encoded as an array of gray levels, as was done for the kinoform, displayed on the same CRT, and photographed through a blue-green transmitting filter, thus exposing the blue and green absorbing layers of the same frame of film used previously. After processing, the layer exposed to red light becomes absorbing in the red, but the layers exposed to blue-green light are transparent in the red. However, these layers do introduce a phase shift in the transmitted red light, due to thickness variations. Thus the color photographic transparency controls both the amplitude and the phase of the transmitted red light, and as such creates an on-axis image of the desired object. Again proper phase matching is critical, and errors in this regard result in a bright spot of light on axis.

Note that both the kinoform and the ROACH are more efficient than detour-phase holograms in their utilization of the space-bandwidth product of the plotter or display used, since only one resolution cell is required for each Fourier coefficient, whereas many cells are required for the binary hologram. However, both the kinoform and the ROACH require that the phase matching problem be solved, whereas no such problem exists for the detour-phase hologram.

Phase Contour Interferograms

When phase variations exceeding 2π radians are to be created by the hologram, as is often required for holographic optical elements used in optical testing, detour-phase holograms have the disadvantage that subapertures near the 2π phase boundaries may partially overlap. For such applications, other representational approaches have some advantages. We discuss here only one such approach, namely a method due to Lee [218].

We focus here on the problem of generating elements that control only the phase of the transmitted wavefront. Consider an optical element with ideal amplitude transmittance

$$tA(x,y)=121+\cos[2\pi\alpha x-\phi(x,y)].$$

$$t_A(x, y) = \frac{1}{2}[1 + \cos[2\pi\alpha x - \phi(x, y)]].$$

(11-100)

This is a carrier frequency hologram which generates two reconstructed waves of interest, one with a pure phase distribution $\phi(x,y)$ and the other a conjugate term with the negative of this phase distribution. Since this amplitude transmittance contains a continuum of gray levels, it would not be easy to display directly on a plotter and record with high fidelity. We prefer some form of binary pattern for this purpose. Let the plotter or printer create a contour plot of t_A , with one contour line per period, each located at a maximum of the distribution. Such contours are defined by the equation

$$2\pi\alpha x - \phi(x,y) = 2\pi n,$$

$$2\pi\alpha x - \phi(x, y) = 2\pi n,$$

(11-101)

where each integer n defines a different contour line. Such a plot, when photographically reduced, has been shown by Lee to generate both the desired phase distribution and its conjugate, each in a different first diffraction order [218].

[Figure 11.43](#) shows such a plot generated for the case of a quadratic-phase approximation to a lens, i.e. a phase distribution

$$\phi(x, y) = \pi\lambda f x^2 + y^2,$$

$$\phi(x, y) = \frac{\pi}{\lambda f} (x^2 + y^2),$$

where the constant λf has been chosen to be unity for convenience, and $\alpha=2.5$. The photoreduction of such a pattern will yield an optical element that provides the effect of a positive lens in one first diffraction order and the effect of a negative lens in the other first order.

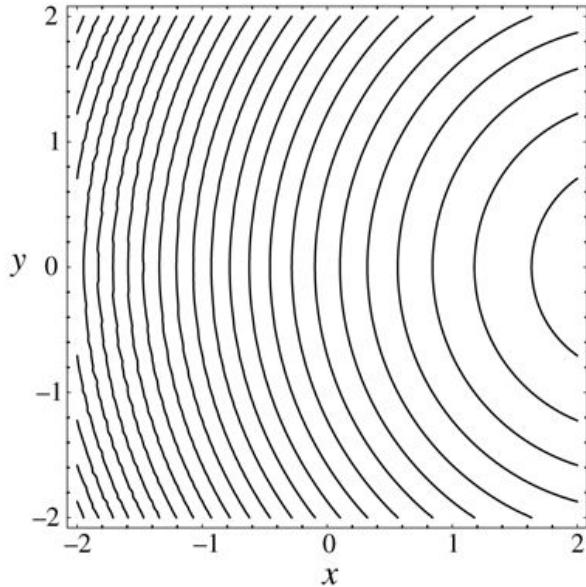


Figure 11.43

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.43 Plot of a phase contour interferogram for a quadratic-phase approximation to a spherical lens.

The illustration shows the horizontal axis labeled x with markings starting from -2 to 2 with intervals of 1 and vertical axis labeled y with marking from -2 to 2 with intervals of 1. Inside the square are concentric semicircles around the midpoint on the right border of the square.

Generalizations of this procedure to allow the incorporation of amplitude information in addition to phase information have been demonstrated by [Lee \[220\]](#). The reader is referred to the original reference for more details.

11.10 Degradations of Holographic Images

Holographic imaging, like other imaging approaches, suffers from certain degradations that limit the quality of the images that can be obtained. Some degradations, such as that caused by diffraction, are common to all systems. Others, while having a counterpart in conventional photography, manifest themselves in distinctly different ways in holography. In this section we review some of the common sources of image degradations and discuss the effects expected and found in holography.

Holography, like other imaging processes, can suffer from all of the classical aberrations encountered in optics. Consideration of such aberrations is beyond the scope of our treatment. The interested reader can find an excellent discussion in the work of [Meier \[248\]](#). We mention only that if a hologram is reconstructed with an exact duplicate of the original reference wave at the same wavelength used for recording, and no auxiliary optical elements exist between the object and the hologram and between the hologram and the image plane, the image obtained will be aberration-free (provided there has been no swelling or shrinking of the emulsion on which the hologram was recorded).

The holographic imaging process almost always uses coherent light (for exceptions, cf. [Section 11.12](#)). Under usual circumstances, such as operation in the linear region of the t_A versus E curve for thin holograms, the imaging process has been argued to be linear in complex amplitude, as long as attention is focused on one of the twin images. Under such circumstances it is possible to characterize the holographic process by an amplitude transfer function $H(f_X, f_Y)$. As with non-holographic systems, the amplitude transfer function is determined by the amplitude transmittance of exit pupil of the imaging system. Thus in the absence of effects associated with limited film MTF or film nonlinearities, we would characterize the holographic imaging process by an amplitude transfer function of the form

$$H(f_X, f_Y) = P(\lambda z_i f_X, \lambda z_i f_Y),$$

$$H(f_X, f_Y) = P(\lambda z_i f_X, \lambda z_i f_Y),$$

(11-102)

where P is the effective exit pupil function, usually determined by the finite size of the hologram. The amplitude transfer function fully accounts for the limitations to image quality posed by diffraction, and therefore we concentrate on other effects in what follows.

11.10.1 Effects of Film MTF

It has been seen previously that the holographic process in general places heavy requirements on the resolving power of recording materials, requirements that may not always be perfectly met in some applications. It is therefore of some interest to understand the effects of a limited spatial frequency response (MTF) of a recording material used in holography. It is important to remember that the MTF is a property of the recording medium, and limits its spatial frequency content. This

limitation of spatial frequency content affects the image of the object in a way that depends on the recording geometry.

We begin with an analysis of the Fourier transform and lensless Fourier transform geometries, and then generalize those results to other geometries. For more detailed consideration of the subject, the reader is referred to the classic work of [van Ligten \[352\], \[353\]](#).

Fourier Transform and Lensless Fourier Transform Holograms

The first type of hologram we consider here is one in which each object point is encoded (paraxially) as a fringe pattern of a unique and constant spatial frequency. Such is the case for the Fourier transform hologram and the lensless Fourier transform hologram discussed earlier. For both of these types of holograms the reference wave originates from a point that is in the same plane as the object, and each object point is encoded in a fringe with a constant spatial frequency that is proportional to the distance of that point from the reference point.

For both types of holograms, the intensity distribution falling upon the recording medium when the object is a point source at coordinates (x_o, y_o, z) and the reference is at coordinates (x_r, y_r, z) is

$$J(x, y) = A^2 + |a|^2 + 2A|a|\cos(2\pi(x_o - x_r)x\lambda_1 z + 2\pi(y_o - y_r)y\lambda_1 z + \phi).$$

$$\mathcal{I}(x, y) = A^2 + |a|^2 + 2A|a|\cos\left[2\pi\frac{(x_o - x_r)x}{\lambda_1 z} + 2\pi\frac{(y_o - y_r)y}{\lambda_1 z} + \phi\right].$$

(11-103)

Here z is the focal length of the lens in the case of the Fourier transform hologram, or the common perpendicular distance of the object and reference points from the recording plane in the case of the lensless Fourier transform hologram. ϕ is a phase angle that, for the lensless Fourier transform hologram, depends on the locations of the reference and the object points but not on the coordinates in the recording plane. For the true Fourier transform hologram, ϕ depends only on the relative phases of the object and reference points.

When the limited extent of the MTF plays a role, the *effective* intensity exposing the hologram is obtained by applying the MTF to the sinusoidal fringe in the interference pattern,

$$J_{\text{eff}}(x, y) = A^2 + |a|^2 + 2A|a|M\cos(2\pi(x_o - x_r)x\lambda_1 z + 2\pi(y_o - y_r)y\lambda_1 z + \phi).$$

$$\begin{aligned} \mathcal{I}_{\text{eff}}(x, y) &= A^2 + |a|^2 \\ &+ 2A|a|M\left(\frac{x_o - x_r}{\lambda_1 z}, \frac{y_o - y_r}{\lambda_1 z}\right)\cos\left[2\pi\frac{(x_o - x_r)x}{\lambda_1 z} + 2\pi\frac{(y_o - y_r)y}{\lambda_1 z} + \phi\right]. \end{aligned}$$

(11-104)

If the factor M in this expression is less than unity, then the amplitude of the fringe generated by this object point will be reduced, the diffraction efficiency will be lower, and the light amplitude incident on the twin images of this particular point will have been reduced by the MTF

of the recording medium. Since object points furthest from the reference point generate the highest spatial frequencies, these points will be attenuated the most.

In the cases of both the Fourier transform hologram and the lensless Fourier transform hologram, the twin images and the image of the reference point lie in a common plane. The effect of the MTF in the image space is representable by an attenuating mask, centered on the image of the reference point and extending over the twin images, with most attenuation for the image points that are furthest from the image of the reference point. If the wavelength used during reconstruction is λ_2 and the distance from the hologram to the image of the reference point is $z \sim \tilde{z}$, then the effective amplitude transmittance of the mask in the image space is given by

$$t_A(x_o, y_o) = M\left(\frac{\lambda_2 \tilde{z}}{\lambda_1 z}(x_o - x_r), \frac{\lambda_2 \tilde{z}}{\lambda_1 z}(y_o - y_r)\right).$$

(11-105)

The intensity transmittance of this mask is of course $|t_A|^2$. Thus for these two cases, the effect of the MTF of the recording medium is seen to restrict the *field of view* about the image of the reference point, but not to affect the resolution attained within that field of view.

Generalization of the Geometry

Suppose that the object is moved closer to the hologram or further from the hologram than the reference point. The reference point is now distance z_r from the hologram recording plane and the object is now z_o from the recording plane. A single object point produces a wave that interferes with the wave from the reference point to produce an intensity that has a varying frequency across the hologram. This variation gives the hologram focusing power, with the result that the twin images no longer reside in the same plane. Rather, during reconstruction (again with a positive lens and wavelength λ_2), one image is closer to the hologram than the image of the reference point while the other image is further from the hologram than the image of the reference point. The MTF mask described above still exists in the image space in the plane of the image of the reference point. Now the observer must look through that mask to view the image closest to the hologram, while the observer sees the image furthest from the hologram shadowed by that mask. The mask remains described by (11-105). In this case, the mask affects both the field of view and the resolution simultaneously, for the mask obscures some object points and at the same time reduces the range of angles that reach some the image points. The range of angles reaching an image point will, of course, affect the resolution achieved at that point.

If the reference wave is collimated, i.e. a plane wave, while the object exists at some finite distance from the hologram, and if no reconstruction lens is used, then the MTF mask exists at infinity and limits solely the range of angles reaching each image point. Thus in this case the effect of the finite MTF is to limit the resolution in the twin images, but not the field of view.

For further generalizations, the reader is referred to the work of Van Ligten referenced above.

11.10.2 Effects of Film Nonlinearities

Throughout our discussions of holography, we have repeatedly assumed that, at least for thin holograms, the recording medium is exposed in such a way as to assure operation within a linear region of the amplitude transmittance versus exposure curve. However, real recording media are never perfectly linear in this respect, the deviation from linearity depending to a large degree on the magnitude of the *variations* of exposure to which the medium is subjected and the exposure bias point selected. In this section we present a brief discussion of the effects of recording medium nonlinearities on the reconstructed image. It should be emphasized that, when the average exposure produced by the object is comparable with that produced by the reference, nonlinear effects can present a serious limitation to image quality. This case may be contrasted with that of a very weak object, for which film-grain noise or other scattered light is generally the limiting factor.

In what follows we omit any detailed analysis of nonlinear effects, preferring to give the reader a set of references that will provide an excellent overview of previous work. Almost all previous work has been devoted to *thin* holograms.

In discussing the effects of nonlinearities on the reconstructed images, it is important to distinguish between two different classes of object. One class consists of objects that contain a collection of isolated point sources; another class consists of diffuse objects, such as a transparency illuminated through a diffuser or a three-dimensional object with an optically rough surface. For both types of objects, the analyses use a model of the nonlinearities first introduced by [Kozma \[210\]](#). A simpler model was subsequently introduced by [Bryngdahl and Lohmann \[47\]](#).

For objects consisting of collections of point sources, the first analysis was that of [Friesem and Zelenka \[118\]](#), who demonstrated both analytically and experimentally several important effects. First, a phenomenon found with all types of objects, nonlinearities introduce higher-order images, i.e. images in the second, third, or higher diffraction orders. Such extraneous images are not of great concern since they generally do not overlap the first-order image. More important effects occur in the first-order image itself. If the object consists of two point sources, one of greater amplitude than the other, small-signal suppression effects are anticipated and observed. That is, the image of the weaker object point source is suppressed relative to that of the stronger point source. In addition, owing to intermodulation effects, false images may be generated within the first-order image by nonlinear interaction of the two point sources, yielding apparent images of point sources that are not actually present on the object itself.

The effects of film nonlinearities for diffuse objects have also been investigated [\[139\], \[212\]](#). In this case the exposure is most properly treated as a random process, employing techniques that are somewhat more complex than required for points-source objects. For further details, the reader may wish to consult the previous references. In this case it is found that the effects of nonlinearities are primarily to introduce a diffuse halo on and around the images of the object. If the diffuse object has fine structure, then the diffuse halo will have related structure. The effects can be quite severe, as illustrated in [Fig. 11.44](#).

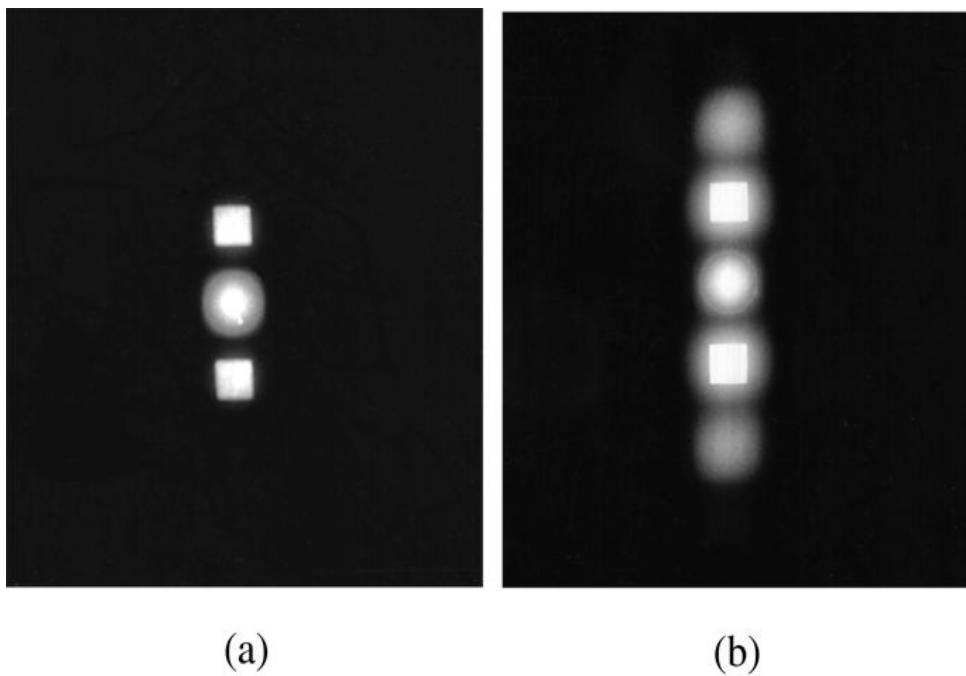


Figure 11.44
Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 11.44 Nonlinear effects in holography for a diffuse object. (a) Images obtained under nearly linear recording conditions, and (b) images obtained under highly nonlinear recording conditions. [From [\[139\]](#). Copyright 1967 by the Optical Society of America. Reprinted with permission.]

Illustration a shows a square patch of darkness with a circular bright dot at the center surrounded by a ring of reduced brightness and a square bright dot is shown both above and below the circular dot. Illustration b shows a circular dot at the centre and a square bright dot is shown both above and below the circular dot. All the three dots are surrounded by a ring of reduced brightness and at the upper and lower ends of the dots, is a circular patch with reduced brightness intensity.

11.10.3 Effects of Film-Grain Noise

When the object wave is very weak compared with the reference wave, the primary source of image degradations is often grain noise arising from the film or plate on which the hologram is recorded. The effects of film-grain noise have been analyzed by [Goodman](#) [\[133\]](#) and by [Kozma](#) [\[211\]](#).

The effects of finite grain size in the photographic emulsions used for holography manifest themselves in a spatial power spectrum of scattered light which overlaps the locations of the desired images. This noise spectrum reduces the contrast of the images obtained, and because it is coherent with respect to the images, it interferes with them to cause unwanted fluctuations of image brightness. Such effects are most noticeable with low resolution films, for which the grain size is relatively large and the scattered light spectrum correspondingly strong. They are also most important when the light arriving from the object during recording is very weak, in which case the statistics of the detection process become quite important.

A particularly important fact, first recognized by Gabor, is that holographic recording can in some circumstances provide greater signal detectability than can conventional photography of the

same coherent object. This enhancement comes about from the interference of a strong reference wave with the weak object wave, and the resulting enhancement of the strength of the fringes, a phenomenon analogous to “heterodyne conversion gain” observed in heterodyne detection. Experimental work has shown that such advantages exist in practice [141].

11.10.4 Speckle Noise

For diffuse objects illuminated with coherent light, granularity arising from speckle can be expected in the resulting images, regardless of the imaging method. Since holography is nearly always a process that depends on coherent light, speckle is of special concern in holographic imaging.

When viewing a virtual image, the pupil of the eye is the limiting aperture, and a speckle pattern appears on the retina of the observer. When detecting a real image, using either film or an electronic detector, the aperture of the hologram is the limiting aperture, and together with the distance of the image from the hologram, defines the size of the speckles. If D is the size of the hologram, and z_i is the image distance, then the speckle size is of the order of the diffraction limit, $\lambda z_i / D$. Speckle has been found to reduce the detectability of image detail by a significant factor, particularly when the size of that detail is comparable with the speckle size. Its effects can be suppressed only by smoothing or averaging the intensity over several speckle sizes, for example with detector elements that are several times larger than the diffraction limit. However, if such smoothing is performed, the resolution in the image is reduced accordingly, so whether the reduction of detail comes from smoothing in the detection process or from the inability of the human visual system to extract details when their size is comparable with a speckle, the results are essentially the same.

For a more detailed discussion of the effects of speckle on the ability of the human observer to resolve image detail, see [12] and the references contained therein.

11.11 Digital Holography

Photographic film and plates were the most common materials used for detection of holograms in the early days of the field. As we have seen, such materials require wet processing and, as a consequence, there is a significant time delay between the recording and reconstruction steps. The advantage of photographic materials, however, is their ability to record extremely fine detail. For example, a $4'' \times 5''$ $4'' \times 5''$ plate with 3000 cycles/mm resolution can record in excess of 35 billion resolvable pixels.

As early as the 1960's, experiments were performed in which holograms were detected on electronic detectors [104], vidicons in these cases because CCD detectors and CMOS detectors with adequate resolution were developed much later. In one case, not only was the hologram detected electronically, the image was formed digitally using the newly popular FFT algorithm [140].

With the development of high-resolution CCD and CMOS electronic detectors for digital photography, the possibility of recording holograms on electronic detectors and reconstructing images digitally gained momentum. Such an approach is free from wet processing and can have short delay times between recording and reconstruction of images. We can identify two different forms of digital holography, one that uses the usual angularly-offset reference wave, which we shall call "offset reference-wave digital holography," and one that uses a sequence of measurements with an in-line reference wave, which we shall refer to as "phase-shifting digital holography." We consider both of these cases in what follows.

11.11.1 Offset Reference-Wave Digital Holography

To record a hologram with an angularly offset reference wave in the usual Leith-Upatnieks geometry, the resolution of the detector must be high enough to accommodate the high-frequency carrier on which amplitude and phase modulations ride. Thus detector pixel counts that are higher than those required to record a two-dimensional image of the same object are needed. Suppose that the object wave at the hologram plane must contain $N \times N$ $N \times N$ detector pixels for adequate resolution. Then in the direction of the offset reference wave the hologram in principle must contain $4N$ $4N$ pixels, $2N$ $2N$ for the autocorrelation function of the object wave, and N N for each of the waves that generate the twin images. In the direction orthogonal to the direction of the offset reference wave, only $2N$ $2N$ pixels are required, as determined by the autocorrelation of the object wave.

In the event that is possible to use a strong reference wave, when compared with the strength of the object wave, the autocorrelation term may be vanishingly weak, in which case in the direction of the offset reference wave, only $2N$ $2N$ pixels are needed, N N for each of waves that lead to the twin images. In the direction orthogonal to the direction of the reference wave offset, only N N pixels are required.

The relation between the number of pixels required to represent a hologram and the number of pixels required to represent the object itself was explored by Macovski [237]. We will not present Macovski's results in detail here but only mention that minimum pixel requirements for the hologram occur in the Fourier transform recording geometry. While a price in terms of

required pixels is paid for using offset reference-wave holography for detecting the hologram, this type of hologram has an advantage for dynamic objects, for the hologram can be recorded with a single fast laser pulse and the image formed digitally for examination.

11.11.2 Phase-Shifting Digital Holography

A solution that eliminated the need for an angularly offset reference wave and its concomitant high-frequency carrier was derived from work of Carré [52], Bruning et al. [46] and Creath [79] in the field of interferometry and applied in the late 1990s by Yamaguchi and Zhang [378] to general holography, and in particular holographic microscopy [382]. In effect, this technique removes the requirement that the reference wave be angularly offset, at the sometimes tolerable cost of requiring a short sequence of holograms of the same object with on-axis reference waves that are changed in phase between exposures. The various holograms recorded electronically in sequence are then combined digitally in such a way that the amplitude and phase of the wave incident on the detector can be recovered, following which an image of the object can be reconstructed digitally.

The detector pixel count for an object wave requiring $N \times N$ samples is $N \times N$ for each of the recorded holograms. As described below, a minimum of 4 measurements must be made, and more often 5 are used.

To understand how this technique is able to recover the amplitude and phase of the incident object wave, start with (11-3), which we repeat here with only a slight modification:

$$\begin{aligned} \mathcal{I}\psi(x, y) &= |A(x, y)|^2 + |a(x, y)|^2 + 2|A(x, y)||a(x, y)|\cos[\phi(x, y) - \psi(x, y)]. \\ \mathcal{I}_\psi(x, y) &= |A(x, y)|^2 + |a(x, y)|^2 + 2|A(x, y)| |a(x, y)| \cos[\phi(x, y) - \psi(x, y)]. \end{aligned} \quad (11-106)$$

Here A and ψ are the amplitude and phase of the reference wave, while a and ϕ are the amplitude and phase of the object wave. Let the reference wave be a normally-incident plane wave with uniform phase ψ , and constant amplitude A_0 . Then

$$\begin{aligned} I\psi(x, y) &= A_0^2 + |a(x, y)|^2 + 2A_0|a(x, y)|\cos[\phi(x, y) - \psi]. \\ I_\psi(x, y) &= A_0^2 + |a(x, y)|^2 + 2A_0|a(x, y)|\cos[\phi(x, y) - \psi]. \end{aligned} \quad (11-107)$$

Our goal is to determine $a(x, y)$ and $\phi(x, y)$. The amplitude modulation $a(x, y)$ is a non-negative quantity, with all sign changes absorbed in the phase $\phi(x, y)$, and therefore $a(x, y)$ can be determined by blocking the reference wave ($A_0 = 0$) and recording the intensity incident on the detector. After sampling by the detector pixels and digitization of the result, a sampled version of $a(x, y)$ can be obtained by taking the positive square root of the samples.

To determine the object phase $\phi(x, y)$, a phase modulator is introduced in the reference beam (before it is combined with the object beam), and the reference phase ϕ is

stepped through different values, for example $0, \pi/2, \pi, 3\pi/2$. For each value of the reference phase ψ , a hologram is recorded and digitized. The intensity distributions incident on the detector in the four cases are therefore

$$\begin{aligned} I_0(x,y) &= A_0 + a(x,y)2 + 2A_0a(x,y)\cos\phi(x,y) \\ I_{\pi/2}(x,y) &= A_0 + a(x,y)2 + 2A_0a(x,y)\cos\phi(x,y)- \\ &\quad \pi \\ I_{3\pi/2}(x,y) &= A_0 + a(x,y)2 + 2A_0a(x,y)\cos\phi(x,y)-3\pi/2 \end{aligned}$$

$$\begin{aligned} I_0(x,y) &= A_0 + |a(x,y)|^2 + 2A_0 a(x,y)\cos[\phi(x,y)] \\ I_{\pi/2}(x,y) &= A_0 + |a(x,y)|^2 + 2A_0 a(x,y)\cos[\phi(x,y) - \pi/2] \\ I_\pi(x,y) &= A_0 + |a(x,y)|^2 + 2A_0 a(x,y)\cos[\phi(x,y) - \pi] \\ I_{3\pi/2}(x,y) &= A_0 + |a(x,y)|^2 + 2A_0 a(x,y)\cos[\phi(x,y) - 3\pi/2] \end{aligned} \tag{11-108}$$

The object phase can be obtained, for example, from the equation

$$\phi = \tan^{-1} \frac{I_{\pi/2} - I_{3\pi/2}}{I_0 - I_\pi}$$

$$\phi = \tan^{-1} \left\{ \frac{I_{\pi/2} - I_{3\pi/2}}{I_0 - I_\pi} \right\} \tag{11-109}$$

Other choices of the number of phase steps and their values are possible, with some offering advantages with respect to noise [310]. For example, 3 phase steps and a measurement of $|a(x,y)|^2$ can be used for a total of four recordings, although the more measurements used, in general the better the noise performance.

Digital holography, in some cases with an angularly offset reference wave and in some cases by phase stepping, is now used commonly in applications such as three-dimensional holographic microscopy, with reconstruction of images from measured amplitude and phase performed digitally. Commercially available digital holographic microscopes are available from, for example, Lycée tec.

11.12 Holography with Spatially Incoherent Light

While holography was originally conceived as a means for coherent image formation, certain techniques exist by means of which holograms of incoherently illuminated objects can be recorded. The extension of holographic techniques to the incoherent case was first suggested by [Mertz and Young \[251\]](#). The theory and practice of incoherent holography were later extended by [Lohmann \[231\]](#), [Stroke and Restrick \[335\]](#), and [Cochran \[73\]](#). For additional relevant information, see the book by [Rogers \[299\]](#).

The light from any one point on a spatially incoherent object will not interfere with the light from any other point. Nonetheless, if by means of some suitable optical trick the light from each object point is split into two parts, then it is possible for each pair of waves of common origin to interfere and form a fringe pattern. Thus each object point may be encoded in a suitable pattern of fringes, and if the encoding is a unique one, with no two object points generating identical fringe patterns, then in principle an image of the object can be obtained.

While many optical systems for achieving the required splitting of the object waves are known, we illustrate here with one particular system suggested by [Cochran \[73\]](#). As shown in Fig. [11.45](#), the system consists of a triangular interferometer, in which are placed two lenses L_1 and L_2 with different focal lengths f_1 and f_2 . We assume that both lenses are positive, although a combination of one positive and one negative lens may also be used. The lenses are separated by a path length $f_1 + f_2$, their focal points coinciding at the point P in the figure. Plane A and plane B both lie at path length f_1 from lens L_1 and path length f_2 from L_2 .

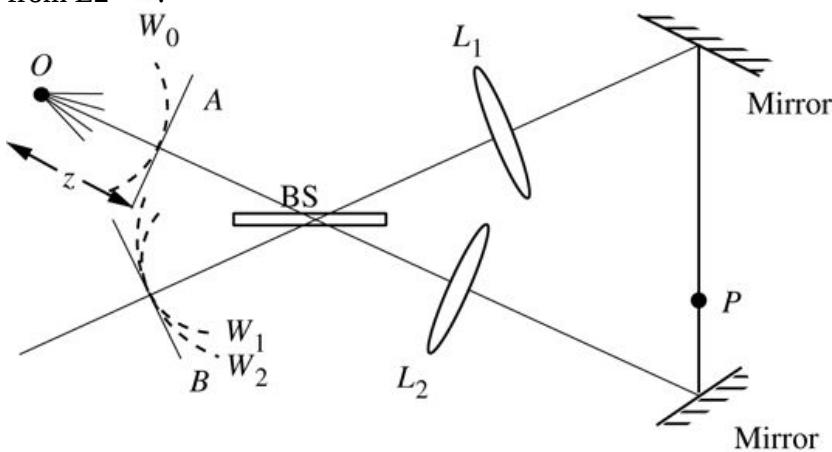


Figure 11.45
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.45 Triangular interferometer for incoherent holography.

The illustration shows two slanting small lines at the extreme right labeled mirror, one at the top and one at the bottom. A line from bottom left points to the center of the upper mirror. A dark spot

labeled 0 is shown on the upper left and small lines emerge from the dot and face in different directions. A line from the dark spot point toward the center of the bottom mirror. A line connects the center of top mirror with the center of bottom mirror and a dark spot labeled P is shown on the line slightly above its lower end. A lengthy rectangular bar labeled BS is shown at the center which is the meeting point of the two lines. Lens L1 is at a point slightly close to the right end of the line that starts from bottom left. Lens L2 is at a point slightly close to the right end of the line that starts from the dark spot on top left. A small slanting vertical line labeled A is on the line that starts from top left and points to the bottom mirror. Line A is at a distance z from the dark spot 0. A dashed line arc labeled W0 is shown with its base resting on Line A. A small slanting vertical line labeled B is on the line that starts from bottom left and points to the top mirror. Two dashed line arcs labeled W1 and W2 are shown with its base resting on line B.

Light may travel from plane A A to plane B B along either of two paths, one clockwise around the interferometer and the second counterclockwise. Considering first the clockwise path, light travels a distance $f_1 f_1$ from plane A A to lens L1 L_1 by means of a reflection at the beam splitter BS BS . From L1 L_1 to L2 L_2 the path length is $f_1 + f_2 f_1 + f_2$, and from L2 L_2 to plane B B (again by means of reflection at BS BS) the path length is $f_2 f_2$. Because of the particular choice of the path lengths in relation to the focal lengths $f_1 f_1$ and $f_2 f_2$, plane A A is *imaged* onto plane B B ; due to the particular sequence in which L1 L_1 and L2 L_2 are encountered on this path, the imaging is performed with magnification $M_1 = -f_2/f_1 M_1 = -f_2/f_1$.

For the counterclockwise path, light is in each case transmitted (rather than reflected) by the beam splitter. Again plane A A is imaged onto plane B B , but for this path the lenses are encountered in opposite sequence, and the magnification is $M_2 = -f_1/f_2 M_2 = -f_1/f_2$.

Consider now the single point O O (see Fig. 11.45) of an incoherent object located at distance $z z$ from plane A A . Regarding the light from that one point as providing a phase reference, we may express the resulting spherical wave (wavefront W0 W_0 in the figure) incident on plane A A as the complex function

$$U_a(x, y) = U_o \exp j\pi \lambda zx^2 + y^2,$$

$$U_a(x, y) = U_o \exp \left[j \frac{\pi}{\lambda z} (x^2 + y^2) \right],$$

(11-110)

where a paraxial approximation has been used. At plane B B we find two spherical waves (wavefronts W1 W_1 and W2 W_2 in the figure), one magnified by $M_1 M_1$ and the second by $M_2 M_2$. Thus the total amplitude is

$$U_b(x, y) = U_1 \exp j\pi \lambda zx M_1^2 + y M_1^2 + U_2 \exp j\pi \lambda zx M_2^2 + y M_2^2.$$

$$U_b(x, y) = U_1 \exp \left\{ j \frac{\pi}{\lambda z} \left[\left(\frac{x}{M_1} \right)^2 + \left(\frac{y}{M_1} \right)^2 \right] \right\} \\ + U_2 \exp \left\{ j \frac{\pi}{\lambda z} \left[\left(\frac{x}{M_2} \right)^2 + \left(\frac{y}{M_2} \right)^2 \right] \right\}. \\ (11-111)$$

The corresponding intensity distribution is

$$J(x, y) = |U_1|^2 + |U_2|^2 + 2|U_1||U_2| \cos \pi \lambda z f_1^4 - f_2^4 / (f_1^2 f_2^2) (x^2 + y^2),$$

$$J(x, y) = \left| U_1 \right|^2 + \left| U_2 \right|^2 + 2 \left| U_1 \right| \left| U_2 \right| \cos \left[\frac{\pi}{\lambda z} \left(\frac{f_1^4 - f_2^4}{f_1^2 f_2^2} \right) (x^2 + y^2) \right], \\ (11-112)$$

where we have used the relation

$$1M12 - 1M22 = f14 - f24f12f22.$$

$$\frac{1}{M_1^2} - \frac{1}{M_2^2} = \frac{f_1^4 - f_2^4}{f_1^2 f_2^2}.$$

If a photographic plate is exposed by the intensity pattern of [Eq.\(11-112\)](#), and processed to produce a positive transparency with amplitude transmittance linearly proportional to exposure, the resulting transmittance may be written

$$t_A(x, y) = t_b + \beta' U_1 U_2^* \exp \left\{ j \left[\frac{\pi}{\lambda z} \left(\frac{f_1^4 - f_2^4}{f_1^2 f_2^2} \right) (x^2 + y^2) \right] \right\} \\ + \beta' U_1^* U_2 \exp \left\{ -j \left[\frac{\pi}{\lambda z} \left(\frac{f_1^4 - f_2^4}{f_1^2 f_2^2} \right) (x^2 + y^2) \right] \right\}. \\ (11-113)$$

We recognize the second and third terms as the transmittance functions of a negative and positive lens, respectively (cf. [Eq.\(5-10\)](#)), each of focal length

$$f = f12f22f14-f24z.$$

$$f = \frac{f_1^2 f_2^2}{f_1^4 - f_2^4} z.$$

$$(11-114)$$

Thus if the transparency is illuminated by a coherent source, both a virtual and a real image of the original object will be formed.

Generalizing now to an object consisting of a multitude of mutually incoherent point sources, each point source generates its own fringe pattern on the recording medium. Since the various sources are not coherent, the total intensity is found simply by adding the various intensity patterns so generated. The (x, y) coordinates of each point source determine the center of the corresponding pattern of fringes, and therefore fix the (x, y) coordinates of the real and virtual images. Similarly, the z coordinate of the point source influences the focal length of its contribution to the transmittance function, as seen in (11-114), and the image formed is thus a three-dimensional one.

Although the possibility of using incoherent illumination, rather than coherent illumination, is an attractive one in many applications, there exists one serious problem that limits the usefulness of incoherent holography. The problem arises because each elementary fringe pattern is formed by two extremely tiny portions of the light incident on the recording medium. Whereas for *coherent* holography light from each object point interferes with all the light contributed by the reference wave, for *incoherent* holography, the interfering waves represent only a minute fraction of the total light. The summation of many weak interference patterns, each with its own bias level of exposure, results in a very large bias level, in general much larger than for a hologram of a similar object formed with coherent light. As a consequence of this bias problem, incoherent holography has been successfully applied only to objects composed of small numbers of resolution elements. This limitation restricts its use significantly.

11.13 Applications of Holography

Holography is a mature scientific field: most of the basic science has been done, and the techniques have undergone a great deal of refinement. During this process, a multitude of applications have been explored, some leading to highly successful businesses, others to important diagnostic tools that are widely used in some branches of both science and engineering. In this section we present a brief summary of the major applications to date.

11.13.1 Microscopy and High-Resolution Volume Imagery

From an historical perspective, microscopy has been the application of holography which has motivated much of the early work on wavefront reconstruction; it was certainly the chief motivating force behind the early works of [Gabor \[120\]](#), [\[121\]](#), [\[122\]](#) and [El-Sum \[103\]](#). Interest in applications to electron microscopy has remained (cf. [\[346\]](#)), and interest in extending holographic microscopy to the X-ray region of the spectrum remains strong as well [\[246\]](#). Interest in both electron and X-ray holography is motivated by the potential for achieving extremely high resolutions, comparable with the wavelength in each case.

In the visible region of the spectrum, holography is not a serious competitor with the conventional microscope in ordinary, run-of-the-mill microscopy. Nonetheless there does exist one area in which holography offers a unique potential to microscopy, namely in *high-resolution volume imagery*. In conventional microscopy, high lateral resolution is achieved only at the price of a limited depth of focus. As seen in [Chapter 7](#), the best lateral resolution achievable by an imaging system is of the order of λ/NA (cf. 7-50), where NA is the numerical aperture. We have seen that with this lateral resolution comes a depth of focus that is limited to an axial distance on the order of λ/NA^2 . Note that for numerical apertures approaching unity, the depth of focus becomes as shallow as one wavelength! Thus there is a limited volume that can be brought into focus at one time.

It is, of course, possible to explore a large volume *in sequence*, by continuously refocusing to explore new regions of the object volume, but such an approach is often unsatisfactory if the object is a dynamic one, continuously in motion.

A solution to these problems can be obtained by recording a hologram of the object using a pulsed laser to obtain a very short exposure time. The dynamic object is then “frozen” in time, but the recording retains all the information necessary to explore the full object volume. If the hologram is illuminated, the real or virtual image can be explored in depth with an auxiliary optical system. Sequential observation of the image volume is now acceptable because the object (i.e. the holographic image) is no longer dynamic.

This approach was fruitfully applied by C. Knox in the microscopy of three-dimensional volumes of living biological specimens [\[200\]](#), and by Thompson, Ward, and Zinky in measurement of the particle-size distributions in aerosols [\[341\]](#). The reader may consult these references for further details.

In recent years, digital holographic microscopes have been introduced, particularly by the company Lyncée tec (see, for example, [\[293\]](#)). Such microscopes detect the hologram on an electronic detector and compute the image in any plane digitally.

11.13.2 Interferometry

Some of the most important scientific applications of holography have proven to arise from the unique modalities of interferometry that it offers. Holographic interferometry can take many different forms, but all are dependent on the ability of a hologram to store two or more separate complex wave fields on the same recording medium, and the subsequent interference of those fields when they are reconstructed together. More detailed treatments of holographic interferometry can be found, for example, in the books by [Vest \[357\]](#) and [Schumann \[311\]](#).

Multiple-Exposure Holographic Interferometry

The most powerful holographic interferometry techniques are based on a property, emphasized by [Gabor et al. \[125\]](#), that, by means of multiple exposures of holograms, coherent additions of complex wavefronts can be achieved. This property can easily be demonstrated as follows: let a photographic recording material be exposed sequentially by N^N different intensity distributions $J_1, J_2, \dots, J_N, I_1, I_2, \dots, I_N$. The total exposure to which the medium has been subjected can be written

$$E = \sum_{k=1}^N T_k |A| + \sum_{k=1}^N T_k |a_k|^2 + \sum_{k=1}^N T_k A^* a_k + \sum_{k=1}^N T_k A a_k^*.$$

$$E = \sum_{k=1}^N T_k |A|^2 + \sum_{k=1}^N T_k |a_k|^2 + \sum_{k=1}^N T_k A^* a_k + \sum_{k=1}^N T_k A a_k^*.$$

(11-116)

Assuming linear operation in the t_A versus E characteristic of the recording medium, we find components of transmittance

$$t_\alpha = \beta \sum_{k=1}^N T_k A^* a_k$$

$$t_\alpha = \beta \sum_{k=1}^N T_k A^* a_k$$

$$t_\beta = \beta \sum_{k=1}^N T_k A a_k^*$$

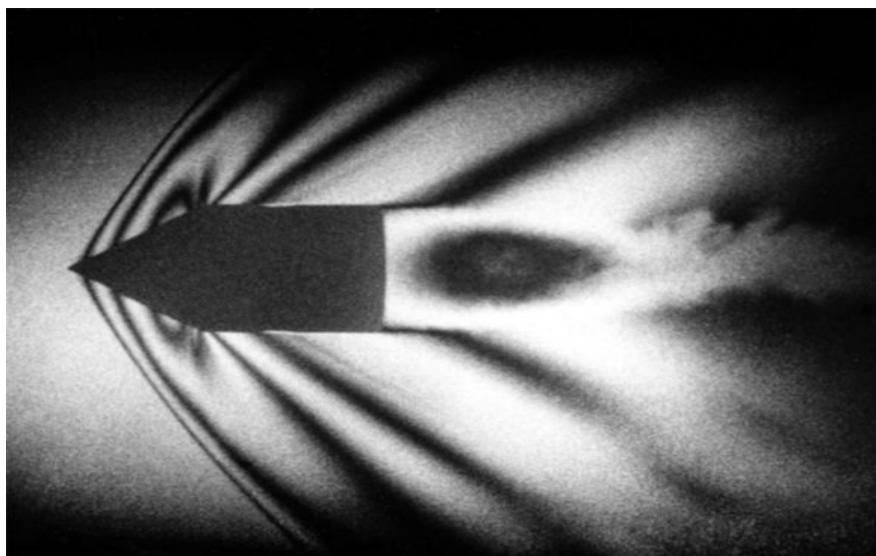
$$t_\beta = \beta \sum_{k=1}^N T_k A a_k^*$$

(11-117)

From (11-117) it is clear that illumination of the processed hologram with a wavefront A will generate a transmitted field component proportional to the product of $|A|^2 |A|^2$ and the sum of the complex wavefronts a_1, a_2, \dots, a_N . As a consequence, N^N coherent virtual images of the objects that gave rise to the N^N wavefronts will be linearly superimposed and will

mutually interfere. In a similar fashion, illumination of the transparency by a wavefront $A^* A^*$ will generate N^N coherent real images which likewise interfere.

The earliest dramatic demonstrations of the potential of this type of interferometry were performed by [Brooks et al. \[43\]](#) using a Q-switched ruby laser. [Figure 11.46](#) shows two photographs obtained in each case by double exposure of a hologram with two laser pulses. In the case of part (a) of the figure, the first pulse records a hologram of only a diffuse background, while the second pulse records a hologram of a bullet in flight in front of the same diffuse background. The shock waves generated by the bullet produce changes in the local refractive index of the air. As a consequence, the two images of the diffuse background, one recorded in the absence of the bullet and the other recorded through the refractive-index perturbations of the air, will mutually interfere, producing interference fringes that outline the shock waves generated by the bullet. These fringes have the appearance of being fixed in three-dimensional space around the bullet.



(a)



(b)

Figure 11.46
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.46 Double-exposure holographic interferometry with a Q-switched ruby laser. [Reproduced from L.O. Hefflinger, R.F. Wuerker and R.E. Brooks, “Holographic Interferometry”, *J. Appl. Phys.* 37, 642–649 (1966) with the permission of AIP Publishing]

Image a shows a square patch with dark corners. At the center of the square is the holographic image of a bullet with a diffused background. Image b shows a vertical rectangular box with shaded corners and a bright center portion. A holographic image of an incandescent bulb is shown at the center.

Part (b) of the same figure is a similarly obtained image of an incandescent bulb. During the first exposure the filament is off, and again a hologram of a diffuse background is recorded, this time through the imperfect glass envelope of the bulb. The filament is then turned on, and a second

laser pulse exposes the hologram. The incoherent light generated by the lamp does not interfere with the laser light, so the filament does not appear lighted in the final image. However, the heating of the gases within the envelope has resulted in changes of the local index of refraction, which again generate fringes of interference in the final image, outlining the patterns of gas expansion. It should be emphasized that these interference fringes have been obtained in the presence of the optically imperfect glass envelope, a feat which would be impossible by other classical methods of interferometry.

Real-Time Holographic Interferometry

Another important type of holographic interferometry depends on interference between a prerecorded, holographically produced wavefront and the coherent waves reflected from or transmitted by the same object in real time [42]. The holographically produced wavefront can be regarded as a reference, representing the reflected or transmitted light when the object is in a “relaxed” state. If the same object, located in the same position relative to the hologram that it occupied when the reference wavefront was recorded, is now perturbed, perhaps by placing it under stress with some form of loading, then the complex fields intercepted from the object change, and upon interference with the reference wavefront, produce fringes that are visible on the image of the object seen through the hologram. Two slightly different coherent images are being superimposed by this process, one the image of the object in its original state, and the second the image of the object in its stressed or modified state. The fringes observed can reveal quantitative information about the nature of the object deformations that have taken place.

Note that we are causing the coherent image of the object *now* to interfere with the coherent image of the object that existed sometime *in the past* (or perhaps, using a computer-generated hologram, with an object that never actually existed previously), a feat that would be impossible to accomplish with conventional interferometry.

Contour Generation

The interference of multiple coherent images described previously has also led to the development of techniques for obtaining three-dimensional images with superimposed constant-range contours. These techniques are applicable to the problems of cross-section tracing and contour mapping. Two distinctly different techniques have been demonstrated by [Hildebrand and Haines \[169\]](#).

In the first of these techniques, the object is illuminated by two mutually coherent but spatially separated point sources. The two object illuminations may be applied simultaneously or the hologram may be double-exposed, with a different position of the object illumination source during each exposure. If the pattern of interference between the two object illumination sources is considered, it is found to consist of interference fringes that follow hyperbolae of constant path-length difference, as shown in [Fig. 11.47](#). If the object is illuminated from the side and the hologram is recorded from above, then depth contours (i.e. the intersection of the object with the hyperbolic fringes) are readily seen on the reconstructed image. Identical results are obtained whether the two object illumination sources were used simultaneously in a single exposure or separately in individual exposures, for in either case the two illuminations add coherently.

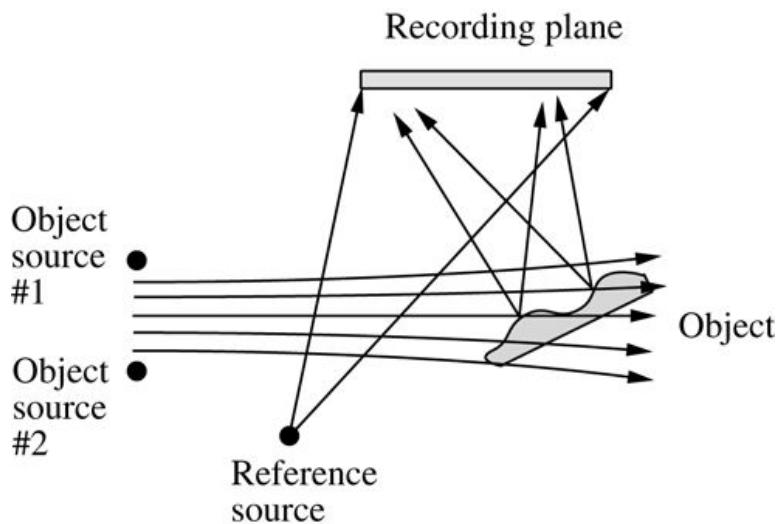


Figure 11.47

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.47 Contour generation by the two-source method.

The illustration shows a lengthy rectangular bar labeled recording plane. A dark spot labeled reference source is at the bottom. Two lines emerge from the reference source point toward the left and right ends of recording plane. To the extreme left are two shaded dots labeled object source 1 and object source 2 which are slightly above the reference source and positioned one above the other. To the extreme right is a slanting irregular shaped structure with two bulges towards its upper and lower ends and is labeled object. An arrow from the right bulge points toward a point on the recording plane slightly before its left end and another arrow from the same point toward a point on the recording plane slightly before its right end. An arrow from the left bulge points toward a point on the recording plane slightly before its left end and another arrow from the same point toward a point on the recording plane slightly before its right end. Five horizontal arrows between the object source 1 and object source 2 point toward the right.

The two-source method of contour generation suffers from the requirement that the directions of illumination and observation must differ by nearly 90° . Thus if the object has significant relief, shadows will be cast and parts of the object will simply not be illuminated. This deficiency is overcome by the two-frequency or two-wavelength method of contour generation. In this case the object and reference illuminations both contain the same two distinct wavelengths, say λ_1 and λ_2 . In effect, each wavelength records a separate and independent hologram on the same recording medium. When the resulting hologram is illuminated by light of a single wavelength, two images with slightly different positions and magnifications are produced. These two images will interfere, and for certain geometries the resulting image contours will be accurate indications of depth. We do not dwell on a detailed analysis of this case; the interested reader may consult the original reference for further details [169]. Figure 11.48 shows the results of contour mapping by the two-wavelength method. In part (a) we see a holographic image of a coin, illuminated in the usual manner with single-wavelength light. When two-wavelength light is used to record the hologram, the image of part (b) is obtained. In this case the two wavelengths were obtained from

two different lines of an argon laser. The two lines are separated by 6.5 nm and the resulting contours on the image are spaced by $20\mu\text{m}$.



(a)



(b)

Figure 11.48

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.48 Contour generation by the two-wavelength method. [From [\[169\]](#). Copyright 1967 by the Optical Society of America, Inc., reprinted with permission.]

Image a shows a square patch inside which is a holographic image of a coin. Image b shows the holographic image of the same coin but is brighter than the first image.

Vibration Analysis

A holographic technique for vibration analysis, first proposed by [Powell and Stetson \[287\]](#), may be regarded as a generalization of multiple-exposure holographic interferometry to the case of a

continuous time exposure of a vibrating object.

With reference to the geometry of [Fig. 11.49](#), we consider a point at coordinates (x_0, y_0) (x_o, y_o) on a planar object which is vibrating sinusoidally with angular frequency Ω . The peak amplitude of the vibration at that point is represented by $m(x_0, y_0)$ $m(x_o, y_o)$, and the fixed phase of the vibration is $\mu(x_0, y_0)$ $\mu(x_o, y_o)$. The light incident at the hologram recording plane coordinates (x, y) (x, y) from that particular point may be regarded as having a time-varying phase modulation

$$\phi(x, y; t) = 2\pi\lambda \cos\theta_1 + \cos\theta_2 m(x_0, y_0) \cos\Omega t + \mu(x_0, y_0),$$

$$\phi(x, y; t) = \frac{2\pi}{\lambda} (\cos\theta_1 + \cos\theta_2) m(x_0, y_0) \cos[\Omega t + \mu(x_0, y_0)],$$

(11-118)

where λ is the optical wavelength of the illuminating source, θ_1 is the angle between the vector displacement of the object at (x_0, y_0) (x_o, y_o) and the line joining that point to (x, y) (x, y) , and θ_2 is the angle between the vector displacement and the direction of propagation of the incident light at (x_0, y_0) (x_o, y_o) .

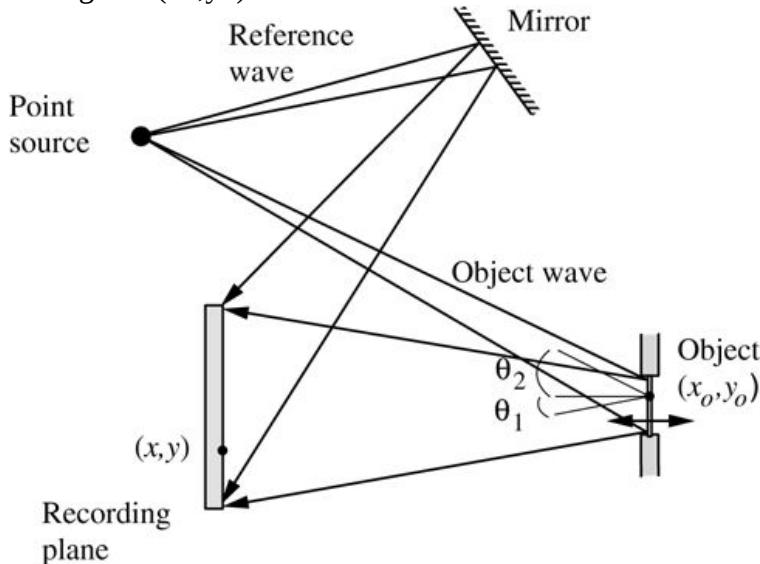


Figure 11.49

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.49 Recording a hologram of a vibrating object.

The illustration shows a vertical rectangular bar at the left bottom labeled recording plane. Above the recording plane is a dark spot labeled point source. On the right, is a slanting line labeled mirror. Two lines from point source labeled reference wave point toward the two points on the mirror near its center and get reflected back toward the top and the bottom ends of recording plane. A dark spot is on the right border of recording plane, slightly above its lower end and is

labeled (x, y) . On the bottom right is a thin vertical strip labeled object and a small vertical rectangular bar is placed above the upper end and below the lower end of the strip. Two lines from the point source point toward the top and bottom ends of the vertical strip. A dark spot is on the vertical strip slightly below the upper end and is labeled (x_0, y_0) . A short horizontal line starts from the dark spot and moves toward the left. Two slanting lines begin from the dark spot, one above the horizontal line and one below the horizontal line. The angle between the line above the horizontal line and the horizontal line is theta 2 and the angle between the line below the horizontal line and the horizontal line is theta 1. A bidirectional horizontal arrow is shown on the vertical strip slightly above its lower end.

Using what by now should be a familiar expansion into Bessel functions, the temporal spectrum of the time-varying phasor representing the modulated light incident at (x, y) can be written

$$F(v) = \mathcal{F}\{\exp[-j\phi(x, y; t)]\} = \sum_{k=-\infty}^{\infty} J_k[2\pi\cos\theta_1 + \cos\theta_2 m(x_o, y_o)] \delta(v - k\Omega/2\pi).$$

$$\begin{aligned} F(\nu) &= \mathcal{F}\{\exp[-j\phi(x, y; t)]\} \\ &= \sum_{k=-\infty}^{\infty} J_k \left[2\pi \frac{\cos\theta_1 + \cos\theta_2}{\lambda} m(x_o, y_o) \right] \delta\left(\nu - \frac{k\Omega}{2\pi}\right). \end{aligned} \quad (11-119)$$

When the exposure time is much longer than the vibration period (which is when $T \gg 2\pi/\Omega$), only the $k=0$ term, which is at the same optical frequency as the reference wave, will cause stable interference fringes to be formed. All other terms will fail to produce such fringes. If the variations of the modulation depth introduced by the term $\cos\theta_1$ are nearly independent of (x, y) (that is, if the angle subtended by the film at (x_o, y_o) is small), then the amplitude of the image at (x_o, y_o) will be suppressed by the factor

$$J_0[2\pi(\cos\theta_1 + \cos\theta_2) m(x_o, y_o)],$$

$$J_0 \left[\frac{2\pi}{\lambda} (\cos\theta_1 + \cos\theta_2) m(x_o, y_o) \right],$$

$$(11-120)$$

and the intensity will be suppressed by the square of this factor. Thus the intensity of the image depends at each point on the depth of vibration of the corresponding object point.

[Figure 11.50](#) shows images of a vibrating diaphragm obtained experimentally by Powell and Stetson. In part (a) of the figure, the diaphragm is vibrating in its lowest-order mode, with a single vibration maximum at the center of the diaphragm. In part (b), the diaphragm is vibrating in a higher-order mode, with two vibration maxima. By counting the number of fringes from the center of the diaphragm to any point in question, it is possible, with the help of [\(11-120\)](#), to determine the vibration amplitude at that point.

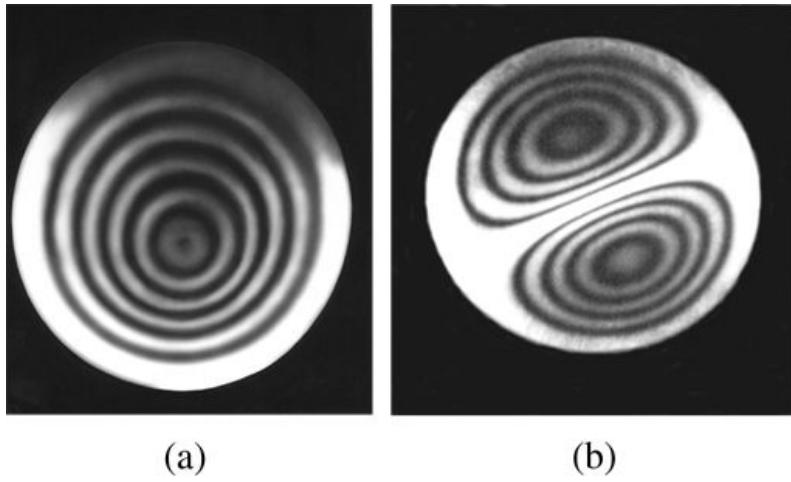


Figure 11.50
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.50 Holographic images of a diaphragm vibrating in two different modes. [From [\[287\]](#). Copyright 1965 by the Optical Society of America, Inc., reprinted with permission.]

Image a shows a square patch of darkness. A bright circular shaped structure is at the center and at the center of it is a dark spot with concentric circles surrounding it. Image b shows a square patch of darkness. A bright circular shaped structure is at the center and two concentric semicircles, one in the upper half and another one in the lower half of the circular structure are shown with a thin gap between them.

11.13.3 Imaging through Distorting Media

In many cases of practical interest, an optical system may be required to form images in the presence of uncontrollable aberrations. These aberrations may result from imperfections of the image-forming components themselves, or they may be introduced by an external medium, such as the Earth's atmosphere. The techniques of holography offer several unique advantages for problems of imaging in the presence of such aberrations. We discuss here three distinctly different holographic techniques for obtaining high resolution in the presence of severe aberrations.

The first technique ([\[225\]](#), [\[201\]](#)) of interest is applicable only when the distorting medium is constant in time. As illustrated in [Fig. 11.51](#), a hologram of the distorted object waves is recorded with an undistorted reference wave. The processed hologram is then illuminated with an “anti-reference” wave, i.e. a reconstruction wave that duplicates the reference wave but propagates in the reverse direction. A real, conjugate image of the distorting medium will form precisely at the location of the medium itself, between the hologram and the image plane. If the object wave incident on the distorting medium during the recording process is represented by $U_o(\xi, \eta)$

$U_o(\xi, \eta)$ and if the amplitude of transmittance of the distorting medium is $\exp[jW(\xi, \eta)]$ $\exp[jW(\xi, \eta)]$, then the wave falling on the distorting medium during reconstruction is $U_o^*(\xi, \eta) \exp[-jW(\xi, \eta)]$ $U_o^*(\xi, \eta) \exp[-jW(\xi, \eta)]$. Note that when this conjugate wave passes back through the identically same distorting medium that was originally present, the aberrations entirely cancel,

$$U_o^*(\xi, \eta) \exp[-jW(\xi, \eta)] \exp[jW(\xi, \eta)] = U_o^*(\xi, \eta),$$

$$U_o^*(\xi, \eta) \exp[-jW(\xi, \eta)] \exp[jW(\xi, \eta)] = U_o^*(\xi, \eta),$$

leaving a wave $U_o^*(\xi, \eta)$ to propagate on to the image plane, where an aberration-free image appears.

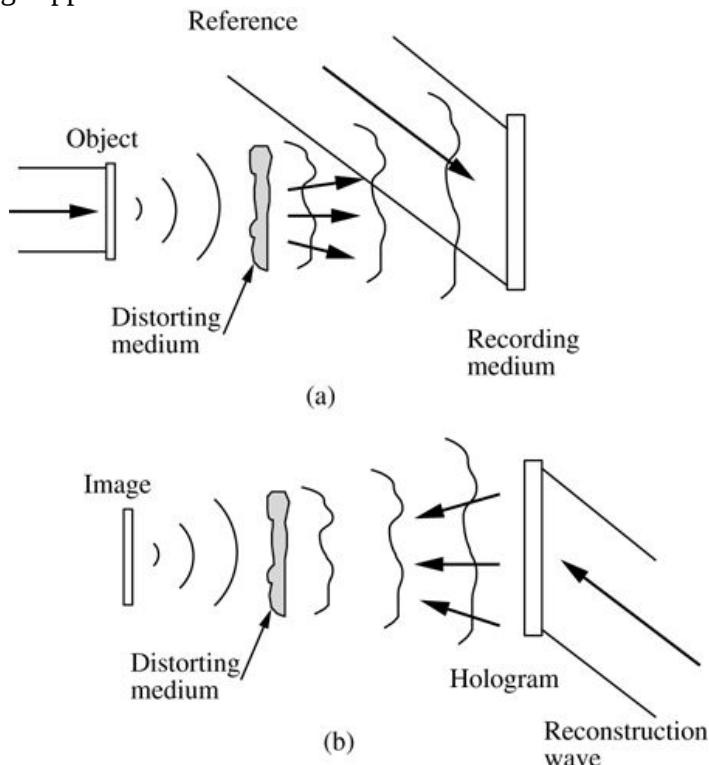


Figure 11.51

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.51 Use of the original distorting medium for compensating aberrations. (a) Recording the hologram and (b) reconstructing the image.

Illustration a shows a thin vertical strip on the extreme left. Two horizontal lines point toward the upper and lower ends of the vertical strip and this structure is labeled object. A horizontal arrow from left pointing toward the right passes through the object. At the right extreme is a vertical rectangular plate labeled recording medium. Between object and recording medium is an irregular shaped structure labeled distorting medium. Three waves from object move toward the distorting medium and three irregular shaped waves from distorting medium move toward recording medium. Three horizontal arrows from distorting medium point toward the recording medium. Two slanting lines labeled reference from top point toward the upper and lower ends of recording medium. A slanting arrow passes between the slanting lines and points toward the center of recording medium.

Illustration b shows a thin vertical strip on the extreme left labeled image. At the right extreme is a vertical rectangular plate labeled hologram. Between image and hologram is an irregular shaped structure labeled distorting medium. Three waves from image move toward the distorting medium and three irregular shaped waves from distorting medium move toward hologram. Three horizontal arrows from hologram point toward the distorting medium. Two slanting lines labeled

reconstruction wave from bottom right point toward the upper and lower ends of hologram. A slanting arrow passes between the slanting lines and points toward the center of hologram.

A limitation of the technique in some applications is that the image must appear where the object originally was situated, whereas in practice it is often desired to obtain an image on the other side of the distorting medium (i.e. to the right of the distorting medium in Fig. 11.51). If the distorting medium is movable, then this difficulty can be overcome.

A second technique of interest is illustrated in Fig. 11.52. Again the distorting medium should be unchanging in time. In this case we record a hologram of the distorted waves transmitted by the medium when it is illuminated by a simple point source (i.e. a record of the point response of the medium). This hologram may now be used as a “compensating plate” to enable a more conventional optical system to form an aberration-free image. Let the waves incident on the recording medium due to the point source be represented by $\exp[jW(x,y)]$. We have assumed that the distorting medium is such that only phase distortions are introduced. The portion of the hologram amplitude transmittance that normally contributes the real image is proportional to $\exp[-jW(x,y)]$. Thus if we replace the point source by a more general object, and reinsert the hologram in the same position it originally occupied, we find that the curvatures of the object waves reaching the hologram are canceled on passage through the hologram, with the waves from different object points producing plane waves traveling at different angles. The lens then forms a distortion-free image in its focal plane.

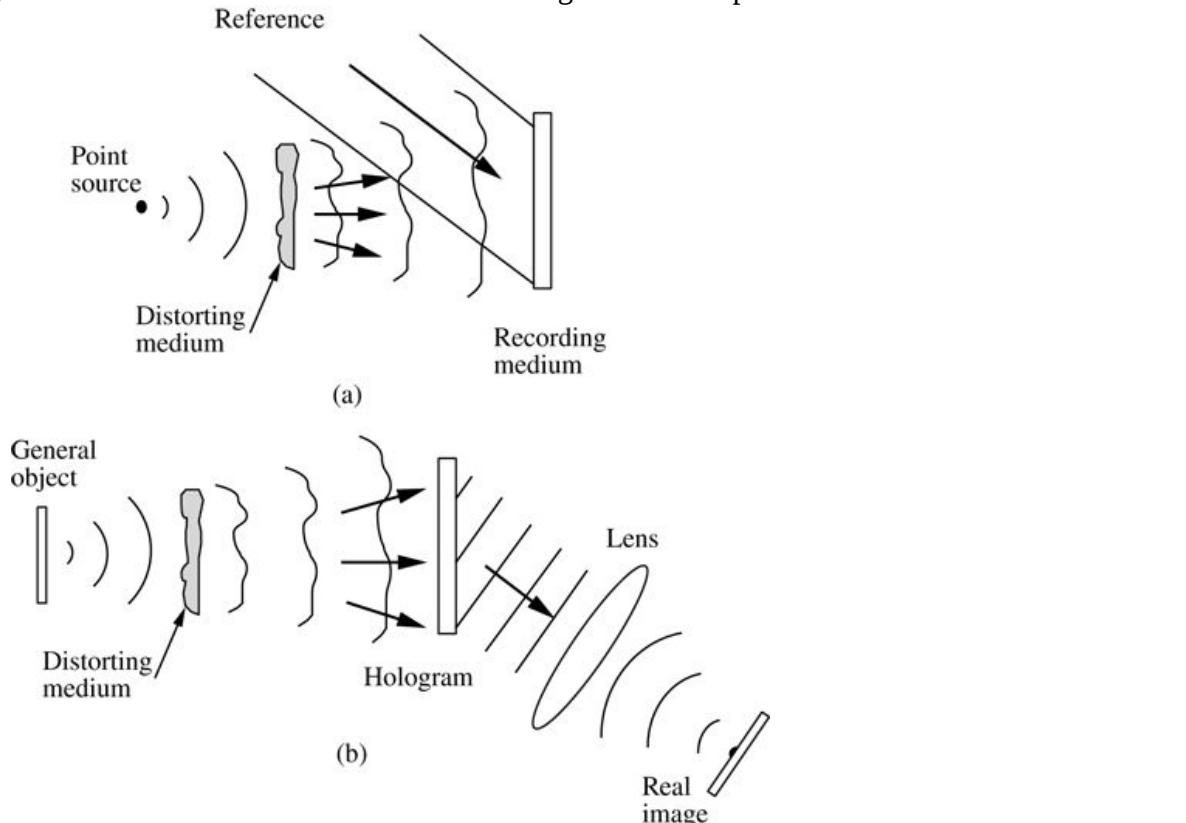


Figure 11.52
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.52 Use of a hologram compensating plate. (a) Recording the compensating plate; (b) cancellation of the aberrations.

Illustration a shows a dark spot labeled point source at the extreme left. At the right extreme is a vertical rectangular plate labeled recording medium. Between object and recording medium is an irregular shaped structure labeled distorting medium. Three waves from the object move toward the distorting medium and three irregular shaped waves from distorting medium move toward recording medium. Three horizontal arrows from distorting medium point toward recording medium. Two slanting lines labeled reference from top point toward the upper and lower ends of the recording medium. A slanting arrow passes between the slanting lines and points toward the center of recording medium.

Illustration b shows a thin vertical strip on the extreme left labeled general object. At the right is a vertical rectangular plate labeled hologram. Between image and hologram is an irregular shaped structure labeled distorting medium. Three waves from general object move toward the distorting medium and three irregular shaped waves from distorting medium move toward hologram. On the bottom right is a slanting rectangular strip labeled real image. A dark spot is at its center. Three waves from real image move toward Lens and five slanting lines from lens move toward hologram and the first three lines from the top touch the right surface of hologram. An arrow from hologram points toward lens.

This technique will work well over only a restricted field of view, for if an object point is too far from the position of the original point source used in recording the hologram, the aberrations imparted to its wave may differ from those recorded on the hologram. This restriction is less severe if the hologram is recorded very close to the distorting medium. [Upatnieks et al. \[351\]](#) have successfully applied this technique to the compensation of lens aberrations, an application to which it is well suited.

A third technique, which may be applied to imaging through media that are time-varying or time-invariant, is accomplished by passing *both* the reference wave and the object wave through the same distorting medium [\[137\]](#). As indicated in [Fig. 11.53](#) the lensless Fourier transform recording geometry is most commonly used, with a reference point source existing in the same plane or nearly the same plane as the object of interest. For simplicity it is assumed that the distorting medium is located immediately in front of the recording plane, although this restriction can be relaxed with some loss of the field of view over which compensation is effective. The reference and object waves reaching the recording medium can be written as $A(x,y)\exp[jW(x,y)]$ $A(x, y) \exp [jW(x, y)]$ and $a(x,y)\exp[jW(x,y)]$ $a(x, y) \exp [jW(x, y)]$, where A A and a a are the waves that would have been present in the absence of a distorting medium. Interference of the two distorted waves yields a pattern of intensity that is unaffected by the presence of the distorting medium,

$$J(x,y)=A(x,y)\exp[jW(x,y)]+a(x,y)\exp[jW(x,y)]^2=|A|^2+|a|^2+A^*a+Aa^*,$$

$$\begin{aligned} J(x, y) &= |A(x, y) \exp [jW(x, y)] + a(x, y) \exp [jW(x, y)]|^2 \\ &= |A|^2 + |a|^2 + A^* a + A a^*, \end{aligned}$$

and distortion-free twin images can be obtained from the hologram.

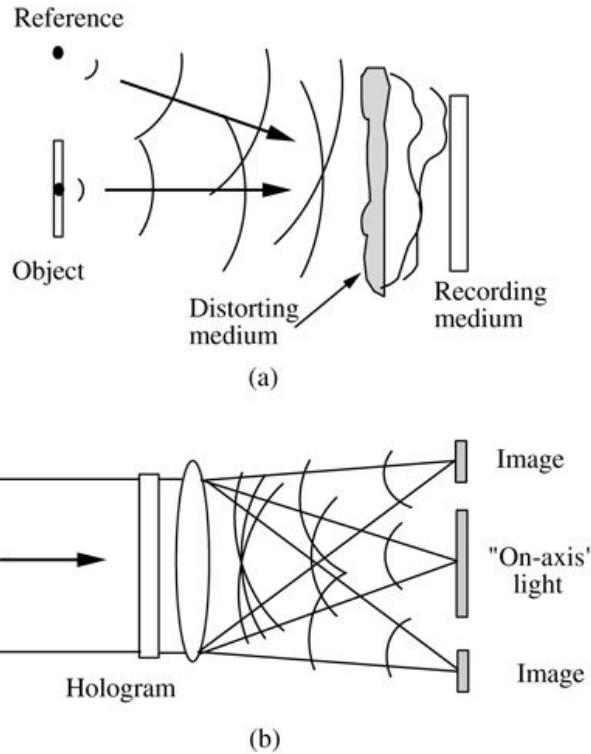


Figure 11.53
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.53 Aberration-free imaging when the object and reference waves are identically distorted. (a) Recording the hologram; (b) obtaining the image.

Illustration a shows a vertical rectangular bar labeled object with a dark spot at the center of it at the extreme left. A dark spot labeled reference at the upper left. At the right extreme is a vertical rectangular plate labeled recording medium. Between object and recording medium is an irregular shaped structure labeled distorting medium. Four waves from object move toward the distorting medium and four waves from reference move toward the distorting medium. The third and fourth waves from object and reference overlap with each other. Two irregular shaped waves from the distorting medium move toward recording medium. Two arrows, one from object and one from reference point toward the distorting medium.

Illustration b shows a thin vertical strip on the extreme left. Two horizontal lines point toward the upper and lower ends of the vertical strip and the structure is labeled hologram. A horizontal arrow from left pointing toward the right passes through center of hologram. Next to hologram is a lens. On the right extreme, a small vertical rectangular structure is shown at the top which is labeled image. Below image, is another lengthy vertical rectangular structure labeled “On-axis” light and below it is another small vertical rectangular bar labeled Image. Two lines from upper and lower ends of hologram, point toward the points near the upper and lower ends of lens, respectively. A line from a point near the upper end of lens point toward the center of upper image and another line from the same point moves toward the center of lower image. A line from a point near the lower end of lens point toward the center of upper image and another line from the same point moves toward the center of lower image. Two lines from upper and lower ends of lens, point toward the center of “On-axis” light. Three waves from upper image move toward lens, three waves from “On-axis” light move toward lens and another three waves from upper image move

toward lens. The second and third layers of waves from the above mentioned three sources overlap with each other.

Again the technique will work over only a limited object field, since points too far from the reference may produce waves with aberrations that differ significantly from those of the reference wave. The working field is largest when the aberrations are introduced close to the recording plane.

This method is an example of a more general set of techniques in optics known as “common path interferometry.” It has been applied to the problem of obtaining high-resolution images of distant objects through the Earth’s atmosphere [\[126\]](#), [\[127\]](#), [\[138\]](#).

11.13.4 Holographic Data Storage

There are many attractive properties of holography as a potential data storage technique, and as a consequence, much attention has been given over the years to this application. Most obvious, perhaps, is the highly diffused nature of holographic storage, in the sense that a single pixel of an analog image or a single bit in a binary data array is stored in a distributed fashion over a considerable area of the hologram. Nonlocalization is most complete for a Fourier transform hologram, and least complete for an image hologram, with Fresnel holograms falling in between these two extremes. When there is a large amount of nonlocalization, a dust speck or a defect in the recording medium that obscures or destroys a very localized area on the hologram will not create a localized defect in the image, and therefore there will not be a localized loss of stored data.

A second advantage, associated particularly with the Fourier transform recording geometry, arises from the fact that a shift in the hologram domain results in only a linear phase tilt in the Fourier domain, and therefore has no effect on the location of the image intensity distribution. As a consequence, Fourier holograms are extremely tolerant to misalignment or registration errors. This property is extremely important for high-density memories, especially those that have high magnification in the sense that a small hologram produces a much larger image.

A third attraction of holography as a storage method comes from our ability to use the third dimension of a three-dimensional recording material, such as a thick recording film or a photorefractive crystal, for recording. Thus holography offers one method of three-dimensional optical storage, and by utilizing the third dimension the volume storage density that can be achieved is quite high.

Early work on holographic storage concentrated on thin holograms and storage of two-dimensional binary arrays [\[8\]](#). [Figure 11.54](#) shows a typical arrangement. Separate two-dimensional pages of binary data are stored in a two-dimensional array of holograms. The light from a CW laser is deflected by a pair of acousto-optic beam deflectors to a particular hologram in the array. The particular hologram selected generates an array of binary spots on a two-dimensional detector array. Thus to determine the state of one particular binary element in the memory, a combination of the right hologram and the right detector element must be interrogated.

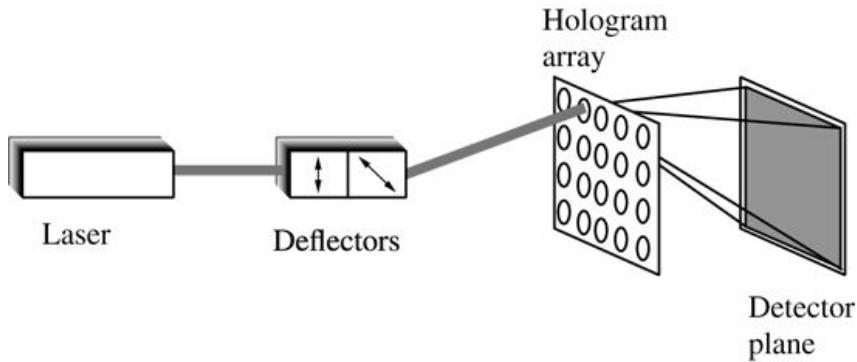


Figure 11.54

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.54 Page-oriented holographic storage.

The illustration starts with a rectangular bar labeled laser. Next to it is another rectangular bar labeled deflectors with a partition at the center. In one partition, is a vertical bidirectional arrow and in another partition is a slanting bidirectional arrow. Next to it is a rectangular plate labeled hologram array with holes arranged in five columns and four rows. Next to the hologram array is another detector plane, which is another rectangular plate. A thick shaded line from laser points toward deflectors and another shaded line from deflectors points toward the hole which corresponds to first row and second column. A line from a point which is slightly right to the midpoint of hologram array points toward the upper left corner of deflector plane and another line from a point near the upper right end of hologram array points toward the upper right corner of deflector plane. Two lines from the lower right and left corners of detector plane point toward two points which are near the upper right corner of hologram array.

More recent emphasis has been on three-dimensional storage media, such as photorefractive crystals (see, for example, [162]), which are capable of providing Bragg selectivity. Multiplexing of holograms within the crystal and selective recall of the data recorded in those holograms can be achieved by means of angle multiplexing, wavelength multiplexing, or multiplexing with phase-coded reference beams. A typical geometry for such a system (from [162]) is shown in Fig. 11.55. A spatial light modulator serves to generate an array of binary data, and the reference wave is introduced at a particular angle designated for that page of data. The reference beams are introduced from the side of the crystal, an orientation that maximizes angular selectivity. The hologram is recorded, and the data can be read out onto a CCD detector array by illumination of the crystal with a duplicate of the reference beam. Other holograms are superimposed in the crystal by use of other reference angles, and must be read out with duplicates of those reference

beams. The diffraction efficiency associated with a single bit falls as $1/N^2$ when N holograms are superimposed, due to the partial erasure of early holograms caused by the recording of later holograms. Superposition of several thousand holograms by angular multiplexing has been demonstrated experimentally [49].

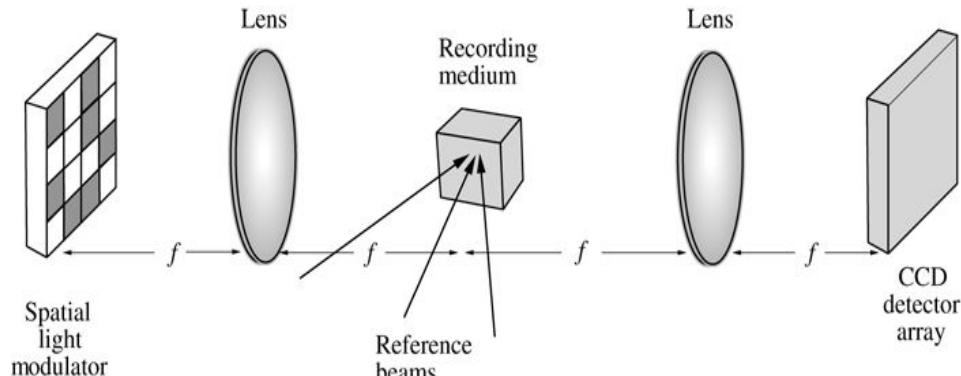


Figure 11.55

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 11.55 A volume holographic storage system. The case of angle multiplexing is illustrated.

The illustration shows a square box labeled spatial light modulator at the extreme left. Inside it are several square partitions wherein few of the partitions are shaded. To its right, is a Lens at a distance f from spatial light modulator. To the right of lens, is a cube labeled recording medium. The distance between Lens and the center of recording medium is f . Three arrows from bottom point toward recording medium and are labeled reference beams. To the right of recording medium is another lens. The distance between the center of recording medium and the second lens is f . To the right of lens is another square shaped structure labeled CCD detector array at a distance f from the second lens.

Finally, mention should be made of the use of holography for associative memories, an idea first described by [Gabor \[123\]](#). Discussion of related ideas can be found in [\[197\]](#), [\[198\]](#), [\[199\]](#), and [\[129\]](#).

11.13.5 Holographic Weights for Artificial Neural Networks

Neural network models provide an interesting and powerful approach to many pattern recognition and associative memory problems. One approach to constructing an artificial “neural” processor is through the use of volume holography. In this section we provide the briefest introduction to this subject, together with references that will allow the reader to pursue the ideas further. The terminology used to describe networks of this type is borrowed from the neurological sciences, but it is important to understand that the models used in artificial neural networks contain only the simplest extraction of the essence of the types of processing that are believed to take place in real biological neural systems. An introduction to the field of neural computing can be found in, for example, [\[164\]](#).

Model of a Neuron

Neural networks consist of a multitude of nonlinear elements referred to as *neurons*, highly interconnected with each other. A simple model of a neuron is illustrated in [Fig. 11.56\(a\)](#). The summation of a multitude of different *weighted* binary inputs is applied to the input of a nonlinear element, usually taken to have a “sigmoid” nonlinear characteristic described by the input-output relation

$$z=g(y)=1/(1+e^{-y}),$$

$$z = g(y) = \frac{1}{1 + e^{-y}},$$

(11-121)

which is illustrated in Fig. 11.56(b).

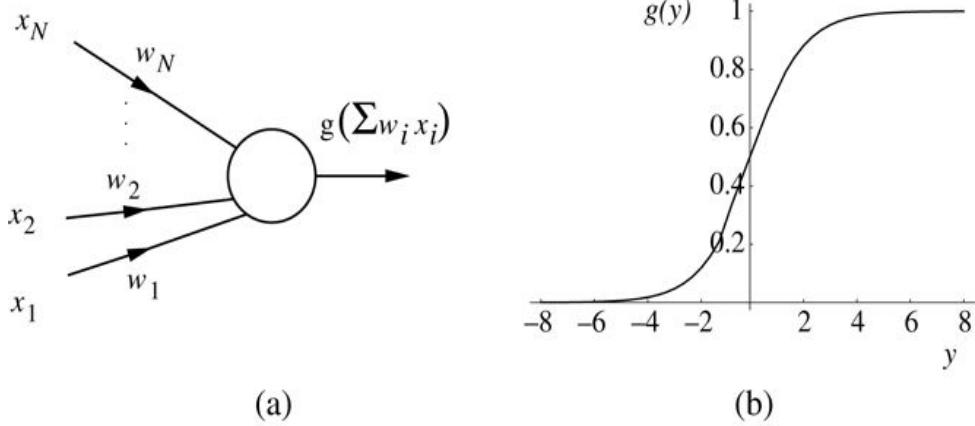


Figure 11.56

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 11.56 (a) Model of a single neuron; (b) sigmoidal nonlinearity.

Illustration a shows a circular shaped structure. A line x_N from top left points toward the circular structure with an arrow facing the circular structure and the text corresponding to the arrow reads w_N . A line x_2 below x_N points toward the circular structure with an arrow facing the circular structure and the text corresponding to the arrow reads w_2 . A line x_1 from bottom left points toward the circular structure with an arrow facing the circular structure and the text corresponding to the arrow reads w_1 . A series of vertical dots are shown between x_N and x_2 . A horizontal arrow from the circle points toward right and is labeled $g(\sum w_i x_i)$.

Illustration b shows a graph with vertical axis labeled $g(y)$ with markings from 0.2 to 1 with equal increments of 0.2 and horizontal axis labeled y with marking from 2 to 8 on positive axis with equal increments of 2 and markings from -2 to -8 on negative axis with equal increments of 2. The curve is s shaped and starts at -8 and moves along the horizontal axis up to -5 after which it increases gradually and reaches the point 1 on vertical axis which corresponds to 4 on horizontal axis and stabilizes and moves parallel to horizontal axis up to the point 8. The values of the graph mentioned above are approximate.

The input y to the nonlinearity is the sum of N weighted inputs x_i , as described by the relation

$$y = \sum_{i=1}^N w_i x_i = \vec{w} \cdot \vec{x},$$

$$y = \sum_{i=1}^N w_i x_i = \vec{w} \cdot \vec{x},$$

(11-122)

where the w_i are the weights applied to those inputs.

A single neuron can be trained to produce either a 1 or a 0 in response to a particular input vector $\mathbf{x} \rightarrow \vec{x}$ by adjusting the weight vector so that it is either co-directional with or orthogonal to that input vector, respectively. In the former case a large positive input to the sigmoid nonlinearity drives the output to a result very close to unity, and in the latter case a large negative input drives the output result very close to zero. In this way, by adjusting the weights, it is possible to “train” the neuron to recognize a particular input vector. Extending this idea, one finds if the neuron is to be presented with an entire set of vectors, each of which is to be classified into one of two possible sets, by training the neuron with examples of the two classes, it can be taught to separate the input vectors by means of a simple hyperplane in the N -dimensional space of the vectors. Classes of vectors that are separable with a hyperplane will then be distinguished by the neuron, while those that are not separable with a hyperplane cannot be distinguished.

Networks of Neurons

To obtain functionality that is more complex than that possible with a single neuron, collections of such elements are joined together to form a *neural network*. An example of such a network having four layers of interconnected neurons is shown in [Fig. 11.57](#). The layer of neurons on the far left can be thought of as containing input neurons. In an image-recognition problem, for example, each such neuron might receive a single input representing the value of one pixel of an image that is to be classified by the network. The layer on the right can be thought of as containing output neurons. Each such neuron represents one of the possible classes into which the image is to be classified. Ideally, when a single image is presented at the input of the network, with appropriate training, the network will cause a single output neuron to produce a value at or near unity and the rest to produce values at or near zero. The particular neuron that has unity output indicates the class of which the input image is a member. The middle layers of neurons are referred to as “hidden” layers. The number of hidden layers determines the complexity of the dividing surfaces in N -dimensional space that can separate input images into classes.

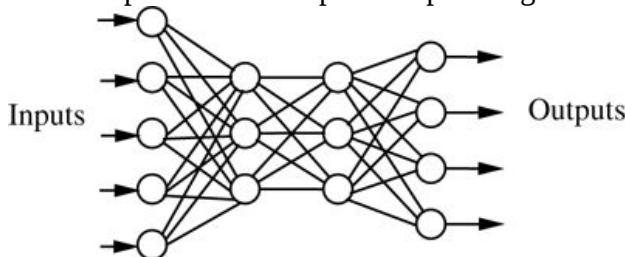


Figure 11.57

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 11.57 A four-layer neural network.

The illustration shows five circles arranged vertically on the extreme left. Five arrows from left, each one pointing toward each circle are labeled inputs. On the extreme right, four arrows are arranged vertically and an arrow from each circle points toward the right and the arrows are labeled outputs. Between the two vertical rows of circles, are circles arranged in two columns and three rows. All circles are interconnected with each other.

The neural network must be trained by presenting it with samples from the various classes of input images, and adjusting all of the weights according to some predetermined algorithm. A variety of training algorithms exist, all of which involve the minimization of an error metric. We

mention in particular the LMS algorithm [369] for single-layer networks and the backpropagation algorithm [303] for multilayer networks, but must refer the reader to the references for details.

Optical Neural Networks Based on Volume Holographic Weights

One popular implementation of neural networks using optics is based upon storage of weights in an erasable, thick holographic medium. Photorefractive crystals are most commonly used. [Figure 11.58](#) illustrates one manner in which a hologram can introduce a weighted interconnection. We assume that the input to the neural network consists of a spatial light modulator that generates a coherent amplitude distribution proportional to the input to be processed. The lens L_1 Fourier transforms the input, and thus each pixel in the spatial light modulator generates a plane wave with a unique k vector at the crystal. We assume that a collection of sinusoidal volume gratings has been written into the crystal; the exposure times or the strengths of the waves used in recording these gratings determine their diffraction efficiencies and therefore control the weights that they will apply to incident Bragg-aligned plane waves. By means of a second Fourier transforming lens, all plane waves traveling in a common direction are summed onto a single output pixel. There are many output pixels, each corresponding to a different direction of light diffracted from the crystal.

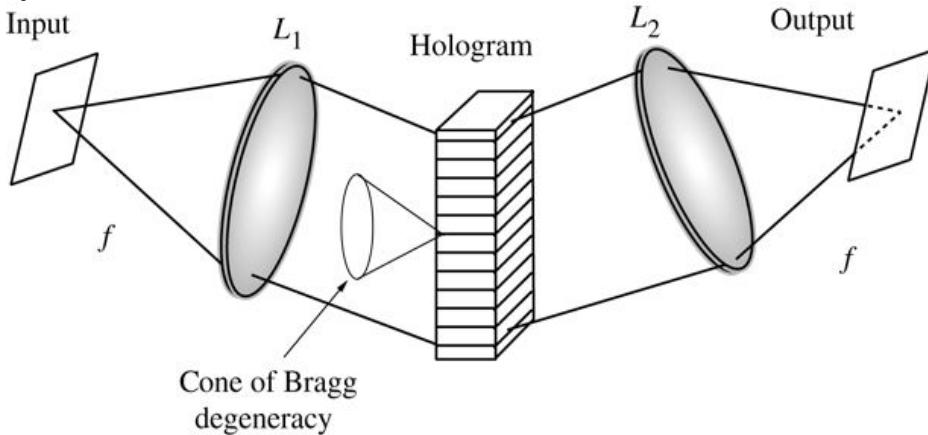


Figure 11.58

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 11.58 Illustration of a single weighted interconnection using a hologram. In practice, many such interconnections would be realized simultaneously.

The illustration shows a square shaped structure at the extreme left labeled input. To its right is Lens L_1 which is in slanting position. To the right of lens is a vertical rectangular structure labeled hologram with horizontal lines inside. To its right is Lens L_2 in slanting position. To the right of L_2 is another square shaped structure labeled output. Two lines from the center of Input point toward the upper and lower end of L_1 . A line from upper end of L_1 points toward the upper last line of hologram and a line from lower end of L_1 points toward the last horizontal line of hologram. A line from upper last horizontal line of hologram points toward the upper end of L_2 and a line from the lower last horizontal line of hologram points toward the lower end of L_2 . Two lines from the upper and lower end of L_2 point toward the center of object. A cone is shown with its pointed end resting at the center of hologram and is labeled cone of Bragg degeneracy. The distance between input and L_1 and between L_2 and output are each f .

Thus a multitude of volume gratings are written into the crystal, each grating representing one weight. A weighted sum of input pixel values then results at each output pixel. A training procedure can be implemented under computer control that changes the strengths of the volume gratings in accord with a chosen training algorithm.

The attraction of optics, and in particular volume holography, in this application comes from the very large number of gratings (or weights) that can be superimposed in a single crystal, coupled with the fact that large numbers of pixels (or neurons) can be realized with SLM technology. The thickness of the recording material is important if a multitude of different gratings are to be angularly multiplexed (using the Bragg effect) in the medium. The goal of achieving large numbers of weights is hindered by two phenomena. One, known as Bragg degeneracy, refers to the fact that the Bragg condition can be satisfied by an entire *cone* of angles, rather than just a single angle, and therefore there is the potential for significant crosstalk to exist between weighted interconnections. This problem can be combated by utilizing only a properly chosen subset of the possible gratings, such that light can pass from one input pixel to one output pixel by means of one and only one volume grating [291]. A second solution is to break the Bragg degeneracy by forcing a single path from an input pixel to an output pixel to diffract from more than one volume grating [272].

A second limitation arises from the fact that, for photorefractive crystals subjected to a sequence of exposures, the later exposures partially erase the early exposures. This limits the total number of gratings that can be superimposed; however, experiments have demonstrated storage of several thousands of exposures [258]. Actually, the tendency of the photorefractive medium to “forget” early exposures can be used to advantage in some learning processes.

We have only touched the subject of optical neural networks in the above discussion. Other approaches to utilizing optics for neural-like computation exist. We mention in particular the realization of Hopfield neural networks using the matrix-vector architecture of [Section 10.8.2](#) [290] and the use of competitive and cooperative phenomena in systems utilizing nonlinear optical elements [7]. For additional references, see the March 1993 issue of *Applied Optics*, which was devoted to the subject of optical neural networks.

11.13.6 Other Applications

Many other applications of holography exist, but space limitations prevent us from reviewing them all here. In this section we give brief mention of several areas that are particularly important and present some references for further study.

Holographic Optical Elements

Holography has found considerable application to the construction of waveshaping optical elements, which are referred to as *holographic optical elements* (HOEs). Such elements are one further example of diffractive optical elements. Again it is the light weight and compact volume associated with such elements that make them particularly attractive. Holographic optical elements have been used, for example, for optical scanning [214], [21], for heads-up displays in aircraft cockpits [71], and in many other applications.

The reader is referred to [Section 7.3](#) for further discussion of diffractive optical elements. References [64], [65], [66], [67], and [68] all contain examples of applications.

Holographic Display and Holographic Art

The striking character of three-dimensional holographic images has been the factor most responsible for interest in holography on the part of the nontechnical public. Holography has been applied to advertising, and a multitude of artists have taken up holography as a medium of choice. A Museum of Holography, containing many holographic works of art, was first established in New York City but is now located in the MIT Museum. Holographic jewelry can be found in many shops around the world.

Holograms for Security Applications

The application that brings holography into direct contact with the largest number of people is the use of holograms for prevention of counterfeiting and fraud. The ubiquitous embossed hologram on the credit card is the most common example in the United States, although in Europe its use has been extended even further. Holography is used in such applications to provide a deterrent to counterfeiting, since the presence of a hologram as an integral part of a credit card or a bank note makes the unauthorized duplication of that item considerably more difficult than would otherwise be the case.

To gain a better appreciation for the variety of applications of holography in the security field, the reader may wish to consult [\[105\]](#), which contains many papers on the subject.

Problems - Chapter 11

1. 11-1. A hologram is recorded using a spherical reference wave that is diverging from the point (x_r, y_r, z_r) , and the images from that hologram are played back with a reconstruction beam that is diverging from the point (x_p, y_p, z_p) . The wavelength used for both recording and reconstruction is λ_1 . The hologram is taken to be circular, with diameter D . See Fig. P11.1(a) below. It is claimed that the image of an arbitrary three-dimensional object obtained by this method is entirely equivalent to that obtained by a lens of the same diameter and at the same distance from the object, as shown in part (b) of the figure, and a prism (for simplicity, not shown), where again the wavelength is λ_1 . What are the two possible focal lengths for the lens that will produce equivalence?

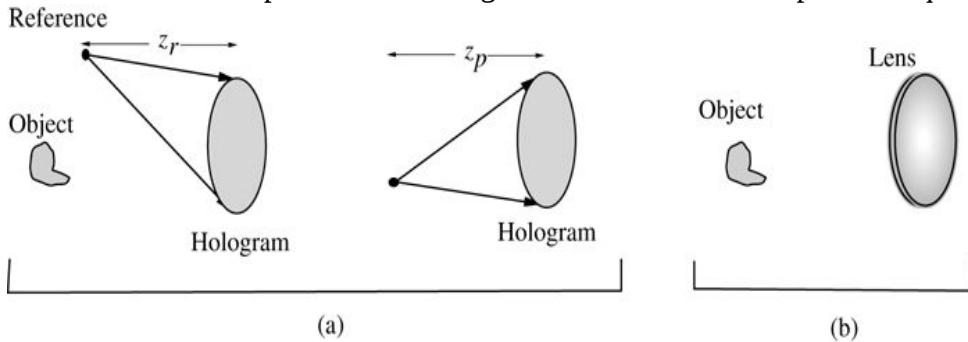


Figure P11.1

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure P11.1

A dark spot labeled reference is shown on top. To its right is an oval shaped structure labeled hologram. Two arrows from reference point toward the upper and lower ends of hologram. The distance between reference and hologram is z_f . To the right of hologram is another dark spot. A hologram is at a distance z_p from dark spot. Two lines from the dot point toward the upper and lower ends of hologram. Illustration b shows an irregular shaped structure labeled object and to its right is a lens.

2. 11-2. A hologram is recorded with light from an argon laser at 488 nm wavelength, and the images are reconstructed with light from a HeNe laser with wavelength 632.8 nm. There is no scaling of the hologram.

1. Assuming $z_p = \infty$, $z_r = \infty$, and $z_o = -10$ cm, what are the axial distances z_i of the twin images? What are the transverse and axial magnifications of the images?
2. Assuming $z_p = \infty$, $z_r = 2z_o$, $z_r = 2z_o$, and $z_o = -10$ cm, what are the axial distances and the transverse and axial magnifications of the twin images?

3. 11-3. A hologram is recorded, and its images reconstructed with the same wavelength λ . Assuming $z_o < 0$, show that when $z_p = z_r$ there results a virtual image with unity transverse magnification, whereas with $z_p = -z_r$ there results a real image with unity transverse magnification. What is the transverse magnification of the twin image in each case?
4. 11-4. The lensless Fourier transform geometry (see [Fig. 11.14](#)) is used to record a hologram of an object consisting of a square transparency of width L . The amplitude transmittance of the object is $t_A(x_o, y_o)$, and the distance of the object from the recording plane is $|z|$. The reconstruction wavelength is the same as the recording wavelength. The images are obtained by illuminating the hologram with a plane wave, followed by a positive lens of focal length f . For simplicity, both the object illumination and the reconstruction wave may be taken to have amplitude unity.

1. What is the transverse magnification M_t of the first-order images?
2. Show that the amplitude of the zero-order (i.e. on-axis) image term can be expressed as

$$U_f(u, v) = \lambda f \int_{-\infty}^{\infty} \int U'_o(x_o, y_o) U'^*(x_o + \frac{u}{M_t}, y_o + \frac{v}{M_t}) dx_o dy_o$$

$$U_f(u, v) = \frac{1}{\lambda f} \int_{-\infty}^{\infty} \int U'_o(x_o, y_o) U'^*(x_o + \frac{u}{M_t}, y_o + \frac{v}{M_t}) dx_o dy_o$$

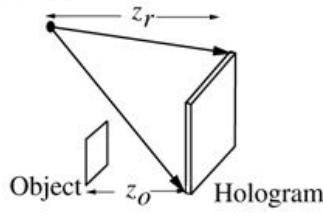
(plus a central diffraction-limited spot), where

$$U'_o(x_o, y_o) = t_A(x_o, y_o) e^{j\pi\lambda|z|(x_o^2 + y_o^2)}.$$

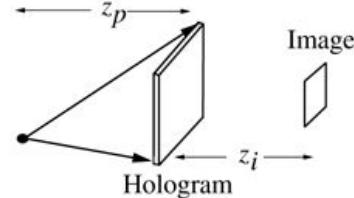
3. How far from the center of the object transparency should the reference point source be placed in order to assure no overlap of the zero-order light with the first-order images?
5. 11-5. We wish to make a holographic display that will project a real image of a planar transparency object. The recording and reconstruction geometries are shown in [Fig. P11.5](#). The reference and object are constrained to lie to the left of the hologram during recording. The reconstruction source must lie to the left of the hologram, and the projected image must lie to the right of the hologram. The hologram is not turned around or changed in size between recording and reconstruction. The recording wavelength is 632.8 nm, the reconstruction wavelength is 488 nm, the object transparency size is 2×2 cm, the desired image size is 4×4 cm, the axial distance from the hologram to the image must

be 1 m, and the axial distance of the reconstruction source to the hologram is constrained to

Reference



(a)



(b)

Figure P11.5

Goodman, *Introduction to Fourier Optics*, 4e,

© 2017 W. H. Freeman and Company

be 0.5 m.

Figure P11.5

A dark spot labeled reference is at the top. To its right is another square shaped structure labeled hologram. Two arrows from reference point toward the upper right and lower left corners of hologram. The distance between reference and hologram is z_r and the distance between object and hologram is z_o . Illustration b shows a dark spot at the left bottom. A hologram is at a distance z_p from dark spot. Two lines from the dot point toward the upper right and lower left corners of hologram. To the right of hologram is another square shaped structure labeled image and the distance between hologram and image is z_i .

1. Subject to the above constraints, specify all possible axial object and reference distances z_o and z_r that will together yield the desired image.
2. Repeat part (a), but with the hologram rotated 180° left to right (i.e. back and front interchanged) between the recording and reconstruction steps.
6. 11-6. It is proposed to record an X-ray hologram using coherent radiation of wavelength 0.1 nm and to reconstruct images optically using light of wavelength 600 nm . The object is a square transparency with a pattern of absorption at the X-ray wavelength. The lensless Fourier transform recording geometry is chosen. The width of the object is $100\text{ }\mu\text{m}$, and the minimum distance between the object and the reference is to be $200\text{ }\mu\text{m}$ to ensure that the twin images will be separated from the “on-axis” interference. The X-ray film is placed 2 cm from the object.
 1. What is the maximum spatial frequency (cycles/mm) in the interference pattern falling on the film?
 2. Assume that the film has sufficient resolution to record all of the incident intensity variations. It is proposed to reconstruct the images in the usual manner, i.e. by looking in the rear focal plane of a Fourier transforming lens. Why will this experiment fail?
7. 11-7. A thick unslanted transmission phase grating is to be produced by bleaching the recording that results from interfering two plane waves in a photographic emulsion. The wavelength of the exposing radiation in air is 488 nm and the angle between the two interfering beams, also in air, is 60° . The thickness of the emulsion is $15\text{ }\mu\text{m}$. The average refractive index of the emulsion, both before exposure and after bleaching, is 1.52 (the index of gelatin). The same wavelength is used for reconstruction as for recording.

- What are the wavelength and the angle between the two beams *inside* the emulsion during recording?. How does the period of the grating predicted by the angle and wavelength outside the emulsion compare with the period predicted using the wavelength and angle inside the emulsion?
- Assuming Bragg matched conditions, what peak refractive index modulation $n_1^{n_1}$ is required in order to reach the first 100% peak of the diffraction efficiency curve for a thick transmission phase grating?
- Assuming operation at this same first maximum of diffraction efficiency, and assuming no error $\Delta\theta \Delta\theta$ in the illumination angle, what wavelength error $\Delta\lambda \Delta\lambda$ (external to the emulsion) will result in the diffraction efficiency dropping to 50%?
- Again assuming operation at the first maximum of the diffraction efficiency, and assuming no error in the reconstruction wavelength, what angular error $\Delta\theta \Delta\theta$ (external to the emulsion) will result in a reduction of the diffraction efficiency to 50%?
- 11-8. A holographic plate of thickness $15 \mu m$ records a hologram by interference of two plane waves with equal but opposite angles to the emulsion normal. The wavelength for both recording and reconstruction is 633 nm, and the refractive index of the emulsion is 1.52 before and after development. For what angle (in air) between the two interfering beams will the thickness parameter $Q Q$ of (9-44) have value $2\pi 2\pi$?
- 11-9. Consider a thick transmission, unslanted sinusoidal absorption grating. Let the absorption modulation have its maximum possible value. Assume that the interfering plane waves are separated in angle by $60^\circ 60^\circ$. Under Bragg matched conditions, what average density $D D$ of the transparency yields the maximum possible diffraction efficiency of 3.7%?
- 11-10. Using (11-66), show that, in the absence of wavelength mismatch, the angular selectivity of a volume grating is maximized when the object and reference waves are separated by an angle of $90^\circ 90^\circ$. Hint: remember that $K K$ depends on $\theta \theta$.
- 11-11. For a certain binary detour-phase hologram, square cells (size $L \times L L \times L$) are allocated to each Fourier coefficient, and the amplitudes $|a_{pq}| |a_{pq}|$ of those coefficients are represented by opening a rectangular subcell within each cell. The width $w_X w_X$ for all transparent subcells is constrained to be 1/10th of the cell width to satisfy the approximations used. The width $w_Y w_Y$ can range from 0 to the full size of a cell, depending on the amplitude to be represented. The hologram is uniformly illuminated by a normally incident plane wave, and no light is lost by the Fourier transforming lens that follows the hologram. For the purposes of this problem, the object is taken to be a point source located at the center of the object space, yielding Fourier coefficients that are all the same constant, say of value $a a$, which we shall take to be a number somewhere between 0 and 1. When the Fourier amplitude is to be $a a$, the vertical height of all subcells is set to $w_Y = aL w_Y = aL$.
 - For a given value of $a a$, find the coefficients of the two-dimensional Fourier series representation of the amplitude transmittance of the binary hologram.
 - Calculate the fraction of the total light intensity incident on the hologram that ends up in the zero-frequency spot on the axis in the image plane.

3. Calculate the fraction of total incident light that is blocked by the opaque portions of the hologram.
 4. Find the diffraction efficiency for both of the two first-order images.
12. 11-12. A certain film has a nonlinear t_A versus E curve which, over its region of operation, may be described by

$$t_A = tb + \beta E_1^3,$$

$$t_A = t_b + \beta E_1^3,$$

where E_1 represents the variations of exposure about the reference exposure.

1. Assuming a reference wave $A \exp(-j2\pi\alpha x)$ and an object wave

$$a(x,y) \exp -j\phi(x,y)$$

$$a(x, y) \exp [-j\phi(x, y)]$$

at the film, find an expression for that portion of the transmitted field that generates the twin first-order images.

2. To what does this expression reduce if $A \gg |a|$?
3. How do the amplitude and phase modulations obtained in the previous parts of the problem compare with the ideal amplitude and phase modulations present when the film has a linear t_A versus E curve?

12 Fourier Optics in Optical Communications

12.1 Introduction

In this final chapter, we briefly discuss some of the applications of Fourier optics to devices and techniques related to modern optical communications. In some cases, the devices or techniques are inspired by topics discussed earlier in this book, while in others, adoption of a Fourier optics point of view facilitates understanding of how the device or technique functions. In this chapter we merely scratch the surface of the optical communications field, to which many other books have been devoted (see, for example, [187] and [188]).

The extremely high data rates achieved in modern optical communications systems are realized by either or both of two different techniques: achievement of extremely high speeds on a single signal channel (e.g. time-domain multiplexing of many slow speed channels into a single high-speed channel), or simultaneous transmission of many orthogonal channels onto a single common medium (e.g. by wavelength-division multiplexing in fibers).

Both free-space and waveguide-based devices and techniques are relevant to our topic. It is important to realize at the start that the Fourier techniques emphasized earlier in this book are appropriate primarily in free-space geometries, and less so to waveguide devices. The reason is quite fundamental—plane waves of infinite extent traveling in different directions are natural “modes” of free space. These modes are in fact the Fourier components of the propagating signals. However, in bounded dielectric media, such as integrated-optic waveguides and optical fibers, the natural modes are not plane wave components, but rather are uniquely determined by the cross-sectional shapes and the refractive index profiles of the waveguides themselves, as well as the wavelength of light in the guide. In addition, instead of an infinite set of orthogonal modes as exists for free space, waveguide devices support a finite set of orthogonal modes. Nonetheless, in some cases the reasoning used for analyzing free-space geometries can provide first-order insight into the operation of waveguide devices. In the material that follows, the use of accurate mode decompositions will not be necessary, although in some cases more accurate results will require such decompositions.

In several sections of this chapter we will be dealing with waveguiding structures having refractive indices greater than unity. For this reason, we will distinguish between free-space

wavelengths and wavelengths in the waveguides by using λ for the former and $\tilde{\lambda}$ for the latter, reminding the reader about this distinction from time to time.

12.2 Fiber Bragg Gratings

In 1978, [Hill and coworkers \[170\]](#) at the Communications Research Center of Canada obtained some surprising experimental results while studying the nonlinear properties of a special fiber using blue light. They hypothesized and later proved that the phenomenon they were observing was caused by the writing of a relatively permanent, photoinduced index grating in the glass fiber itself. This was the birth of a new technology now known as Fiber Bragg Gratings (FBGs). In this section we cover some of the properties and applications of such gratings. For an excellent review of this field, see [\[332\]](#).

Much additional work by many individuals took place before FBGs became a commercial reality. Some of this work included the use of UV lasers for writing the gratings, the photosensitization of the glass by diffusion of molecular hydrogen into a standard fiber before exposure, and the use of phase masks for creating the proper interfering beams during exposure. The reader is referred to the article cited above for a more detailed recounting of the history of this technology. It is now possible to induce essentially permanent refractive index changes of magnitude 10^{-4} to 10^{-2} in a glass fiber using these methods.

A FBG is basically a thick hologram recorded down the length of a section of glass fiber. The chief advantage of a FBG arises from the fact that the grating resides in a fiber, which can be spliced to ordinary fiber, and therefore provides a compact and low-loss method for introducing in-fiber devices such as narrowband filters, dispersion compensators, and other types of filter structures.

12.2.1 Introduction to Optical Fibers

We begin by giving a brief description of a glass fiber (for additional background, see [\[305\]](#), [Chapter 8](#)), a short section of which is shown in [Fig. 12.1](#). A cylindrical glass cladding of index n_2 and radius b surrounds a cylindrical glass core of index n_1 and radius a ($a < b$ and $n_2 > n_1$). In general, such a structure supports multiple propagating modes, present predominantly in the core, but with tails extending into the cladding. The lowest order mode, which is the only propagating mode in a single-mode fiber, has a shape that resembles a Gaussian distribution and is generally referred to as the LP_{01} mode. For a single mode fiber, the cladding diameter is generally much larger than the core diameter.

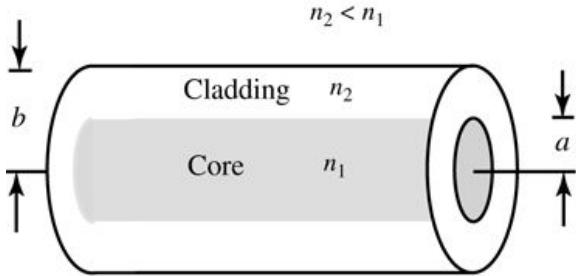


Figure 12.1

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 12.1 A short section of fiber.

The illustration shows a cylinder consisting of an inner cylindrical core of index n_1 wrapped in a uniform layer of cladding of index n_2 , where n_2 is less than n_1 . The radius of the cylindrical core is a , and the radius of the fiber as a whole is b .

The most important property of properly designed optical fiber is its extremely low attenuation for optical signals. At the wavelength of lowest loss, 1550 nm, single mode fibers introduce as little as 0.16 dB loss per km.

When viewed from air, the angular spread of the light emerging from a fiber (as well as the angular spread of light that can be efficiently coupled into a fiber) is described by the numerical aperture, which can be shown to be

$$NA_{air} = \sin\theta_a = (n_1^2 - n_2^2)^{1/2} \approx n_1(2\Delta)^{1/2},$$

$$NA_{air} = \sin\theta_a = (n_1^2 - n_2^2)^{1/2} \approx n_1(2\Delta)^{1/2},$$

(12-1)

where θ_a is the maximum half angle from the fiber axis, and $\Delta = (n_1 - n_2)/n_1$

$\Delta = (n_1 - n_2)/n_1$ is the fractional difference of refractive index between the core and the cladding. Within the core, the corresponding expression for numerical aperture is

$$NA_{core} = n_1(2\Delta)^{1/2},$$

$$NA_{core} = \sqrt{\frac{n_1^2 - n_2^2}{n_1^2}} \approx (2\Delta)^{1/2},$$

(12-2)

as can easily be derived with the help of Snell's Law. Note that refractive index n_1 typically lies between 1.44 and 1.46, and Δ typically is between 0.001 and 0.02.

Unfortunately, different wavelengths of light propagate with slightly different speeds in a single-mode fiber, a consequence of both the material dispersion of glass and waveguide dispersion. In most cases material dispersion is the dominant effect, but if dispersion is to be perfectly compensated, both effects must be taken into account. See [305], p. 351. Since a short pulse of light has a spectrum containing a range of wavelengths, pulse broadening occurs by an

amount that depends on the particular type of single-mode fiber used, the central wavelength of the light, and the length of the fiber. To describe this effect in more detail, consider the propagation of a broadband signal in a single-mode fiber. Neglecting the spatial profile of the signal in a fiber, the complex representation of the signal $u(t)$ can be written

$$u(t) = U(t) \exp[-j(\omega t - \beta(\omega)L)],$$

(12-3)

where $U(t)$ is a complex time-varying phasor representing the amplitude and phase modulation of the launched signal, $\omega = 2\pi\nu$ is the angular optical frequency, and L is the length of the fiber over which the signal propagates. Here $\beta(\omega)$ is the propagation constant, and it depends on frequency, due both to the dependence of the refractive index of glass on frequency and the dependence of the mode profile on frequency.¹

Since the spectral width of the signal is generally much less than the center frequency, it is helpful to expand $\beta(\omega)$ in a Taylor series about the center frequency ω_0 . Keeping only four terms in the expansion, we have

$$\begin{aligned} \beta(\omega) &= \beta(\omega_0) + (\omega - \omega_0) \frac{\partial \beta}{\partial \omega} + \frac{1}{2}(\omega - \omega_0)^2 \frac{\partial^2 \beta}{\partial \omega^2} + \frac{1}{6}(\omega - \omega_0)^3 \frac{\partial^3 \beta}{\partial \omega^3} \\ \beta(\omega) &= \beta(\omega_0) + (\omega - \omega_0) \frac{\partial \beta}{\partial \omega} + \frac{1}{2}(\omega - \omega_0)^2 \frac{\partial^2 \beta}{\partial \omega^2} + \frac{1}{6}(\omega - \omega_0)^3 \frac{\partial^3 \beta}{\partial \omega^3} \end{aligned}$$

(12-4)

where the derivatives are all evaluated at frequency ω_0 . The first term in this series leads to a phase shift that is constant over frequency and can be ignored. The second term, containing a linear phase shift with frequency, leads to a simple delay of the signal, with no internal change of time-structure. This term is useful in defining the *group velocity*, or the speed at which a pulse

propagates down the fiber. The time delay of the pulse is $\tau = L \frac{\partial \beta}{\partial \omega}$ and the group

velocity is therefore $v_g = L / \tau = \omega_0 / \frac{\partial \omega}{\partial \beta}$, evaluated at the center frequency ω_0 . The third term introduces a quadratic-phase distortion across the frequency spectrum of the signal and is generally the dominant dispersion term. The fourth term corresponds to the slope of the dispersion curve as a function of ω and can make an important contribution in some applications.

The time-spreading of a pulse, $\Delta\tau$, caused by the quadratic-phase term depends on the length L of fiber traveled and the spectral width $\Delta\omega$ of the signal through

$$\Delta\tau = \frac{\partial^2 \beta}{\partial \omega^2} L \Delta\omega.$$

$$\Delta\tau = \frac{\partial^2 \beta}{\partial \omega^2} L \Delta\omega.$$

The group velocity dispersion coefficient D (measured in picoseconds per kilometer-nanometer) is defined as time spread per unit length due to wavelength variation and is given by

$$D = -2\pi c \lambda^2 \partial^2 \beta / \partial \omega^2$$

$$D = -\frac{2\pi c}{\lambda^2} \frac{\partial^2 \beta}{\partial \omega^2}$$

(12-5)

where λ is the wavelength in air, from which it can be seen that time spread can be expressed as²

$$\Delta\tau = DL\Delta\lambda.$$

$$\Delta\tau = |D|L \Delta\lambda.$$

(12-6)

Dispersion can be combated in optical fiber communications through a variety of techniques. Most common is the use of dispersion-shifted fiber, which is fiber that, through geometry and index profile changes, has the zero-dispersion point shifted from its usual wavelength near 1300 nm to wavelength 1550 nm where fiber attenuation is minimum. Another approach is to use dispersion-compensating fiber, which, again through design, has had the sign of its dispersion changed, and therefore introduces dispersion opposite to that of normal fiber. When normal fiber and dispersion compensating fiber are spliced together, dispersion is reduced. Finally, it is possible to introduce separate devices in the fiber path that serve to compensate dispersion. One approach that uses FBGs is discussed in a later sub-section.

12.2.2 Recording Gratings in Optical Fibers

A phase grating can be recorded in a glass fiber by either of two methods, one directly interferometric and the other using a phase grating to generate two interfering beams. [Figure 12.2\(a\)](#) illustrates one of several possible interferometric methods. A section of fiber is shown, illuminated from the side by two mutually coherent beams, generated by splitting the light from a UV laser. The two beams suffer approximately equal path length delays (thus preserving their mutual coherence), and interfere in a region surrounding a portion of the fiber. In the illustration shown, the fringes run perpendicular to the long axis of the fiber. Since the UV wavelength of the recording laser is different from the near-IR wavelength where the filter will be used in a communications system, the angle of the interfering beams must be adjusted to achieve the fringe spacing that will be appropriate for the IR wavelength.

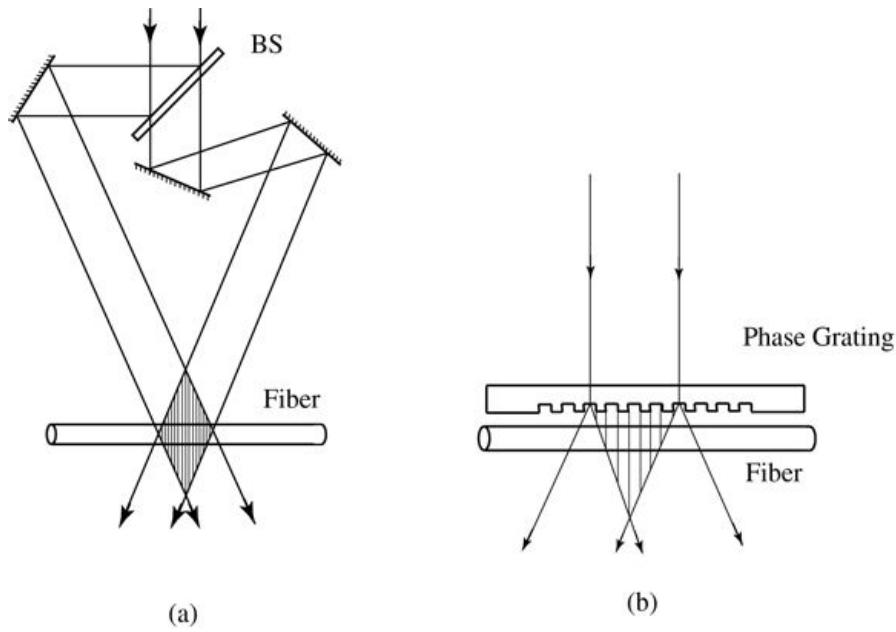


Figure 12.2

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 12.2 Two methods for recording an FBG: (a) interferometric method, and (b) phase grating method.

Illustration a shows a pair of downward pointing vertical rays falling on a UV laser and passing through it to two comb like interfaces, one to the left and one below that transmits to another to the right. Thereafter the beams, a pair from the left and another from the right, converge at the center of a horizontal fiber where they form a diamond shaped intersection. Illustration b shows a rectangular phase grating located horizontally just above and parallel to a cylindrical piece of fiber. The lower edge of the grating is a series of grooves. Two downward rays pass through the grating and on reaching a groove, each splits in two, thus appearing like an inverted capital letter Y. The split rays enter and cross the fiber beneath. From the corners of the grooves that lie between the two rays, vertical lines are drawn extending up to where the split rays cross their paths.

A second method for generating an FBG is illustrated in Fig. 12.2(b). For this method, a “master” phase grating is made, usually by etching grooves in a plate of glass. Typically the phase grating will have a close approximation to a square-wave profile, with an optical path-length difference of π radians between groove peaks and troughs. Such a grating has no zero order and no even orders, the transmitted light being dominated by the two first orders, which contain more than 80% of the transmitted light (c.f. Prob. 4-16). The two first orders interfere in the fiber, creating a fringe pattern with a period that is half the period of the master grating. Advantages of the phase grating approach are that it minimizes recording laser coherence requirements, the fringe period created is not affected by small changes of the laser wavelength, and the method appears more amenable to mass manufacturing of FBGs than the interferometric method, although both have been used in practice. The disadvantage of the phase grating method is that once the master grating has been generated, there is no easy way to change the period of the FBG.

12.2.3 Effects of an FBG on Light Propagating in the Fiber

A common method used to analyze FBGs is an extended version of the coupled mode theory discussed in Section 11.7.5. For a detailed treatment of this extended theory, see [332], Section

IIC.

Here we will not be as thorough and complete as the cited analysis, but instead will try to use results already derived in [Chapter 11](#) to obtain a first-order understanding of the properties of gratings in fibers. We will assume that the refractive index perturbations are weak, and we shall consider primarily the lowest-order mode propagating in a single-mode fiber, the LP_{01} mode. The angular spread of the mode is determined by the numerical aperture of the light in the core, as given by [\(12-2\)](#), for which a typical value would be $NA_{core} \approx 0.15$, corresponding to a relatively small angular spread of the light in the fiber.

Phase Reflection Gratings

Consider first a uniform sinusoidal phase reflection grating recorded in a fiber, with the grating lines perpendicular to the axis of the core of the fiber. Referring back to the discussion surrounding [Fig. 11.28](#), when the grating lines are perpendicular to the direction of propagation of the light, wavelength sensitivity is maximized and angular sensitivity is less severe. Hence the small spread of angles, implied by the low numerical aperture of the light in the core, can be ignored, and the results derived in [Chapter 11](#) for the response of such a grating to a plane wave of infinite extent can be used as a reasonable approximation.

With reference to [Fig. 11.29](#), for the case of interest here, the following parameter values hold: $\theta=0$, and $\psi=\pi/2$. In addition, the angular mismatch of the illumination, $\Delta\theta$, is zero. It follows that the parameters Φ and χ have the values

$$\Phi = \pi \delta n \ell \tilde{\lambda}^{-1}, \quad \chi = -\pi \ell \Delta \tilde{\lambda} / 2\Lambda,$$

$$\begin{aligned}\Phi &= \frac{\pi \delta n \ell}{\tilde{\lambda}} \\ \chi &= -\frac{\pi \ell \Delta \tilde{\lambda}}{2\Lambda},\end{aligned}$$

(12-7)

where ℓ is the length of the uniform fiber grating, δn is the peak change of core refractive index caused by the grating, $\tilde{\lambda}$ is the wavelength of the light in the fiber core, $\Delta \tilde{\lambda}$ is the difference between the Bragg-matched wavelength $\tilde{\lambda}_B = 2\Lambda$ and the actual wavelength, and Λ is again the period of the grating.

The maximum diffraction efficiency occurs when the wavelength is Bragg matched, i.e. when $\tilde{\lambda}_B = 2\Lambda$, in which case the diffraction efficiency of the grating is given by (c.f. [\(11-79\)](#))

$$\eta = \tanh 2\Phi = \tanh 2\pi \delta n \ell / 2\Lambda.$$

$$\eta = \tanh^2 \Phi = \tanh^2 \frac{\pi \delta n \ell}{2\Lambda}.$$

(12-8)

The diffraction efficiency of the gratings when the wavelength is de-tuned from the Bragg wavelength is given by (c.f. (11-78))

$$\eta = 1 + \frac{1 - \chi^2 / \Phi^2}{\sinh^2 \left(\Phi \sqrt{1 - \frac{\chi^2}{\Phi^2}} \right)}.$$

$$\eta = \left[1 + \frac{1 - \frac{\chi^2}{\Phi^2}}{\sinh^2 \left(\Phi \sqrt{1 - \frac{\chi^2}{\Phi^2}} \right)} \right]^{-1}.$$

(12-9)

The crux of the problem now is to find expressions for χ^2 / Φ^2 and η when the wavelength is de-tuned from the Bragg wavelength $\tilde{\lambda}_B = 2\Lambda$ by $\Delta\lambda / \tilde{\lambda}_B$. This task is simplified by the fact that the range of wavelengths within which the filter will have effect is small compared with the Bragg-matched wavelength, i.e. $\Delta\lambda / \tilde{\lambda}_B \ll 1$. Using this fact, as well as a number of other manipulations (see Prob. 12-1), it is possible to show that the expression for diffraction efficiency can be reduced to

$$\eta = 1 + \frac{1 - \frac{4x^2}{\delta n^2}}{\operatorname{csch}^2 \left(\frac{\pi \delta n N}{2} \sqrt{1 - \frac{4x^2}{\delta n^2}} \right)},$$

$$\eta = \left[1 + \left(1 - \frac{4x^2}{\delta n^2} \right) \operatorname{csch}^2 \left(\frac{\pi \delta n N}{2} \sqrt{1 - \frac{4x^2}{\delta n^2}} \right) \right]^{-1},$$

(12-10)

where csch is a hyperbolic cosecant, $N = \ell / \Lambda$ is the number of periods in the grating, and $x = \Delta\lambda / \tilde{\lambda}_B = \Delta\lambda / \lambda_B$ is the fractional wavelength de tuning from the Bragg wavelength. Note that when $x=0$, i.e. the wavelength is the Bragg wavelength, the diffraction efficiency reduces to the equivalent of (12-8),

$$\eta = \tanh^2 \left(\frac{\pi \delta n N}{2} \right).$$

$$\eta = \tanh^2 \left(\frac{\pi \delta n N}{2} \right).$$

(12-11)

Referring to Fig. 11.34, we see that the quantity $\pi \delta n N / 2$ drives the diffraction efficiency towards unity as it increases; in fact, when this quantity grows to value 3, the diffraction efficiency is 99%. At this point the diffraction efficiency can not be increased significantly by increasing the length of the grating, since the incident power has, for all practical purposes, already been transferred to the backward-propagating wave. The *effective* number of grating lines N_0 and the *effective* length ℓ_0 at this point are given by

$$N_0 \approx 6\pi\delta n\ell \approx 6\Lambda\pi\delta n.$$

$$N_0 \approx \frac{6}{\pi \delta n}$$

$$\ell_0 \approx \frac{6\Lambda}{\pi \delta n}.$$

(12-12)

This kind of behavior is also true of the more general expression of [Eq. \(12-10\)](#), with the added complication of the quantity $1 - 4x^2 / \delta n^2$. This factor, appearing in two places, drives the diffraction efficiency down due to the de tuning from the Bragg wavelength. In fact, when $4x^2 / \delta n^2 > 1$, or when $x > \delta n / 2$, this quantity becomes imaginary, and an oscillatory drop of diffraction efficiency appears. [Figure 12.3](#) shows a three-dimensional plot of diffraction efficiency η versus both grating length ℓ (in meters) and fractional de tuning x for the following specific parameters: $\lambda_B = 1550$ nm, $n_1 = 1.45$, and $\delta n = 10^{-4}$.

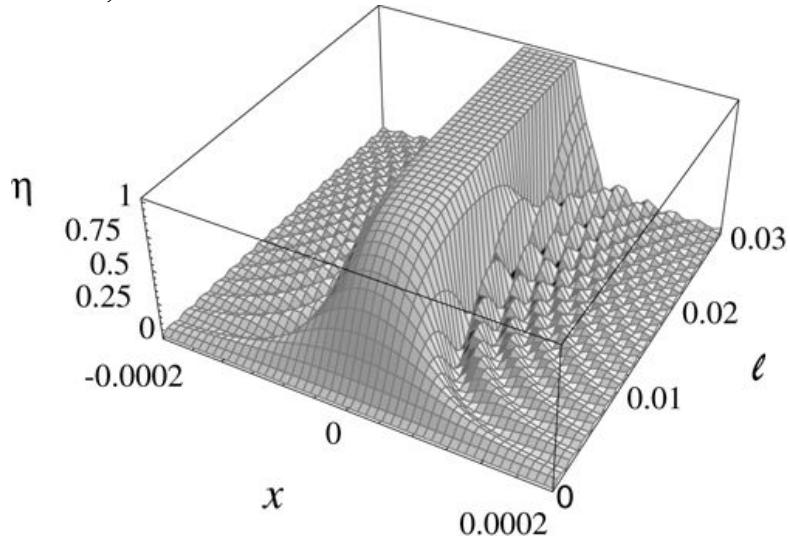


Figure 12.3

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 12.3 Plot of diffraction efficiency η versus grating length ℓ and fractional wavelength de-tuning x for $\lambda_B = 1550$ nm, $n_1 = 1.45$, and $\delta n = 10^{-4}$.

The 3 dimensional graph shows horizontal axis x , marked from 0.0002 to minus 0.0002, extending from right to left and the other horizontal axis ℓ , marked from 0 to 0.03, extending from left to right. The vertical axis η is marked from 0 to 1. The 3D graph resembles a symmetric mountain with steep cliffs on either side. The flat peak of the mountain is perpendicular to the 0 mark on the x axis, turning into a slope as it approaches the axis.

To help understand the changing character of the grating response with length, Fig. 12.4 shows four plots for four different lengths of a grating with $\delta n = 10^{-4}$ and (free-space) Bragg wavelength 1550 nm. In part (a), the length is too short to achieve significant diffraction efficiency, and the grating response curve is broad. In part (b), the grating is only long enough to achieve 80% diffraction efficiency, but the response curve has narrowed. In part (c), the grating is long enough to achieve near 100 % diffraction efficiency (i.e. just barely deplete the forward-propagating wave); the length is approximately equal to the maximum effective length. In part (d), the grating length is far longer than the effective length, and as a result the diffraction efficiency curve has taken on a flat top; a broader range of wavelengths is able to fully deplete the forward-propagating wave. Note that the response curve does not get narrower when the length exceeds the effective length. The width of the curve is determined primarily by the number N_0 of grating planes in the effective length. Equation (12-12) implies that the effective length of this grating is between 1 and 2 cm.

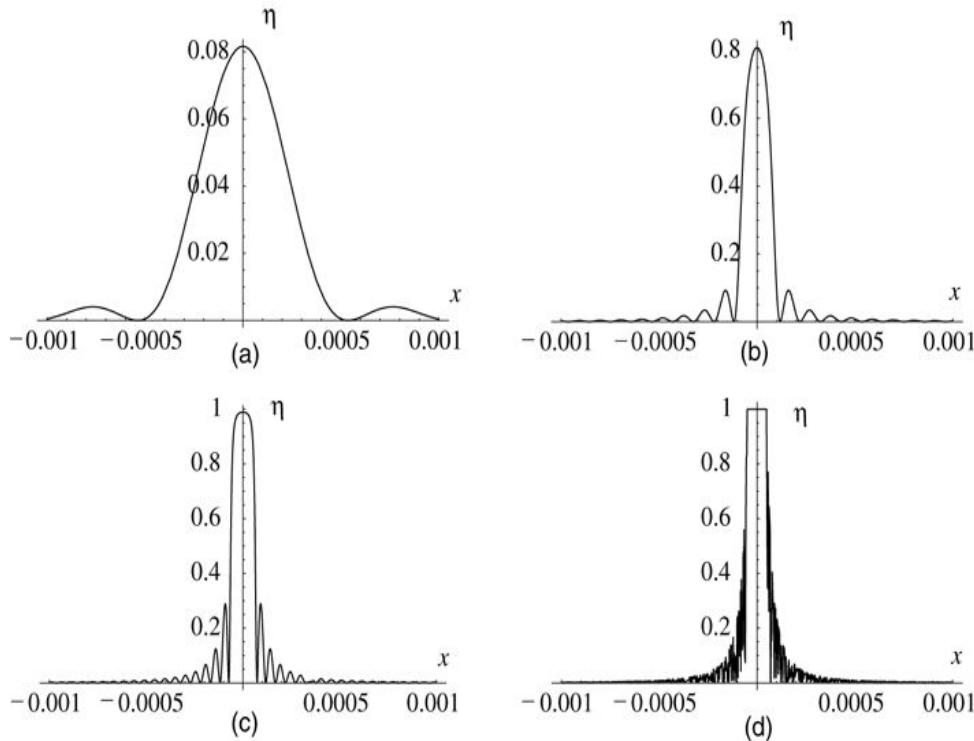


Figure 12.4

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 12.4 Response of a reflection grating with $\delta n = 10^{-4}$ and (free-space) Bragg wavelength 1550 nm for (a) length 1 mm, (b) length 5 mm, (c) length 1 cm, (d) length 1 m.

In graph a, horizontal axis x is marked from +0.001 to minus 0.001, and vertical axis η is marked from 0 to 0.08. The curve rises slightly in the leftward direction from the +0.001 mark on the horizontal axis and slopes down to reach the +0.0005 mark on the horizontal axis, where it rises steeply to reach the 0.08 mark on the vertical axis. The graph thus far is reflected onto the other side across the vertical axis. Graph b is the same as graph a but here the curve rises and falls slightly several times on the horizontal axis between + 0.001 and to the right of +0.0001, where it rises very steeply to reach the 0.8 mark on the vertical axis. The graph thus far is reflected onto the

other side across the vertical axis. Graph c is the same as graph b but here the curve rises and falls slightly more number of times on the horizontal axis between +0.001 and to the left of +0.0001, where it rises very steeply, almost perpendicularly, to reach the 1 mark on the vertical axis. The graph thus far is reflected onto the other side across the vertical axis. Graph d is the same as graph c but here the curve rises and falls slightly many, many more times, almost vertically, on the horizontal axis between +0.001 and to the left of +0.0001, where it rises perpendicularly to reach the level of the 1 mark on the vertical axis. The curve turns left horizontally and reaches the vertical axis at the +1 mark. The graph thus far is reflected onto the other side across the vertical axis.

We shall discuss the use of FBGs as narrowband filters in the sub-section to follow. Note that the response curves plotted against $x \propto$, the fractional wavelength de tuning, can also be viewed as plots versus fractional frequency de tuning, since $\Delta\lambda/\lambda_B = \Delta\nu/\nu_B$. Note also that the sidelobes of the grating response versus de tuning can be reduced by proper apodization of the grating strength along the fiber.

12.2.4 Applications of FBGs

While there are a great many applications of FBGs in the optical communications field, here we will discuss two applications of the reflection FBGs described above. One area of application is realization of narrowband filters for use in add/drop multiplexers. The second is to realization of filters for wavelength dispersion compensation.

Narrowband Filters for Add/Drop Multiplexers

A common method for achieving very high optical data rates is through the use of dense wavelength-division multiplexing (DWDM). Many different data streams are multiplexed onto a single fiber by assigning a unique wavelength to each stream. Typically the wavelengths are arranged in a dense comb, with adjacent channels separated by 100 GHz, 50 GHZ, or even 25 GHz. As many as several hundred channels can be multiplexed onto a single fiber in practice.

A key device or subsystem in such a system is an add/drop multiplexer (ADM), which is capable of extracting a single wavelength from the fiber without disturbing other wavelengths, or adding a single wavelength to the fiber without disturbing others. Many different architectures for realizing ADMs have been discussed in the literature. Here we focus only on one that uses an FBG.

[Figure 12.5](#) shows a typical geometry for an ADM. The only unfamiliar devices in this figure are the circulators. An optical circulator is a non-reciprocal device that allows light to propagate from an input port to an output port in only one direction (forward propagation). Light traveling in the opposite direction is sent to a separate port where only the backward propagating light appears. The isolation between forward and backward propagating signals is generally very high in such devices (~ 50 dB). Light entering the first circulator passes through to the FBG, which has been designed to be a narrow-band reflection filter, reflecting only wavelength λ_2 and passing all other wavelengths to the second circulator. Meanwhile, wavelength λ_2 passes in the reverse direction to the “drop” port, where the signal on this particular wavelength channel can be detected. Returning to the second circulator, the wavelength grid, now missing λ_2 passes through to the output port undisturbed. A new wavelength channel λ_2' is applied to the second

input port of the last circulator and travels back to the FBG, where it is reflected, then passes through the second circulator, filling the empty channel space where λ_2 is missing. Thus it is possible with such an architecture to extract a particular wavelength and to add a new wavelength.

If two FBGs are cascaded in the middle, the first tuned to λ_2 and the second to λ_2' , then wavelength λ_2 and λ_2' need not be the same.

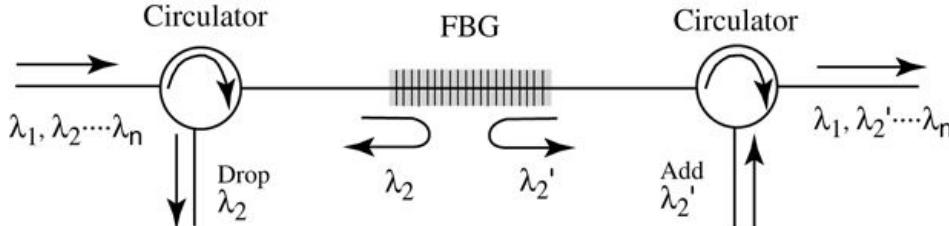


Figure 12.5

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 12.5 Typical structure of an FBG add/drop multiplexer.

The illustration shows a horizontal line connecting a circulator, an FBG, and a circulator, in that order. A circle with a clockwise arrow represents the circulators while a series of equal vertical lines represents the FBG. A rightward arrow at the left extreme points at the first circulator. Below the arrow a line reads, "Labda 1, lambda 2, and so on till lambda n." From the first circulator, a line drops vertically; a downward vertical arrow reads, "Drop lambda 2." At the bottom left corner of the FBG is a U turned arrow pointing to the left. It is labeled lambda 2. At the bottom right corner of the FBG is a U turned arrow pointing to the right. It is labeled lambda 2 dash. From the second circulator, a line drops vertically; an upward vertical arrow reads, "Add lambda 2 dash." A rightward arrow at the right extreme points away from the second circulator. Below the arrow a line reads, "Lambda 1, lambda 2 dash, and so on till lambda n."

Given the extremely close spacing of wavelengths in typical DWDM systems, it is important that the FBGs be designed to be very narrowband. To obtain a very narrowband filter, it is essential that δn be small, so that the effective number of planes in the grating can be very large. The optical signal should propagate as far as possible before all the light has been transferred to the backward-propagating wave. Thus it is generally *not* desirable to realize the largest possible index perturbations in this application.

FBG Dispersion Compensators

A second application for which FBGs have been used is dispersion compensation in fiber systems. As we have seen earlier, dispersion arises most commonly because different wavelengths in an optical fiber propagate at different speeds. In the usual case, the higher frequency (shorter wavelength) components of light travel faster than the lower frequency (longer wavelength) components.

While dispersion compensating fiber can be used to overcome this distortion, generally long lengths of such fiber are needed to provide adequate compensation. FBGs offer the promise to provide similar compensation in much shorter lengths.

[Figure 12.6](#) shows the basic idea behind the use of an FBG for this purpose. A chirped period grating is manufactured. Ideally this grating is designed to introduce a time delay as a function of frequency that exactly compensates for the delays implied by [Eq.\(12-4\)](#). A simpler understanding

can be gained by realizing that the long wavelengths, which were delayed the most by the dispersive fiber, are delayed the least by the chirped grating, while the opposite is true of the short wavelengths. The result is a compensated signal pulse that has dispersion largely removed.

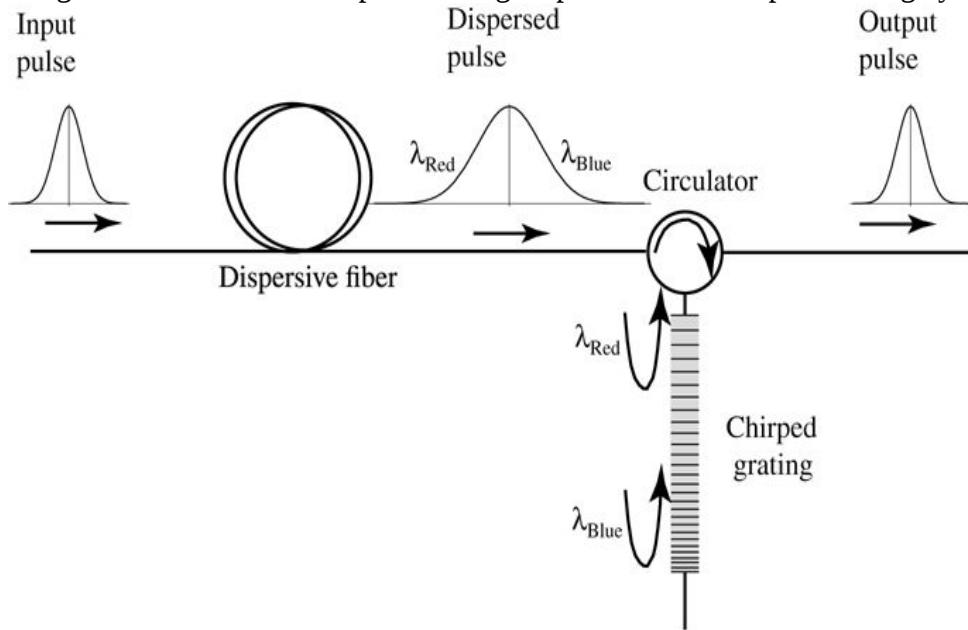


Figure 12.6

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 12.6 Dispersion compensation using a chirped FBG.

The illustration shows a horizontal line with three rightward pointing arrows, each with a bell curve. The curves are input pulse (to the left), dispersed pulse (in the center), and output pulse (to the right). The bell curve for the disperse curve is labeled lambda subscript Red on the left and lambda subscript Blue to the right. Between the input pulse and the dispersed pulse is a circular dispersive fiber. Between the dispersed pulse and the output pulse is a circulator, from which a chirped grating extends downward. It is represented by a column of short horizontal lines. Near the top end of the grating is a U turned arrow pointing upward. It is labeled lambda subscript Red. Near the lower end of the grating is a similar arrow labeled lambda subscript Blue.

It is possible to tune a grating such as this by heating or stretching the FBG. In this way the grating planes are moved slightly further away from one another, shifting the phase delay that is imparted to each wavelength. Thus a small amount of tuning of the dispersion compensation can be achieved, if needed.

12.2.5 Gratings Operated in Transmission

There do exist some applications in which the reflection grating geometry is inappropriate, and a transmission grating can serve a useful purpose. Such gratings are often called either “slanted gratings” or “long-period” gratings, depending on the type of grating.

The term *slanted grating* is applied to an FBG that has grating planes at an angle to the fiber axis. Typical angles of 2 to 3 degrees will nearly eliminate the main reflection peak. However, coupling to cladding modes that propagate backwards remains. If the grating period is chirped, the envelope of the cladding mode response defines the width of the loss peak in the forward direction, with a typical “rejection” bandwidth of 10 to 20 nm.

The term *long-period grating* is used for a transmission grating that has a period that induces coupling between the single mode in the core and multiple forward-propagating cladding modes, which eventually are scattered by the coating of the fiber. The period is typically in the range 100 μm to 1 mm, and the length typically 1 to 10 cm. Long-period gratings have wider rejection peaks than FBGs, with rejection bandwidths of a few 100 nm being typical in standard telecommunications fiber.

Slanted FBGs and long-period gratings are typically used for gain flattening in conjunction with fiber amplifiers, and for filtering in communications.

12.3 Ultrashort Pulse Shaping and Processing

Since the invention of the laser, the duration of optical pulses that can be generated in practice has been pushed shorter and shorter. Of particular interest have been pulses in the range from

picoseconds ($1 \text{ psec} = 10^{-12} \text{ sec}$) to femtoseconds ($1 \text{ fsec} = 10^{-15} \text{ sec}$). Pulses with durations of 100 fsec were demonstrated in 1981 [116]. Further progress pushed pulse durations to only a few fsec, corresponding to a small number of optical cycles.

Following success in the generation of ultrashort pulses, interest naturally grew in methods for changing simple short pulses into more complicated waveforms. This led to the invention of a variety of methods for performing such waveform shaping; here we focus attention on the most successful of these methods, pioneered by [Froehly \[119\]](#) and [Weiner and Heritage \[363\]](#). For general reviews of this method of ultrashort pulse shaping see [361] and [362].

12.3.1 Mapping of Temporal Frequencies to Spatial Frequencies

Pulses in the femtosecond range have spectra that occupy significant portions of the optical spectrum. For example, at the common long-distance fiber communication wavelength of 1550 nm, a 100 fsec pulse has a ratio of bandwidth to center frequency, $\Delta\nu/\nu$, of more than 5%, while a 10 fsec pulse has a similar ratio of more than 50%. These enormous optical bandwidths allow common dispersive elements, such as gratings, to provide sufficient spatial spreading of frequencies to facilitate a useable mapping of temporal frequencies into spatial positions. We briefly discuss this mapping in this subsection.

The simplest case to consider is that of a transmissive amplitude grating shown in [Fig. 12.7\(a\)](#). For plane wave illumination, the angle θ_2 of the $-1 - 1$ diffraction order is related to the period Λ of the grating, the angle θ_1 of illumination and the wavelength λ through the grating equation³, derived in [Appendix D](#),

$$\sin\theta_2 = \sin\theta_1 - \frac{\lambda}{\Lambda}.$$

$$\sin\theta_2 = \sin\theta_1 - \frac{\lambda}{\Lambda}.$$

(12-13)

In the case of a reflection geometry, as shown in [Fig. 12.7\(b\)](#), the same relationship holds. Only the $-1 - 1$ diffracted first order is shown, it being assumed that the blaze⁴ on the grating suppresses the $+1 + 1$ order and that the grating depth is such that the zero order is negligible.

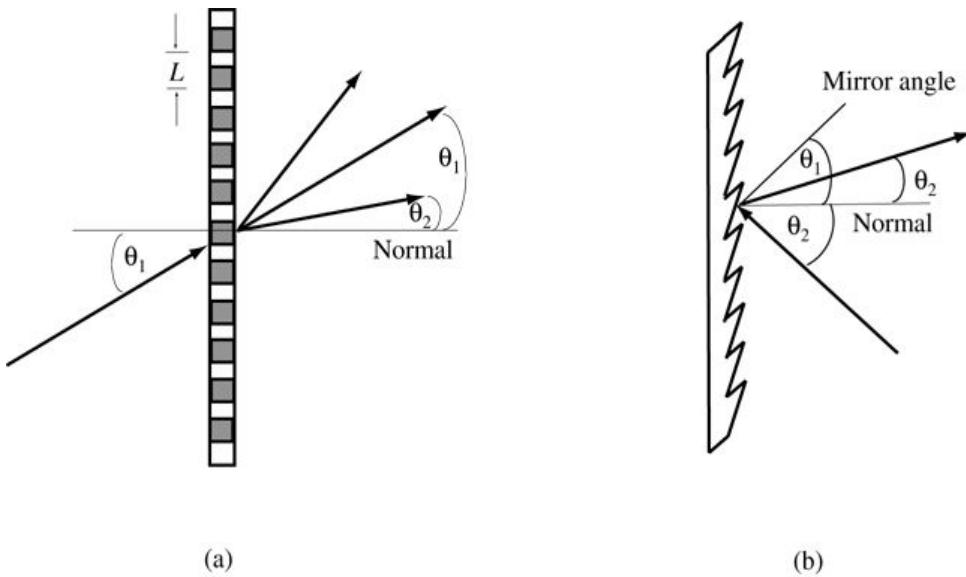


Figure 12.7

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 12.7 (a) Simple amplitude transmission grating. (b) Reflection grating.

Illustration a shows a grating represented by a vertical column of gray and white rectangles such that the combined height of a gray rectangle and its neighboring white rectangle is L. The normal is a horizontal line that cuts through the center of the grating. On the left side, below the normal is an upward sloping, rightward pointing arrow incident at the point where the normal intersects the grating. The arrow makes an angle measuring theta subscript 1 with the normal. The arrow extended to the other side and above the normal makes angle theta subscript 1 with the normal. On each side of the extended arrow, an arrow emerges from the intersection. The lower one makes an angle measuring theta subscript 2 with the normal; theta subscript 2 is less than theta subscript 1. The other arrow is higher. Illustration b shows a vertical grating represented by a shape with a straight line along its left edge and a saw-like jagged edge along its right edge. The normal is a horizontal axis line on the right side from a point near the center of the grating. From below the normal and downward sloping, leftward pointing arrow incident at the intersection of the grating and the normal. This angle is mirrored above the normal and measures theta subscript 1. An upward sloping, rightward pointing arrow from the intersection makes angle theta subscript 2 with the normal.

One additional element is needed to complete the time-to-space mapping—a lens. Let the grating be placed in or near the front focal plane of the lens, and observe light across the back focal plane. The lens in this geometry maps angles into positions in the back focal plane. The angle of the diffracted light depends on the direction of illumination and the wavelength of the light (or equivalently on the optical frequency). Different frequencies are thus mapped to different positions in the focal plane. The geometry is illustrated in [Fig. 12.8](#).

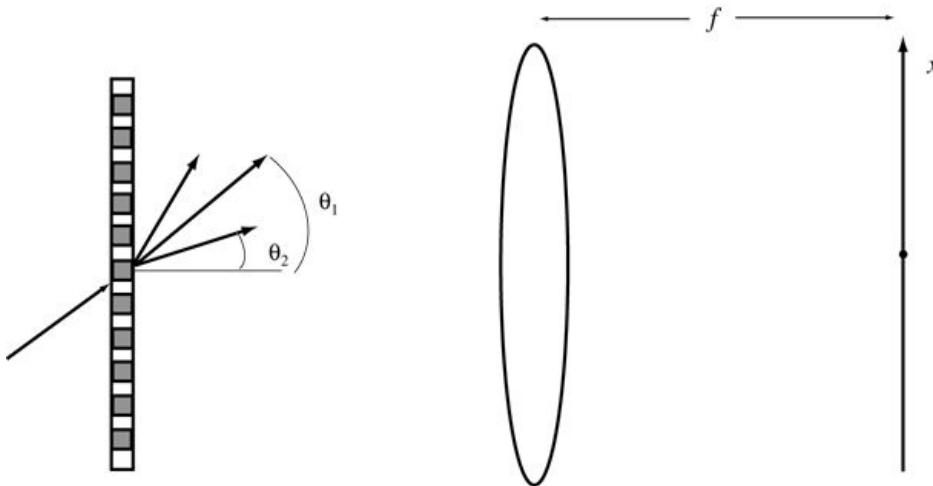


Figure 12.8

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 12.8 Geometry for mapping optical frequency onto space.

The first illustration shows a grating represented by a vertical column of gray and white rectangles. The normal is a horizontal line perpendicular at the center of the grating on the right side. On the left side, below the normal level is an upward sloping, rightward pointing arrow incident at the point where the normal intersects the grating. The arrow extended to the other side and above the normal makes angle theta subscript 1 with the normal. On each side of the extended arrow, an arrow emerges from the intersection. The lower one makes an angle measuring theta subscript 2 with the normal. The other arrow is higher.

The second illustration shows a vertically elongated oval at a distance of f from an upward pointing axis x .

To understand this mapping in more detail, we begin with the grating equation above. For small angle θ_2 of the -1^{-1} diffraction order away from the normal, this equation can be approximated as

$$\theta_2 \approx \sin\theta_1 - \frac{\lambda}{\Lambda}.$$

$$\theta_2 \approx \sin\theta_1 - \frac{\lambda}{\Lambda}.$$

(12-14)

Again for small angle θ_2 , the position x in the focal plane is related to the angle θ_2 of the diffracted component through⁵

$$x \approx f\theta_2,$$

$$x \approx f\theta_2,$$

(12-15)

obtained by tracing a ray at angle θ_2 through the center of the lens to the back focal plane. Substituting, we obtain

$$x = f \sin \theta_1 - f \lambda \Lambda = x_0 - f \lambda \Lambda,$$

$$x = f \sin \theta_1 - \frac{f \lambda}{\Lambda} = x_0 - \frac{f \lambda}{\Lambda},$$

(12-16)

where $x_0 = f \sin \theta_1$. The equivalent expression in terms of frequency $v = c/\lambda$ is

$$x = x_0 - f v \Lambda.$$

$$x = x_0 - \frac{f c}{v \Lambda}.$$

(12-17)

Knowing the parameters f , c , and Λ , we can now specify where each temporal frequency component (or wavelength component) of an incident plane wave falls in the focal plane. The results derived above for a transmission grating apply equally well for the reflection grating shown in Fig. 12.7(b).

12.3.2 Pulse Shaping System

A system capable of mapping an ultrashort pulse into a more complex signal is shown in Fig. 12.9 (after [362]). A plane wave pulse is input to the system from the lower right, travels to the first grating, is dispersed and mapped to the focal plane of the first lens, passes through a mask that modifies the amplitude and (in some cases) the phase of the temporal Fourier components of the pulse. The modified spectrum is then collapsed by the second lens and the second grating back to a plane wave, but with modified temporal spectral components. The final temporal signal then exits to the lower left.

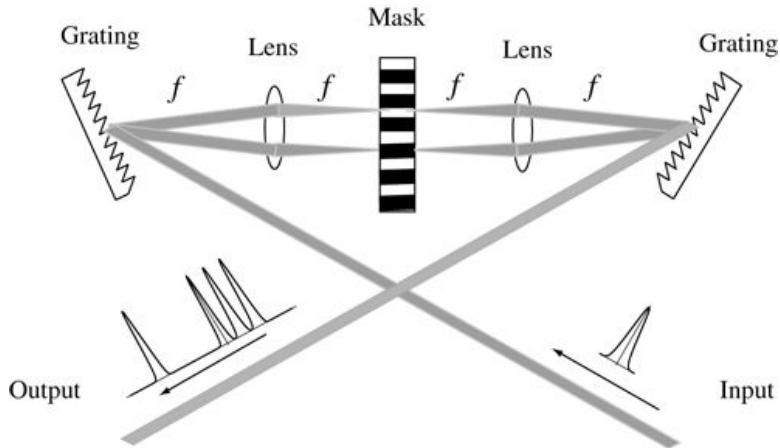


Figure 12.9
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 12.9 Pulse shaping by spectral filtering.

The illustration shows a downward sloping grating on the top left corner and an upward sloping grating on the top right corner. Between the two is a mask represented by a horizontally striped vertical column. On either side is a lens in the middle of the distance between the grating and the mask. The grating to lens and the lens to mask distance are each of length f .

From the bottom right corner, the input, represented by a narrow bell curve, moves in an upward leftward direction to the center of the grating in the top left corner. From the grating, two rays diverge to the extremes of the lens, beyond which the rays run parallel and perpendicular to the mask. They continue their horizontal path beyond the mask and cross the lens on the other side; thereafter they converge on the grating in the top right corner. From here in a downward leftward path to the bottom left corner is the movement of the output represented by a graph with four sharp spikes, three in a row and the fourth one at a distance.

The optical system in this case uses a tilted input reflection grating in order to direct the diffracted components closer to the optical axis of the lenses than would otherwise be the case. The output grating is likewise tilted, and the “4f” optical system forms a telescopic imaging system. Since the focal lengths of the two lenses are identical, the magnification is unity, and the input grating is imaged onto the output grating.

The mask in the focal plane can be of several different types, similar to the possibilities in coherent optical processing. An absorbing mask will modify the amplitudes of the temporal Fourier components, while a phase mask will modify their phases. Two such masks used together can control the complex amplitude of the spectral components. A spatial light modulator can be used to dynamically change the amplitude, phase, or both amplitude and phase of the filter. Programmable liquid crystal SLMs [364] and acousto-optic cells [171] have been used for these purposes.

If the goal is to synthesize a temporal filter with transfer function $H(v)$, then the amplitude transmittance of the mask required in the focal plane can be found by solving (12-17) for v and substituting this expression as the argument for H . The relation is⁶

$$v = cf\Lambda(x_0 - x).$$

$$\nu = \frac{cf}{\Lambda(x_0 - x)}.$$

(12-18)

Using the above result, it is clear that the amplitude transmittance in the focal plane should be

$$t(x) = H\left(\frac{cf}{\Lambda(x_0 - x)}\right).$$

$$t(x) = H\left(\frac{cf}{\Lambda(x_0 - x)}\right).$$

(12-19)

Note that it is only necessary to achieve this transmittance for the particular diffraction order that is actually used, in this case the -1^{-1} order.

12.3.3 Applications of Spectral Pulse Shaping

Shaping of ultrashort pulses by the method described above has found applications in several different scientific fields, including nonlinear optics, femtosecond spectroscopy, and ultrafast laser-materials interactions. Here, in concert with our chapter title, we focus on an application in the field of optical communications.

Application to Code Division Multiple Access

The application considered here is code division multiple access (CDMA) waveform generation and decoding. CDMA is an encoding and decoding process that assigns to each user on a multiuser communication channel a unique coded signal that is (ideally) orthogonal to the coded signals assigned to all other users. The orthogonality of the coded signals allows one user to address a message to another user using the specific coded waveform appropriate for the recipient. The original message consists of a sequence of ultrashort pulses, with a binary “1” represented by the presence of a pulse in a given time slot, and a binary “0” represented by the absence of a pulse in that time slot. Each binary “1” is encoded by the spectral coding process discussed in the previous section, into an extended waveform appropriate for the intended recipient for this particular message. Each sender must be equipped with a changeable mask (i.e. an SLM) so that the coded waveform appropriate for any of the possible destinations can be generated.

Note that with full complex encoding of the amplitudes of the temporal Fourier components, optical waveforms with both amplitude and phase modulations can be realized.⁷ In practice, however, the advantages of complex encoding are not great, and binary phase SLMs and spectral codes consisting of a spatial sequence of 0 and π phase shifts are often used. A sequence of such phase shifts is a code. A single location on the network has a unique spectral code associated with that location. Any other user can address this location by using this particular code.

If a particular user wishes to receive messages addressed to him/her, then that user loads into the local SLM a mask that is the complex conjugate of spectral encoding mask used by any sender for this destination. The extended encoded signal is then compressed into an ultrashort pulse at the local receiver. In effect, the decoding system is being used as a matched filter. If this same user wishes to communicate with another user, then the local SLM is loaded with the spectral mask containing the code appropriate for the user to whom the message is directed. [Figure 12.10](#) illustrates the idea. Four users are shown, connected by a fiber ring. Each user node is coupled to the fiber such that a portion of the circulating signal is tapped off the ring and detected. In addition, each node is coupled to the ring such that a message can be inserted on the ring. Inside each box labeled “user” is a spectral filtering system such as shown in [Fig. 12.9](#), with an SLM providing a dynamic spectral mask. In the above figure, User 1 is shown injecting an ultrashort pulse (i.e. a binary “1”) into the local spectral filtering system (the box labeled “User 1”), which then emits a waveform with a spectral code appropriate for User 3. User 3 is in reception mode, and compresses the coded waveform into an ultrashort pulse, which is then detected. Users 2 and 4 have spectral masks that are their own codes, which are orthogonal to the code of User 3. Thus these two users do not find ultrashort pulses at their outputs. If each receiver has a threshold circuit, then only the properly compressed pulses will be detected, and only User 3 will receive the message.

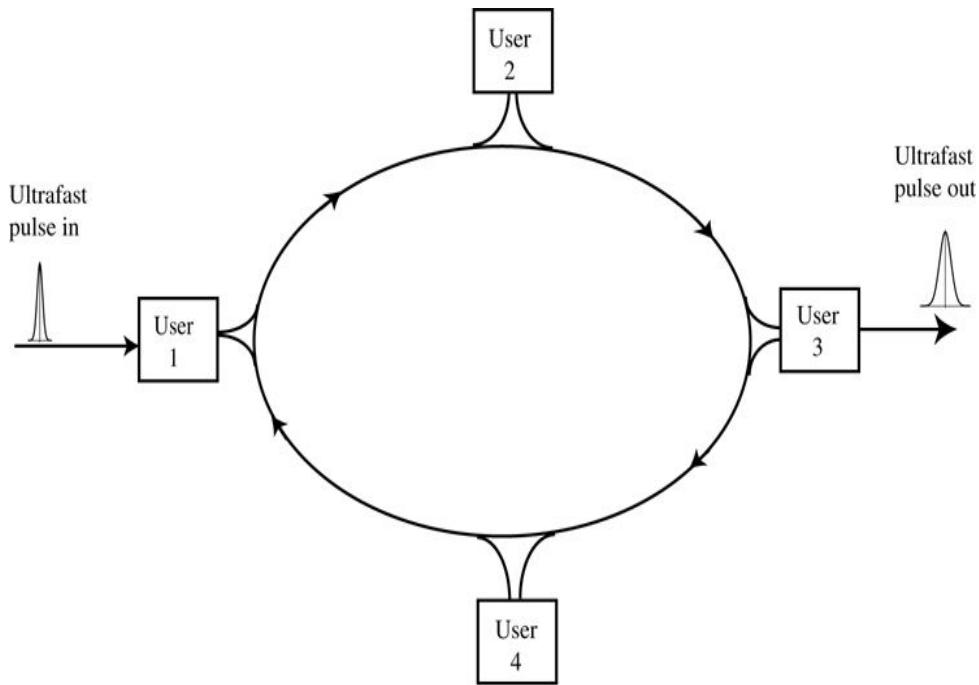


Figure 12.10

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 12.10 Typical CDMA system.

The illustration shows an oval path with a box at each of its four extremes. Beginning at the left extreme and moving in the clockwise direction are User 1, User 2, User 3, and User 4. A rightward arrow pointing at User 1 appears with the label “Ultrafast pulse in” and a narrow bell curve. Four arrows, one between every two neighboring users, all point in the clockwise direction. On the right extreme, a rightward arrow points away from User 3 and appears with the label “Ultrafast pulse out” and a narrow bell curve.

The above discussion is a very simplified example of the ideas behind the use of CDMA on optical fiber networks. There are other network architectures that could be used, and many different types of codes that can be utilized. Indeed the problem of finding the best possible codes has been an active area of research (see, for example, [249]). Many areas have not been touched on here, such as network synchronization and network protocols. For more details, the reader is referred to publications that describe actual systems, such as [289] and [306].

Application to Fiber Dispersion Compensation

As described in (12-4), light propagating over long distances in a single-mode optical fiber suffers dispersion, i.e. different wavelengths propagate at different speeds. The dominant distortion to the propagating signal comes from a quadratic-phase distortion with frequency, but additional distortion can arise from the third-order distortion term as well. One approach to compensating distortion is to use a length of dispersion-compensating fiber to remove the quadratic-phase distortions and a spectral filtering system to remove third-order distortions. Such an approach has been used to restore 500 psec pulses that have been broadened to 400 times their original length by propagation through ordinary single mode fiber. A length of dispersion-compensating fiber narrowed the pulse length to twice its original length, and a spectral filtering system reduced the pulse length to its original 500 psec duration [59].

12.4 Spectral Holography

The concepts associated with ultrashort pulse shaping have been extended to a field known as *spectral holography* [365]. Using techniques to be described, it is possible to record a spatial hologram of a signal temporal waveform, using a femtosecond pulse as a reference signal, and later reconstruct that waveform by addressing the hologram with a femtosecond probe or reconstruction pulse.

12.4.1 Recording the Hologram

A typical geometry for recording the temporal hologram is shown in Fig. 12.11. As in previously described methods for mapping temporal frequencies into spatial positions, a tilted grating is used at the input to the recording system. The grating is tilted horizontally, and the grating lines run vertically. The signal time waveform and a femtosecond reference pulse are incident simultaneously on different confined regions of the grating. In the figure, the reference pulse is incident on a small region near the bottom of the grating, and the signal waveform is incident on a small region near the top of the grating. The locations of these positions determine the angles at which the two beams strike the hologram plane. As the two beams leave the grating, each spreads in the horizontal (x^x) direction due to the grating dispersion, and each spreads a small amount in the vertical (y^y) direction due to diffraction. The propagation distance to the spherical lens is one focal length. After passing through the lens, the two signals propagate to the back focal plane of the lens, where they overlap. It is assumed that the two signals originate from the same laser and are mutually coherent, and therefore they interfere at the hologram plane, where a photosensitive medium is situated. The elliptical area shown on the back of the hologram represents the recording area, which actually exists on the front of the hologram. Because the reference pulse is extremely short, its spectrum is extremely broad, and covers the elliptical recording region shown. The spectrum of the signal waveform is more complex, varying in both amplitude and phase as a function of temporal frequency. These variations are captured by interference with the spectrum of the reference pulse.

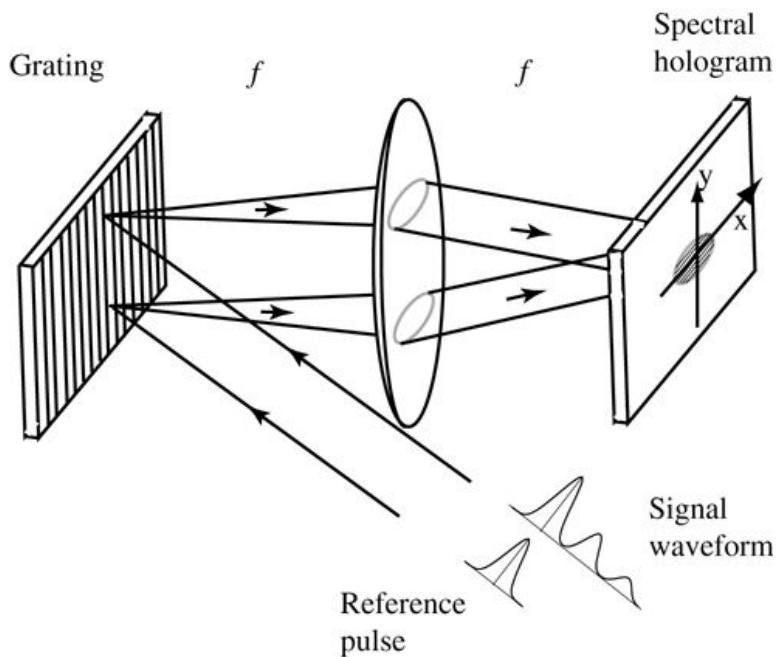


Figure 12.11

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 12.11 Recording a Spectral hologram.

The illustration shows a spherical lens between a titled grating in the top left corner and a special hologram in the top right corner. All three are similarly oriented. The distances from the lens to the grating and from the lens to the hologram are each of length f . From the bottom right corner, two rays are pointed in the top left corner. One, namely signal waveform, reaches the top end of the grating while another, namely reference pulse, reaches the lower end of the grating. From the grating, the signal waveform diverges in a rightward movement to the extremes of an oval area in the top half of the lens, beyond which it runs mutually parallel to reach the center of hologram. Similarly, the reference pulse diverges in a rightward movement to the extremes of an oval area in the lower half of the lens, beyond which it runs mutually parallel to reach the center of hologram. On the rectangular hologram is a coordinate plane with vertical axis y and horizontal axis x .

A mathematical description of the process can be written easily based on the mathematics of holography introduced in [Chapter 11](#). Let $R(\nu)$ and $S(\nu)$ represent the complex temporal spectra of the reference pulse and the signal waveform, respectively.⁸ The intensity incident on the hologram recording plane can then be described by

$$J(x,y) = |R(\nu)|^2 + |S(\nu)|^2 + R^*(\nu)S(\nu)\exp(-j2\pi\theta\nu y/c) + R(\nu)S^*(\nu)\exp(j2\pi\theta\nu y/c),$$

$$\begin{aligned} J(x, y) = & \quad |R(\nu)|^2 + |S(\nu)|^2 + R^*(\nu)S(\nu) \exp(-j2\pi\theta\nu y/c) \\ & + R(\nu)S^*(\nu) \exp(j2\pi\theta\nu y/c), \end{aligned} \tag{12-20}$$

where θ is the vertical angle between the signal and reference beams (assumed to be a small angle, for simplicity). To express the above results in terms of x and y , rather than ν , we

must use (12-18). The small angle assumption allows the substitution

$$v = cf\Lambda(x_0 - x) = \mu/(x_0 - x),$$

$$\nu = \frac{cf}{\Lambda(x_0 - x)} = \mu \Big|_{(x_0 - x)},$$

(12-21)

where Λ is again the period of the grating, x_0 is the x coordinate where the zero order of the grating would strike the focal plane, and $\mu = cf/\Lambda$. Substituting this expression in (12-20) yields

$$\mathcal{J}(x, y) = R\mu/(x_0 - x)^2 + S\mu/(x_0 - x)^2 + R^*\mu/(x_0 - x)S\mu/(x_0 - x)\exp(-j2\pi f\theta y\Lambda(x_0 - x)) + R\mu/(x_0 - x)S^*\mu/(x_0 - x)\exp(j2\pi f\theta y\Lambda(x_0 - x)) = R\mu/(x_0 - x)^2 + S\mu/(x_0 - x)^2 + 2R\mu/(x_0 - x)S\mu/(x_0 - x)\cos(2\pi f\theta y\Lambda(x_0 - x) - \phi\mu/(x_0 - x)),$$

$$\begin{aligned} \mathcal{J}(x, y) &= |R(\mu/(x_0 - x))|^2 + |S(\mu/(x_0 - x))|^2 \\ &\quad + R^*(\mu/(x_0 - x))S(\mu/(x_0 - x))\exp\left(-j\frac{2\pi f\theta y}{\Lambda(x_0 - x)}\right) \\ &\quad + R(\mu/(x_0 - x))S^*(\mu/(x_0 - x))\exp\left(j\frac{2\pi f\theta y}{\Lambda(x_0 - x)}\right) \\ &= |R(\mu/(x_0 - x))|^2 + |S(\mu/(x_0 - x))|^2 \\ &\quad + 2|R(\mu/(x_0 - x))||S(\mu/(x_0 - x))|\cos\left[\frac{2\pi f\theta y}{\Lambda(x_0 - x)} - \phi(\mu/(x_0 - x))\right]. \end{aligned} \quad (12-22)$$

where $\phi(v)$ is the phase angle of the signal waveform spectrum at each value of v . Note that, neglecting the phase modulation ϕ , the carrier frequency fringes are tilted in a kind of radial spoke pattern, due to the fact that the frequency v is changing in the x direction. A line of $n2\pi$ phase of the carrier portion of the argument of the cosine occurs when

$$2\pi f\theta y\Lambda(x_0 - x) = n2\pi$$

$$\frac{2\pi f\theta y}{\Lambda(x_0 - x)} = n2\pi$$

or

$$y = n\Lambda(x_0 - x)f\theta.$$

$$y = \frac{n\Lambda(x_0 - x)}{f\theta}.$$

The slope of this line is $-\frac{n\Lambda}{f\theta}$, which changes with the integer n chosen. [Figure 12.12](#) illustrates a density plot of a portion of the typical fringe structure in the focal plane. The degree to which the slope of the fringes is noticeable depends on the geometry and the grating dispersion.

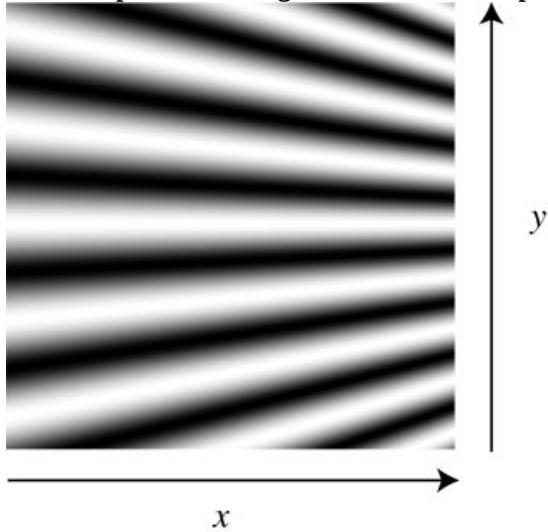


Figure 12.12

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 12.12 Fringes in the focal plane.

12.4.2 Reconstructing the Signals

To reconstruct the signal waveform, the system shown in [Fig. 12.13](#) is used. In the figure, a femtosecond probe (reconstruction) pulse illuminates the input grating, but this time the signal waveform is not present. The probe pulse leaves the grating and propagates to the lens and then to the hologram, where its spectrum is incident, spread in the x direction. As usual, we assume that the material recording the hologram produces an amplitude transmittance that is proportional to the original exposing intensity.

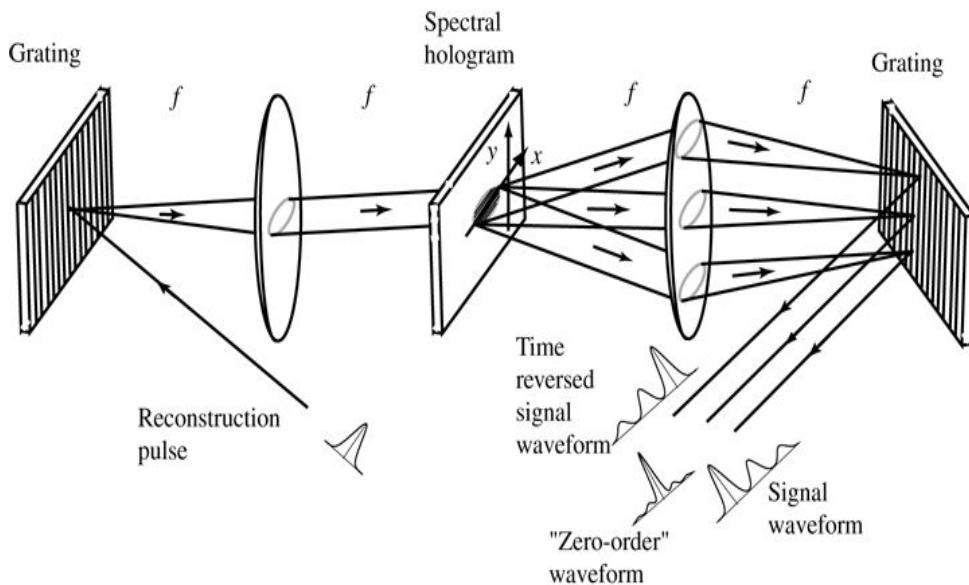


Figure 12.13

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 12.13 Reconstructing the temporal signal.

The illustration shows along a horizontal plane in the left to right direction a grating, a lens, a spectral hologram, another lens, and a grating. On either side of the hologram, the lens is as far from the hologram as it is from the grating, the distance being f . From below, an upward pointing rightward arrow representing a reconstruction pulse reaches the center of the grating in the left extreme. From the point of incidence, rays diverge and pass through the single oval area at the center of the lens, beyond which the rays run mutually parallel and extend up to the hologram. Thereafter, the rays split into three beams, one running straight ahead, another sloping upward and yet another sloping downward. These three beams enter the lens, each through an oval area in the lens, beyond which each converges separately and reaches a point along a vertical line on the grating in the right extreme. The middle beam converging at the middle, the top beam a little above and the bottom beam a little below.

From the top point of convergence, rays are reflected in a downward leftward arrow labeled “time reversed signal waveform.” The one from the middle point is labeled “Zero-order waveform.” And the one from the lower point is labeled “Signal waveform.”

For the moment we assume that the probe pulse has a spectrum $P(\nu)$, possibly different from that of the reference pulse. Neglecting a proportionality constant, we have a field transmitted by the hologram given by three distinct terms

$$U(x,y) = P\mu/(x_0-x)R\mu/(x_0-x)^2 + S\mu/(x_0-x)^2 + P\mu/(x_0-x)R^*\mu/(x_0-x)S\mu/(x_0-x)\exp[-j2\pi f\theta y]\Lambda(x_0-x) + P\mu/(x_0-x)R\mu/(x_0-x)S^*\mu/(x_0-x)\exp[j2\pi f\theta y]\Lambda(x_0-x).$$

$$\begin{aligned}
U(x, y) = & P(\mu / (x_0 - x)) \left[|R(\mu / (x_0 - x))|^2 + |S(\mu / (x_0 - x))|^2 \right] \\
& + P(\mu / (x_0 - x)) R^*(\mu / (x_0 - x)) S(\mu / (x_0 - x)) \exp \left(-j \frac{2\pi f \theta y}{\Lambda(x_0 - x)} \right) \\
& + P(\mu / (x_0 - x)) R(\mu / (x_0 - x)) S^*(\mu / (x_0 - x)) \exp \left(j \frac{2\pi f \theta y}{\Lambda(x_0 - x)} \right).
\end{aligned} \tag{12-23}$$

When both the reference and the probe are single femtosecond pulses, their spectra are approximately flat over the portion of the hologram that contains the signal waveform spectrum, and therefore the transmitted field becomes

$$U(x, y) = P_0 R_0 + S(\mu / (x_0 - x)) \exp(-j2\pi f \theta y \Lambda(x_0 - x)) + P_0 R_0 S^*(\mu / (x_0 - x)) \exp(j2\pi f \theta y \Lambda(x_0 - x)).$$

$$\begin{aligned}
U(x, y) = & P_0 \left[|R_0|^2 + |S(\mu / (x_0 - x))|^2 \right] \\
& + P_0 R_0 S(\mu / (x_0 - x)) \exp \left(-j \frac{2\pi f \theta y}{\Lambda(x_0 - x)} \right) \\
& + P_0 R_0 S^*(\mu / (x_0 - x)) \exp \left(j \frac{2\pi f \theta y}{\Lambda(x_0 - x)} \right).
\end{aligned} \tag{12-24}$$

(12-24)

where P_0 and R_0 are the uniform spectral amplitudes of the probe and reference beams respectively, assumed real valued. These three wave-components⁹ propagate to the lens, and are focused by the lens onto separate spots on the grating, as shown in Fig. 12.13. Consider all three temporal signals that leave the grating after having their spectral dispersion cancelled. The three output signals are physically separated in this illustration, and have different relations to the original waveforms. The signal generated by the first term consists of a combination of the probe or reference pulse (which are identical in this special case) and the autocorrelation of the signal waveform, with the relative strength of these two terms depending on the beam ratio when the hologram was recorded. This signal is analogous to the on-axis term reconstructed by a conventional hologram and is represented by the middle waveform emerging at the bottom of Fig. 12.13. The second term in (12-25) reconstructs a duplicate of the original signal waveform, analogous to the virtual image produced by a conventional hologram. The third term in this equation, which is an amplitude proportional to $S^* S^*$, generates a *time reversed* version of the original signal waveform, which is analogous to the real image of a conventional hologram.

In practice it may be desirable to have the probe pulse incident on the input grating at different positions than shown here. If the goal is to generate a duplication of the original signal waveform, the probe can be introduced at the location where the reference pulse originally hit the grating. A lens of limited size may then capture only two of the three grating orders, allowing only the “on axis” and signal waveform terms to emerge. Alternatively, if the goal is to generate a time-reversed signal waveform, introduction of the probe pulse at the grating location where the signal waveform was originally incident may be advantageous. These strategies are aimed at making best use of lenses of finite aperture.

The assumption used above that both the reference and probe beams are simple femtosecond pulses can be changed to yield a more general temporal signal processing capability. Refer to the more general equation for the field transmitted by the hologram, (12-24), and consider the inverse Fourier transforms of the three major terms (neglecting the exponential terms, which simply lead to offsets in spatial positions):

$$\begin{aligned}\mathcal{F}^{-1}P(v)R(v)2+S(v)2&=p(t)*r(t)*r^*(-t)+p(t)*s(t)*s^*(-t)\mathcal{F}^{-1}P(v)R^*(v)S(v)&=p(t)*r^*(-t)*s(t)\mathcal{F}^{-1}P(v)R(v)S^*(v)&=p(t)*r(t)*s^*(-t),\\\mathcal{F}^{-1}\{P(\nu)[|R(\nu)|^2+|S(\nu)|^2]\}&=p(t)*r(t)*r^*(-t)+p(t)*s(t)*s^*(-t)\\\mathcal{F}^{-1}\{P(\nu)R^*(\nu)S(\nu)\}&=p(t)*r^*(-t)*s(t)\\\mathcal{F}^{-1}\{P(\nu)R(\nu)S^*(\nu)\}&=p(t)*r(t)*s^*(-t),\end{aligned}\tag{12-25}$$

where $p(t)$, $r(t)$ and $s(t)$ are the time waveforms of the probe, reference and signal, respectively. Clearly very general linear signal processing operations can be realized with appropriate choices for $p(t)$ and $r(t)$. It is also possible to realize some nonlinear signal processing operations by taking advantage of the nonlinearities of the holographic recording medium. Note that the hologram could be computer generated, rather than optically generated, adding another flexibility to the process. See [362] for more details.

12.4.3 Effects of Delay between the Reference Pulse and the Signal Waveform

Before closing this topic, some examination of the effects of relative delay of the reference and signal waveforms is worthwhile. Again let $r(t)$ represent the reference pulse, and $s(t)$ the signal waveform. Consider a delay of the signal with respect to the reference, $s(t-\tau_0)$ (note τ_0 could be positive or negative, according to whether the signal is delayed or advanced with respect to the reference pulse). If $S(v)=\mathcal{F}s(t)$, then

$$\mathcal{F}s(t-\tau_0)=S(v)\exp\{-j2\pi v \tau_0\}.$$

$$\mathcal{F}\{s(t-\tau_0)\}=S(\nu)\exp\{-j2\pi\nu\tau_0\}.$$

The spectral resolution at the recording plane is limited by the period of the grating and the finite size of the spot illuminated on the grating by the signal beam. In effect, the spectra falling on the hologram are convolved with an amplitude spread function associated with this limited spectral resolution. As a consequence of this convolution, each point in the spectral plane has a range of optical frequencies present. The reference and signal spectra interfere on a frequency-by-frequency basis. As a result, several simultaneous fringe patterns are present at each point on the hologram; these fringes have spatial frequencies that are nearly the same, but phases that are different due to the presence of the linear phase shift with frequency caused by the time difference between reference and signal. If the phase shift $2\pi v \tau_0$ changes by 2π radians or more within a single resolution cell in the spectrum, then the fringe patterns will largely cancel one another due to their different phases, leaving a constant intensity but no fringe pattern at all. Thus

there is a maximum time separation that can be tolerated between the reference and signal—in effect a finite time window exists around the reference pulse during which signals can be recorded holographically. The time window is wider if the spectral resolution is higher in the hologram plane. See [Prob. 12-3](#) for further exploration of this effect.

Finally, in closing we mention a related subject, called *temporal imaging*, in which temporal analogs of lenses, free space propagation and imaging can be realized. The reader is referred to [\[204\]](#) for some examples, and to [\[22\]](#) for discussion of an application to temporal microscopy.

12.5 Arrayed Waveguide Gratings

With the rise of dense wavelength-division multiplexing techniques in optical communications comes the need to multiplex, demultiplex and route wavelengths with spectral precisions that were previously not required. Naturally, cost and reliability are of extreme importance in choosing solutions for these requirements. These facts have led to consideration of a variety of integrated optics solutions that promise to have the cost and reliability attributes required. In this section we review one such solution, the arrayed waveguide grating (AWG), which has interesting interpretations from a Fourier optics point of view.

The origin of the AWG lies in papers published by [Takahashi \[339\]](#) and [Dragone \[95\], \[94\]](#). An excellent in-depth discussion of such devices can be found in [\[267\]](#). We begin by introducing the various integrated components of an AWG. The overall architecture is then considered, and some applications are described.

12.5.1 Component Parts of an Arrayed Waveguide Grating

An AWG is a fairly complex integrated device made up of several more simple integrated components, as shown in [Fig. 12.14](#). Here we briefly describe these basic components, including waveguides for transporting the optical signals, star couplers for fan-out and fan-in of optical signals, and the waveguide grating for wavelength dispersion.

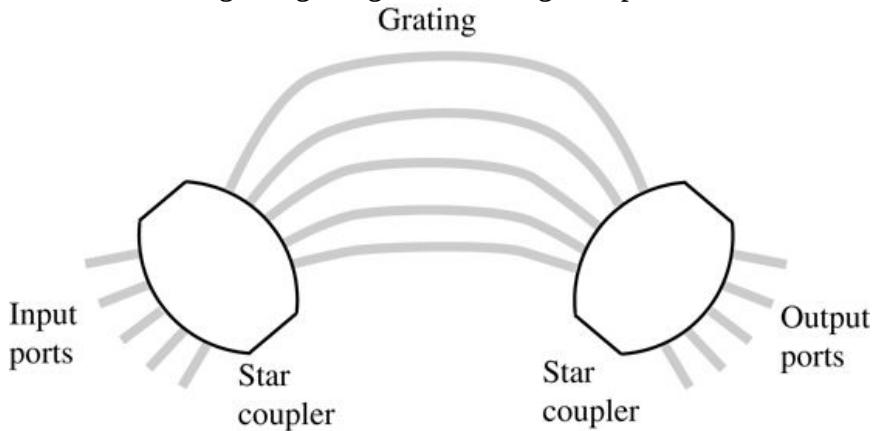


Figure 12.14

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 12.14 Architecture of an arrayed waveguide grating.

The illustration shows two star couplers, each represented by a four sided figure of two identical circular arcs facing outward in opposite directions with their corresponding extremes connected by straight lines. The couplers are connected by five arching curves representing a grating. To the left of left coupler are 5 input ports and to the right of the right coupler are 5 output ports.

Integrated Optics Waveguides

The basic building block of an integrated optical circuit is a waveguide. Since this technology is fundamentally planar, the waveguides are usually rectangular in shape, rather than circular as in the case of optical fibers. [Figure 12.15](#) shows a cross section of a typical rectangular guide.

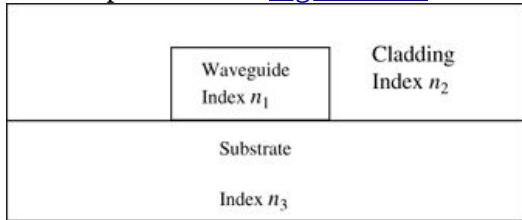


Figure 12.15

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 12.15 Cross section of a rectangular waveguide.

The illustration shows a rectangle with its longer side oriented horizontally. A horizontal line divides it into two halves. The lower half is Substrate, Index n_3 . In the upper half is a small rectangle for Waveguide, Index n_1 located at the center of the dividing line. The area in the upper half surrounding Waveguide is Cladding, Index n_2 .

The theory of propagation in single mode rectangular dielectric waveguides is complex and will not be dealt with in depth here (see [\[177\]](#) for in-depth analysis). Complexity arises from the fact that modal confinement is different horizontally and vertically due to the rectangular

geometry, and different at the top and bottom interfaces of the waveguide when $n_2 \neq n_3$. For our purposes it suffices to characterize the waveguide by an effective propagation constant β_{eff} , that in general depends on the geometry, the polarization of the light, as well as the optical frequency.¹⁰ To accurately find β_{eff} generally requires a numerical approach. To design an AWG device requires accurate modeling of such waveguides, but to understand the general operation of the device it suffices to understand the function of the waveguides.

Rectangular waveguides play the role of wires in electrical circuits, connecting various optical components and transporting optical signals to them, sometimes with carefully controlled phase delays.

Integrated Star Couplers

The purpose of a star coupler is to spread a portion of the signals appearing on each and every input port to all output ports (fan out), and to collect portions of the signals from each input port at each and every output port (fan in). The input and output ports are themselves rectangular waveguides used for transporting signals into and out of the device. In some applications, there may be one input port and N^N output ports, in others N^N input ports and 1 output port. Most generally there are M^M input ports and N^N output ports, and the symmetrical case of $N \times N$ $N \times N$ is probably most common. [Figure 12.16](#) illustrates the fan-out and fan-in operations.

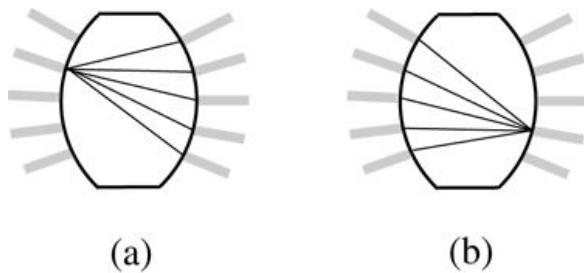


Figure 12.16

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 12.16 Star coupler showing (a) fan out from a particular input port to all output ports, and (b) fan in from all input ports to a particular output port. Similar operations happen for all input ports and all output ports simultaneously.

In both a and b, the coupler is represented by a four sided figure of two identical circular arcs facing outward in opposite directions with their corresponding extremes connected by straight lines. Each of the couplers has five input ports entering its left side and five output ports exiting its right side.

In illustration a, five line segments connect a single input port to the five output ports. In illustration b, a line segment from each of the input ports connects to a particular output port.

A variety of methods can be used to realize integrated-optic star couplers. Here we describe a method for which an understanding of Fourier optics is helpful. This method was pioneered by [Dragone \[93\]](#).

The star coupler consists of a comparatively wide but vertically thin planar waveguide (a so-called “slab” waveguide), with curved end faces that interface to smaller rectangular waveguides at the input and output. The curve of each end face is a circular arc, such that the center of curvature of each arc lies on top of the opposite circular arc at its midpoint. Thus the circular arcs are confocal. The geometry is illustrated in [Fig. 12.17](#). In practice, the waveguides would be much closer to one another than shown, in order to maximize efficiency.

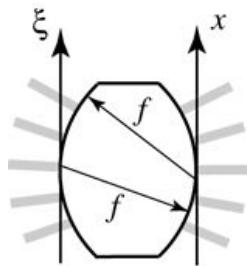


Figure 12.17

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure 12.17 Star coupler geometry. f is the radius of both circular arcs.

The illustration shows a star coupler represented by a four sided figure of two identical circular arcs facing outward in opposite directions with their corresponding extremes connected by straight lines. The coupler is oriented such that its straight line sides are horizontal. Entering its left arc are five input ports and exiting its right arc are five output ports. Tangential to each of the arcs is an upward pointing axis, xi on the left and x on the right. From where the xi axis meets the arc a

downward arrow extends up to the lower half of the right arc. From where the x axis meets the arc an upward arrow extends up to the upper half of the left arc. Each of these arrows is of length f.

This geometry is a one-dimensional analog of the diffraction geometry shown in [Fig. 4.5](#), which considered diffraction between two confocal spherical caps. When diffraction occurs between two such surfaces under paraxial conditions, the result is a two-dimensional Fourier transform relationship between the complex fields on the two surfaces. In a similar fashion, under paraxial conditions, the fields on the two circular arcs of the star coupler are related by a one-dimensional Fourier transform. If $U(\xi)$ represents a coherent complex field on the left surface of the star coupler, and $U(x)$ represents the complex field on the right surface of the star coupler, and light is propagating left to right, then

$$U(x) = e^{j2\pi f/\lambda} e^{-j\lambda f} \int_{-\infty}^{\infty} U(\xi) e^{-j\frac{2\pi}{\lambda} x \xi} d\xi.$$

$$U(x) = \frac{e^{j2\pi f/\lambda}}{\sqrt{j\lambda f}} \int_{-\infty}^{\infty} U(\xi) e^{-j\frac{2\pi}{\lambda} x \xi} d\xi.$$

(12-26)

Note that x and ξ are measured on lines parallel to one another and tangent to the midpoints of the circular arcs that compose the ends of the star coupler; the wavelength λ is the wavelength in the slab waveguide, which will depend on the optical frequency and the effective speed of propagation in that waveguide.

If we ignore coupling between adjacent guides, ignore light in the cladding of the guides, and ignore the vertical structure across the thin dimension of the guide, then a reasonable approximation for the field across the end of one input guide is a truncated Gaussian. The field across the output surface of the star coupler is then a convolution of a sinc function (from the truncation) with a Gaussian (the Fourier transform of a Gaussian is Gaussian). The width of one input waveguide must be small enough to spread the output field over the region of the output surface occupied by output waveguides.

For a stand-alone star coupler, it is usually desired to achieve a distribution of light across the output waveguides that is as close to uniform as possible. However, in the case of star couplers used as part of an AWG, this is not generally true. Some tapering of the distribution, so that it is strongest near the center, supplies an apodization to the Fourier transforming operations of the star couplers and thereby reduces the sidelobes of the arrayed waveguide grating.

As a caution, it should be mentioned that the various optical signals entering on the input ports of an AWG are usually not coherent with one another—they often originate from different mutually incoherent sources. However, the field introduced on the left of the input star coupler by any one input waveguide is coherent across the extent of the waveguide, and its Fourier transform across the output surface of that star coupler (i.e. the input to the grating section) is fully coherent.

For the output star coupler in an AWG, the various waveguides entering this star coupler contain some signals that are coherent with one another and some that are not mutually coherent. Each mutually coherent group of signals is focused by the star coupler onto an output waveguide.

A design constraint that must be placed on such a coupler is that the acceptance angle of the output waveguides be large enough to allow light arriving at the widest possible angle from the input guides to be captured. Another way of stating this constraint rests on the principle of

reciprocity—if light were input to the star coupler from one of the *output* ports, that light should spread over the *input* surface of the coupler sufficiently to cover all input waveguides. This condition in turn places a constraint on how short the star coupler can be made.

Note that the minimum possible loss associated with a $1 \times N$ star coupler is the splitting loss, $10\log_{10}N$ dB, and in practice the loss is more, due to the finite fill factor of the output waveguides. For a 64×64 star coupler, excess losses of 2 to 3 dB above and beyond the splitting loss can be realized in practice [267].

Waveguide Grating

A free-space grating consisting of an opaque screen with equally spaced holes is shown on the left of Fig. 12.18, and on the right the waveguide grating portion of the AWG is illustrated, showing both the waveguides and the end faces of this region. Consider first the free space grating. On the left, an incident wavefront is shown, as well as the rays traveling to the holes in the screen. On the right, rays and a wavefront are shown leaving the grating, corresponding to a grating order deflected downward.¹¹ In accord with the discussion in Appendix D, since $\theta_2 < \theta_1$, the case shown corresponds to a *negative* diffraction order $m < 0$. Using the main result from that appendix, we can apply the grating equation,

$$\sin\theta_2 = \sin\theta_1 + m\lambda/\Lambda$$

$$\sin\theta_2 = \sin\theta_1 + m\frac{\lambda}{\Lambda}$$

(12-27)

to the situation at hand. This equation describes the relationships between the various physical quantities involved, provided the signs of these quantities are properly assigned.

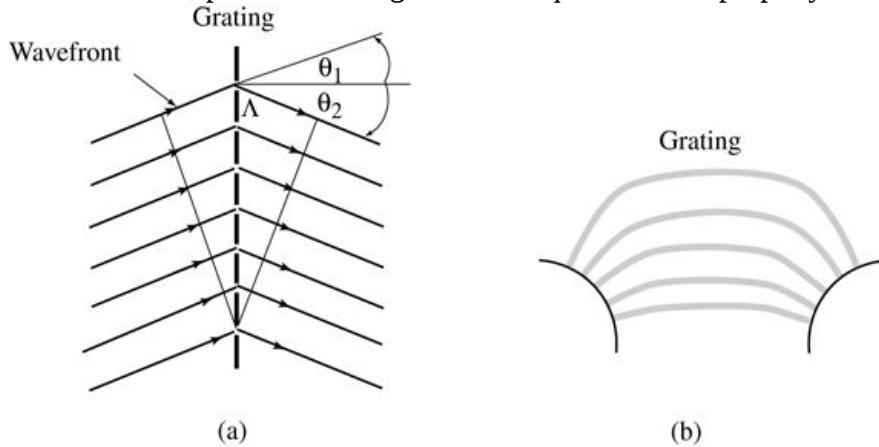


Figure 12.18

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 12.18 Gratings in (a) free space and (b) waveguides.

Illustration a shows a grating represented by a series of identical vertical dashes with equal spaces in between marking the holes. Upward sloping rays are shown to enter the grating from the left side, one in each of the holes. On the right side the rays slope downward, reflecting the rays on the

left. On the left side, the topmost ray, labeled wavefront, is shown extended in a straight line on the other side, making angle theta 1 with the perpendicular at the hole. On the right side, the topmost ray makes an angle theta 2 with the perpendicular. Illustration b shows two arcs on two extremes, their curvatures facing each other. Five thick arching curves connecting the curvatures represent the grating.

Note that, because the holes in the screen are small, many orders of the grating can exist. If the wavelength of the illuminating light changes, the angles of the transmitted orders change in accord with these relationships.

The waveguide grating structure shown on the right of the figure functions in exactly the same way. The waveguide lengths increase by ΔL as we move up the array by one guide, implying that we will be dealing with negative diffraction orders ($m < 0$). Thus $\Delta L = -m\lambda$. $\Delta L = -m\tilde{\lambda}$, where $\tilde{\lambda}$ is the wavelength in the waveguide. A useful correspondence between the free space grating and the waveguide grating is obtained by comparing expressions for the path length difference of the light passing through adjacent grating openings (holes in the case of free space, waveguides in the case of a waveguide grating), taking account of the signs of the angles,

$$\begin{aligned} \Lambda \cdot \sin\theta_2 + \sin\theta_1 &\leftrightarrow \Delta L \\ \Lambda(-\sin\theta_2 + \sin\theta_1) &\leftarrow \Delta L \end{aligned} \quad (12-28)$$

This correspondence can prove useful, provided the fact that the guides have an effective refractive index that is different than that of free space is taken into account.

The Overall System

We turn now to considering the performance of the overall system, as shown in Fig. 12.14. Of particular interest is the change of the output of the system caused by a change of optical wavelength.

To start with the simplest case, an optical wavelength λ_0 is input onto the central input guide of the first star coupler.¹² Suppose that the AWG has been designed such that this wavelength is output on the central output waveguide of the second star coupler. Now we ask what happens to the location of this output when the wavelength is changed from λ_0 to λ_1 . The phase difference $\Delta\phi$ between the output of a waveguide in the grating section and the output of the waveguide just below it is positive and is a function of wavelength given by

$$\Delta\phi(\lambda) = 2\pi n_g \Delta L \lambda,$$

$$\Delta\phi(\lambda) = 2\pi n_g \frac{\Delta L}{\lambda},$$

$$(12-29)$$

where n_g is the effective index in the grating guides. The change in $\Delta\phi$ as the wavelength is changed from λ_0 to λ_1 is

$$\delta\phi = \Delta\phi(\lambda_1) - \Delta\phi(\lambda_0) = 2\pi n_g \Delta L (\lambda_1 - \lambda_0) \approx -2\pi n_g \Delta L \Delta \lambda / 2,$$

$$\delta\phi = \Delta\phi(\lambda_1) - \Delta\phi(\lambda_0) = 2\pi n_g \Delta L \left(\frac{1}{\lambda_1} - \frac{1}{\lambda_0} \right) \approx -2\pi n_g \frac{\Delta L \Delta\lambda}{\lambda_0^2},$$

(12-30)

where it has been assumed that the change in wavelength is small compared with λ_0 and $\Delta\lambda = \lambda_1 - \lambda_0$. $\Delta\lambda$ is positive when $\lambda_1 > \lambda_0$ and negative when $\lambda_1 < \lambda_0$, so $\delta\phi$ is negative for increasing wavelength.

This change of $\Delta\phi$ causes a tilt of the circular wavefront leaving the waveguide grating, and causes a shift in the position of the spot at the output of the second star coupler. The change in position x at the output can be calculated by finding the dispersion of the system,

$$\partial x \partial \lambda = \partial \phi \partial \lambda \cdot \partial x \partial \phi.$$

$$\frac{\partial x}{\partial \lambda} = \frac{\partial \phi}{\partial \lambda} \cdot \frac{\partial x}{\partial \phi}.$$

(12-31)

The first term in this expression can be found from [Eq.\(12-30\)](#),

$$\partial \phi \partial \lambda \approx \delta \phi \Delta \lambda = -2\pi n_g \Delta L \lambda_0^2.$$

$$\frac{\partial \phi}{\partial \lambda} \approx \frac{\delta \phi}{\Delta \lambda} = -2\pi n_g \frac{\Delta L}{\lambda_0^2}.$$

(12-32)

The second term can be found by converting $\delta\phi$ into a wavefront slope change and then calculating the change in x caused by that slope change, with the result

$$\partial x \partial \phi = -\lambda_0 f 2\pi n_s \Lambda,$$

$$\frac{\partial x}{\partial \phi} = -\frac{\lambda_0 f}{2\pi n_s \Lambda},$$

(12-33)

where n_s is the effective index of the slab waveguide in the star coupler. Combining these results, the dispersion of the grating is found to be

$$\partial x \partial \lambda = n_g \Delta L f n_s \lambda_0 \Lambda = -m f n_s \Lambda$$

$$\frac{\partial x}{\partial \lambda} = \frac{n_g \Delta L f}{n_s \lambda_0 \Lambda} = -m \frac{f}{n_s \Lambda}$$

(12-34)

where the last step has assumed that $\Delta L = -m\lambda_0 / n_g$, i.e. the m^{th} negative diffraction order is used. Thus as the wavelength increases, a given output order moves downward on the output surface of the last star coupler, as it should for a negative diffraction order.

Now we consider the resolution of the AWG. Two wavelengths will be barely resolved when the phase difference between the outputs of the top grating guide and the bottom grating guide changes by 2π radians. In that case, with N grating guides, we require a change between adjacent guides $\partial\phi/\partial\lambda \cdot \delta\lambda = 2\pi/N$. Using the previous expression for $\partial\phi/\partial\lambda$, we obtain wavelength resolution $\delta\lambda$ given by

$$\delta\lambda = \lambda_0 N m$$

$$\delta\lambda = \frac{\lambda_0}{N m}$$

(12-35)

and using the previous expression for $\partial x/\partial\lambda$, the spatial resolution is

$$\delta x = \partial x \partial \lambda \cdot \delta \lambda = \lambda_0 f n s N \Lambda.$$

$$\delta x = \left| \frac{\partial x}{\partial \lambda} \right| \cdot \delta \lambda = \frac{\lambda_0 f}{n_s N \Lambda}.$$

(12-36)

In order to achieve this resolution, the output guides from the last star coupler must be no wider than δx .

An additional subject of importance is the *free spectral range* of the overall system. The waveguide grating has many diffraction orders. If again we assume that only the central input waveguide at the input coupler is excited, changes of wavelength will move the output spot through the various waveguides at the system output until the spot passes beyond the last output waveguide (either at the top or the bottom of the output array, depending on whether the wavelength is decreasing or increasing, respectively). At the moment the output spot moves beyond the last output guide, a new spot appears at the waveguide at the opposite end of the output array. One grating order having moved off the output waveguide array, and an adjacent grating order introduces a new spot to take its place, but at the opposite end of the output array. In effect there is “wrap-around” of the output spot due to the multitude of diffraction orders. The total wavelength range that can be accommodated before wrap-around occurs is called the free spectral range of the system.

The free spectral range X of the system can be determined by considering how far the output spot can move before the grating order changes from the m^{th} order to the $(m+1)^{\text{st}}$ order. The grating order changes when $\delta\phi$ of Eq.(12-30) (between adjacent grating waveguides) changes by exactly 2π radians, or when

$$X = \partial x \partial \phi \cdot 2\pi = \lambda_0 f n s \Lambda.$$

$$X = \left| \frac{\partial x}{\partial \phi} \right| \cdot 2\pi = \frac{\lambda_0 f}{n_s \Lambda}$$

(12-37)

An appealing way to state the above results is

$$\delta\lambda\lambda_0=1Nm, \delta x X=1N.$$

$$\frac{\delta\lambda}{\lambda_0} = \frac{1}{Nm},$$

$$\frac{\delta x}{X} = \frac{1}{N}.$$

(12-38)

This concludes our discussion of the general properties of AWGs. We turn now to describing applications of these devices.

12.5.2 Applications of AWGs

There are two primary uses of AWGs. First they have been widely used as multiplexers and demultiplexers for dense WDM signals. Second, they have a rather unique capability to rearrange signals at different wavelengths arriving on different input channels, producing output channels, each with a variety of wavelengths taken from different input channels. We review each of these applications in what follows.

Wavelength Multiplexers and Demultiplexers

[Figure 12.19](#) shows AWGs used as both a demultiplexer and a multiplexer. Considering the demultiplexer first, a single input port carrying equally spaced optical wavelengths $\lambda_1, \lambda_2, \dots, \lambda_N$ arrives at the input to the AWG. The AWG demultiplexes these signals, producing each of the N^N different wavelengths on a separate output port. The wavelength separation between WDM channels must be greater than or equal to the wavelength resolution of the AWG. A minimum of N^N different waveguides is needed in the grating to demultiplex N^N different equally spaced optical wavelengths

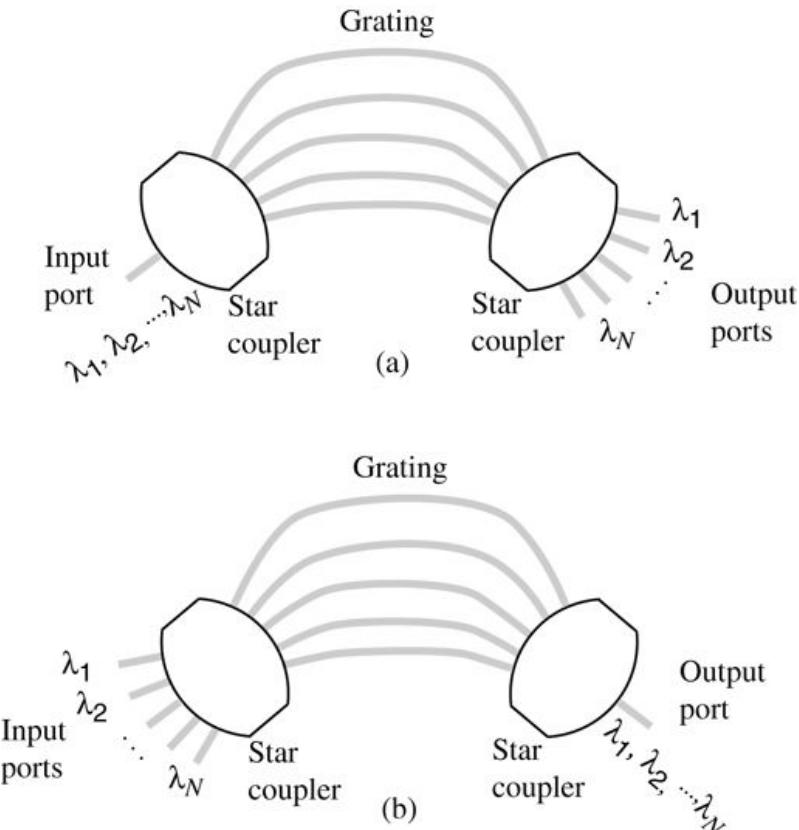


Figure 12.19
Goodman, Introduction to Fourier Optics, 4e,
 © 2017 W. H. Freeman and Company

Figure 12.19 AWG used as (a) demultiplexer and (b) multiplexer

Illustration a shows two star couplers, one on the left and the other on the right. Each coupler is represented by a four sided figure of two identical circular arcs facing outward in opposite directions with their corresponding extremes connected by straight lines. Five thick arching curves connecting the coupler arcs that face each other represent a grating. An input port is connected to the left arc of the left coupler, labeled “lambda 1, lambda 2, and so on till lambda N.” Output ports exit from the right arc of the right coupler; they are labeled “lambda 1, lambda 2, and so on till lambda N.” Illustration b is the same as illustration a but here there are multiple input ports on the left coupler and a single output port on the right coupler.

The multiplexer has a similar geometry, except that there are now N^N different input ports, each carrying only a single optical wavelength, and one output port that carries all of the wavelengths. Again a minimum of N^N different waveguides is needed in the grating to multiplex N^N different equally spaced wavelengths.

A typical commercially available AWG contains up to 40 channels and has an insertion loss for each output channel of 2 to 3 dB. The channel wavelength spacing depends on the design, but can be in the 25 GHz to 200 GHz range.

Wavelength Routers

The wavelength routing function of an AWG can be most easily understood through an analogy with a dispersive free-space imaging system. Consider the imaging system shown in [Fig. 12.20](#).

Two positive lenses, each of focal length f are shown, separated from the grating by distance f along the optical axis of the system. Aside from the grating, this is a “ $4f$ ” imaging system, which will produce an image of the object with unity magnification and with image inversion. The presence of the grating offsets the angle of the second half of the system and makes the system dispersive. Note that each of the lenses (together with the free space before and after) is the analog of a star coupler, and the grating is the analog of the waveguide grating in an AWG.

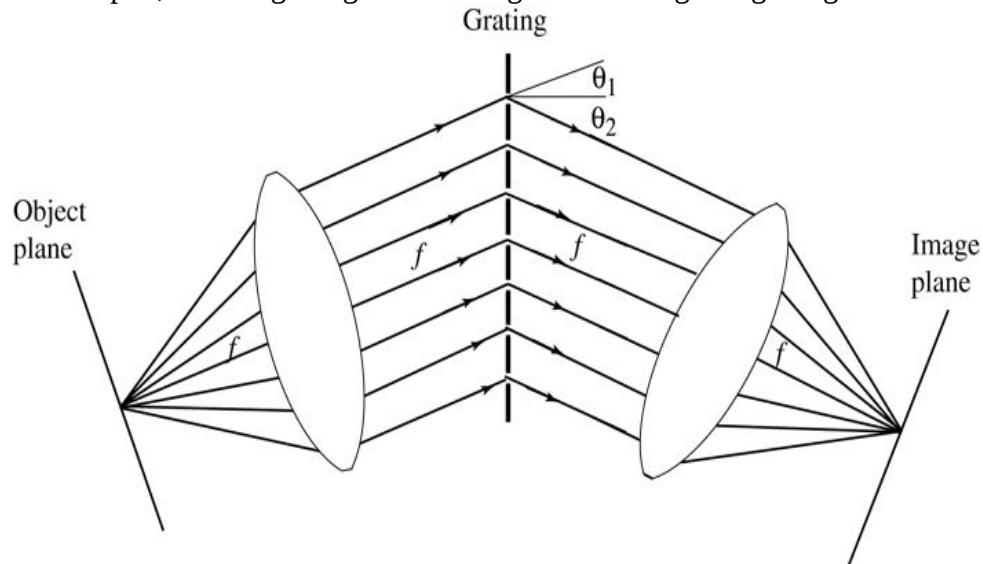


Figure 12.20
Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 12.20 Imaging analog to an AWG.

The illustration shows a grating represented by a series of identical vertical dashes with equal spaces in between marking the holes. On the left extreme is an object plane represented by a downward sloping line. From the center of the object plane, rightward rays diverge to reach a similarly oriented lens, on the other side of which parallel, upward sloping rays extend to the holes in the grating. The rays exit the grating and follow a downward sloping path to an upward sloping lens that converges the rays to the center of the image plane behind, represented by an upward sloping line. On the right side, the topmost ray on the left side extends in a straight line, making angle θ_1 with the perpendicular at the hole. The topmost ray on the right side makes an angle θ_2 with the perpendicular. The distances between the object plane and the left lens; the left lens and the grating; the grating and the right lens; and the right lens and the image plane are each marked f .

Keeping this imaging analog in mind, we now consider an AWG under several different input conditions. [Figure 12.21\(a\)](#) shows an AWG with one wavelength input on its central input port and all other input ports inactive. We label this wavelength λ_0 to indicate that it is the wavelength that the system was designed to image directly from central input port to central output port. Now consider the situation depicted in [Fig. 12.21\(b\)](#). The same wavelength λ_0 has been moved up by one input port. The result, by simple imaging laws, is that the output moves down one port. In

this way, imaging rules can be applied to determine where wavelength λ_0 will appear at the output when applied to any input port.

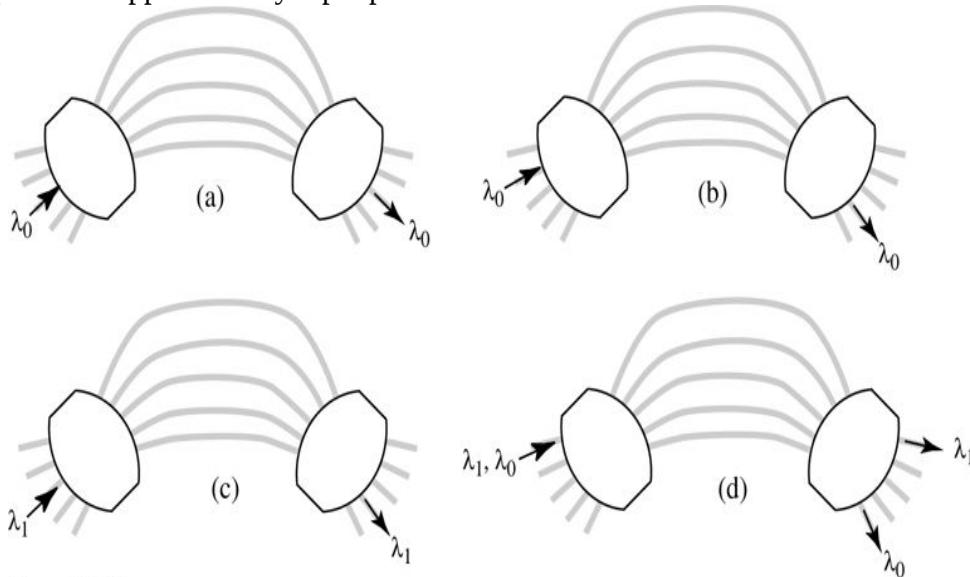


Figure 12.21

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 12.21 Illustration of AWG wavelength routing properties. (a) Imaging λ_0 from central port to central port. (b) Imaging λ_0 from an off-center input port to an output port at the inverted image position. (c) Imaging wavelength $\lambda_1 = \lambda_0 + \delta\lambda$ from the central input port to an offset image port. (d) Imaging λ_1 from a top input port to a “wrapped-around” output port.

Each of the four illustrations shows two star couplers, one on the left and the other on the right. Each coupler is represented by a four sided figure of two identical circular arcs facing outward in opposite directions with their corresponding extremes connected by straight lines. Five thick arching curves connect the coupler arcs that face each other. Five input ports enter the left of the left coupler and five output ports exit the right side of the right coupler. In illustration a, an entry arrow marks the third input port and an exit arrow marks the third output port, both labeled lambda 0. In illustration b, an entry arrow marks the second from top input port and an exit arrow marks the second from bottom output port, both labeled lambda 0. In illustration c, an entry arrow marks the third input port and an exit arrow marks the second from bottom output port, both labeled lambda 1. In illustration d, an entry arrow, labeled “lambda 1, lambda 0,” marks the topmost input port and an exit arrow, labeled “lambda 1,” marks the topmost output port, while another exit arrow, labeled “lambda 0,” marks the lowermost output port.

Now consider the case shown in Fig. 12.21(c). In this case we have increased the wavelength from λ_0 to $\lambda_1 = \lambda_0 + \delta\lambda$, where $\delta\lambda$ is the wavelength change required to move the output down by one output port ($\delta\lambda$ is the wavelength resolution of the AWG). Thus at wavelength λ_1 the output is shifted downward one output port. If the input with wavelength λ_1 is moved to another input port, the output will always appear shifted downward by one output port from the position predicted by simple imaging, unless the downward shift brings the expected

output port past the end of the output array, in which case λ_1 will be found at the top of the output array, as shown in Fig. 12.21(d). In effect, a change of wavelength from λ_0 induces a cyclic shift of the output across the output ports, with the size of the shift being the number of increments $\delta\lambda$ in the change. Lengthening the wavelength causes downward shifts, and shortening the wavelength causes upward shifts, given that we are using negative diffraction orders in the AWG.

We are now prepared to understand the most general wavelength-routing use of an AWG. With reference to Fig. 12.22, consider a wavelength numbering system that labels wavelengths according to both the input port they are applied to and the wavelength offset from the wavelength that images without circular rotation, which we have labeled λ_0 . We number the input ports from 0 to $N-1$ from the bottom input to the top input. We label the wavelengths with two subscripts, the first being the input port it is applied to, and the second being the wavelength offset from λ_0 in units of $\delta\lambda$, the resolution of the AWG. Thus a wavelength labeled $\lambda_{n,m}$ is wavelength $\lambda_0 + m\delta\lambda$ ($m = 0, 1, \dots, N-1$) appearing on input port n ($n = 0, 1, \dots, N-1$).

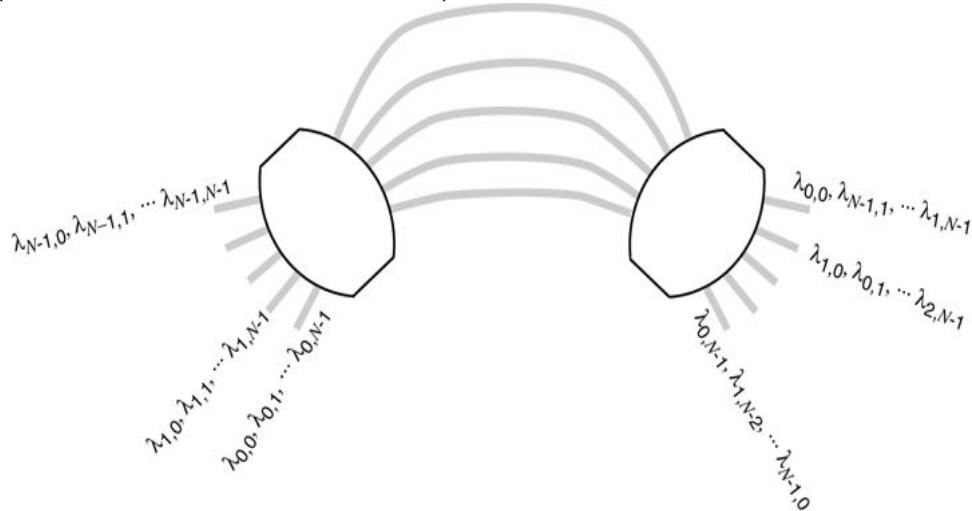


Figure 12.22

Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure 12.22 Wavelength routing properties of an AWG. The first wavelength subscript corresponds to the input port number, and the second subscript corresponds to the wavelength offset from λ_0 in increments of $\delta\lambda$.

The illustration shows two star couplers, one on the left and the other on the right. Each coupler is represented by a four sided figure of two identical circular arcs facing outward in opposite directions with their corresponding extremes connected by straight lines. Five thick arching curves connect the coupler arcs that face each other. Five input ports enter the left of the left coupler and five output ports exit the right side of the right coupler.

The topmost input port is labeled thus: lambda subscript (N minus 1, 0), lambda subscript (N minus 1, 1), ... lambda subscript (N minus 1, N minus 1). The input port second from bottom is labeled thus: lambda subscript (1, 0), lambda subscript (1, 1), ... lambda subscript (1, N minus 1). The lowermost input port is labeled thus: lambda subscript (0, 0), lambda subscript (0, 1), ...

lambda subscript $(0, N \text{ minus } 1)$. The topmost output port is labeled thus: lambda subscript $(0, 0)$, lambda subscript $(N \text{ minus } 1, 1)$, ... lambda subscript $(1, N \text{ minus } 1)$. The second from top output port is labeled thus: lambda subscript $(1, 0)$, lambda subscript $(0, 1)$, ... lambda subscript $(2, N \text{ minus } 1)$. The lowermost output port is labeled thus: lambda subscript $(0, N \text{ minus } 1)$, lambda subscript $(1, N \text{ minus } 2)$, ... lambda subscript $(N \text{ minus } 1, 0)$.

Now suppose that each input port has a full complement of wavelengths, i.e. each carries all N^N possible wavelengths. The situation is as shown at the input in [Fig. 12.22](#). The routing functions described above on a wavelength-by-wavelength basis can now be applied to the full set of inputs. The wavelength labels on the right of the AWG show the wavelengths that appear on each output port. Note that each output port contains all the wavelengths, but a different wavelength from each input port. Thus the AWG acts as a complex re-arranger of wavelengths, filling each output port with a full complement of wavelengths, each wavelength having come from a different input port. Such a routing function is a kind of wavelength exchanger, which can be useful in interconnecting network branches in complex network topologies.

Problems - Chapter 12

1. 12-1. With reference to (12-9), prove that this equation can be equivalently written as (12-10).
2. 12-2. Find the effective length and the effective number of grating lines of a fiber phase reflection grating when the free-space Bragg wavelength is $\lambda_B = 1550 \text{ nm}$, $n_1 = 1.45$, and for $\delta n = 10^{-4}$, 10^{-3} and 10^{-2} .
3. 12-3. In this problem, we consider the effect of a delay (or advance) by τ_0 of the signal waveform with respect to the reference pulse in spectral holography. Suppose that the illumination spot of the signal waveform beam on the input grating during recording of the hologram is square and covers exactly N grating lines in the x direction on the input grating. The reference pulse is assumed to cover the same number of lines as well. Using the result from Prob. 4-13, as well as the arguments in this chapter about the time window in spectral holography, show that the maximum τ_0 that will produce a significant fringe pattern in the hologram is approximately equal to $1/\delta\nu$, where $\delta\nu$ is the frequency resolution of the grating. Show that τ_0 is equivalently limited to N optical cycles of the center frequency of the time signals.
4. 12-4. Find an expression for $v\nu$ in terms of x analogous to that in (12-18) when θ_2 is not small enough to make the approximation $\sin\theta_2 \approx \theta_2$.
5. 12-5. Considering the results depicted in Fig. 12.21, and defining $\lambda_m = \lambda_0 + m\delta\lambda$, predict where the input wavelengths shown in Fig. P12.5 will appear on the output ports of the AWG. Assume the system has been designed such that for wavelength λ_0 , the N input ports are mapped (with inversion) to the top N output ports.

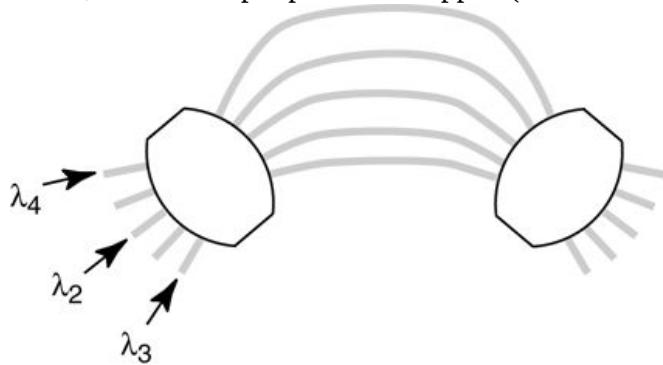


Figure P12.5

Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure P12.5

The illustration shows two star couplers, one on the left and the other on the right. Each coupler is represented by a four sided figure of two identical circular arcs facing outward in opposite directions with their corresponding extremes connected by straight lines. Five thick arching curves connect the coupler arcs that face each other. Five input ports enter the left of the left coupler and five output ports exit the right side of the right coupler. In the top down direction, the first, third, and the fifth input ports are labeled lambda 4, lambda 2, and lambda 3, respectively.

6. 12-6. The input star coupler of an AWG has N^N input waveguides and $2N^{2N}$ output waveguides. The output star coupler has $2N^{2N}$ input waveguides and $2N^{2N}$ output waveguides. There are $2N^{2N}$ waveguides in the grating section. All star-coupler waveguide widths and spacings are the same, so the waveguides at the output of the second star coupler occupy twice as much of the star-coupler surface as do the waveguides at the input of the first star coupler.

1. Specify the wavelength resolution $\delta\lambda \delta\lambda$, the spatial resolution $\delta x \delta x$, and the free spectral range $X X$ for this AWG in terms of any of the parameters N^N , $m m$, ns^{n_s} , ng^{n_g} , $\lambda_0 \lambda_0$, $f f$, and $\Lambda \Lambda$ that may be necessary.
2. Draw a picture similar to [Fig. 12.22](#) for this AWG, showing where various wavelengths from various input ports end up at the output.

A Delta Functions and Fourier Transform Theorems

A.1 Delta Functions

The one-dimensional Dirac delta function, widely used in systems analysis, is in fact not a function at all, but rather is a more general entity, often called a “functional” or a “distribution.” While a function is an entity that maps a *number* (the argument of the function) into a *number* (the value of the function), a functional maps a *function* into a *number*. A simple example of a functional is a definite integral, for example

$$\int_{-\infty}^{\infty} h(\xi) d\xi,$$

$$\int_{-\infty}^{\infty} h(\xi) d\xi,$$

which maps any given function $h(\xi)$ into the value of its area.

In this spirit, the defining characteristic of the delta function¹ is its so-called “sifting” property under integration, namely

$$\int_{-\infty}^{\infty} \delta(\xi - b) h(\xi) d\xi = h(b) \text{ if } b \text{ is a point of continuity of } h \\ 12h(b+) + h(b-) \text{ if } b \text{ is a point of discontinuity of } h.$$

$$\int_{-\infty}^{\infty} \delta(\xi - b) h(\xi) d\xi = \begin{cases} h(b) & b \text{ a point of continuity of } h \\ \frac{1}{2} [h(b^+) + h(b^-)] & b \text{ a point of discontinuity of } h. \end{cases}$$

(A-1)

In this equation, the symbols $h(b^+)$ and $h(b^-)$ represent the limiting values of h as its argument approaches the discontinuity from above and from below, respectively. The mapping of a function h into the values on the right of the above equation defines the functional we call the delta function. The integral is a convenient representation for this mapping, but should not be interpreted literally as an integral. It can be viewed as the limit of a set of integrals, i.e.

$$\int_{-\infty}^{\infty} \delta(\xi - b) h(\xi) d\xi \equiv \lim_{N \rightarrow \infty} \int_{-\infty}^{\infty} g_N(\xi - b) h(\xi) d\xi,$$

(A-2)

where g_N is a sequence of functions that in the limit $N \rightarrow \infty$ exhibit the required sifting property. Such functions must all have unit area and must *in some sense* become narrower and narrower as N grows large.

It has become fairly common practice in the engineering literature to represent the delta function by the limit of the sequence of functions g_N in [Eq.\(A-2\)](#), i.e. to write

$$\delta(x) = \lim_{N \rightarrow \infty} g_N(x).$$

$$\delta(x) = \lim_{N \rightarrow \infty} g_N(x).$$

(A-3)

Although this representation is not strictly correct, the limit of the sequence of integrals being the proper representation, nonetheless we use it here with the understanding that it really means what is expressed in [Eq.\(A-2\)](#). Thus we write, for example that

$$\delta(x) = \lim_{N \rightarrow \infty} N \exp(-N^2 \pi x^2) \quad \delta(x) = \lim_{N \rightarrow \infty} N \operatorname{rect}(Nx) \quad \delta(x) = \lim_{N \rightarrow \infty} N \operatorname{sinc}(Nx).$$

$$\begin{aligned}\delta(x) &= \lim_{N \rightarrow \infty} N \exp(-N^2 \pi x^2) \\ \delta(x) &= \lim_{N \rightarrow \infty} N \operatorname{rect}(Nx) \\ \delta(x) &= \lim_{N \rightarrow \infty} N \operatorname{sinc}(Nx).\end{aligned}$$

A plot of the last of the above functions shows that $N \operatorname{sinc}(Nx)$ does not become a very narrow pulse as $N \rightarrow \infty$, but rather it retains a finite spread and develops ever more rapid oscillations everywhere except at the origin. Such oscillations in the limit ensure that under an integral sign the value of h at the location of the center of the function sequence will be sifted out. Thus it is not necessary that the functions g_N vanish in the limit everywhere except the origin. A somewhat more bizarre example is the function sequence

$$g_N(x) = N e^{j\pi/4} \exp(j\pi(Nx)^2),$$

$$g_N(x) = N e^{j\pi/4} \exp[j\pi(Nx)^2],$$

each member of which has unit area and magnitude N everywhere, but still exhibits a sifting property in the limit.

While the δ function is used in electrical systems analysis to represent a sharp, intense pulse of current or voltage, the analogous concept in optics is a point source of light, or a *spatial* pulse of unit volume. The definition of the δ function in two dimensions is a simple extension of the one-dimensional case, although there is even greater latitude in the possible functional forms of the pulse sequences used. Many possible definitions use separable pulse sequences, e.g.

$$\begin{aligned}\delta(x,y) &= \lim_{N \rightarrow \infty} N^2 \exp[-N^2 \pi(x^2+y^2)] \\ \delta(x,y) &= \lim_{N \rightarrow \infty} N^2 \operatorname{rect}(Nx) \operatorname{rect}(Ny) \\ \delta(x,y) &= \lim_{N \rightarrow \infty} N^2 \operatorname{sinc}(Nx) \operatorname{sinc}(Ny).\end{aligned}$$

$$\begin{aligned}\delta(x, y) &= \lim_{N \rightarrow \infty} N^2 \exp[-N^2 \pi(x^2 + y^2)] \\ \delta(x, y) &= \lim_{N \rightarrow \infty} N^2 \operatorname{rect}(Nx) \operatorname{rect}(Ny) \\ \delta(x, y) &= \lim_{N \rightarrow \infty} N^2 \operatorname{sinc}(Nx) \operatorname{sinc}(Ny).\end{aligned}$$

Other possible definitions use circularly symmetric functions, e.g.

$$\delta(x,y)=\lim_{N \rightarrow \infty} N^2 \pi \operatorname{circ}(Nx^2+y^2) \quad (\text{A-4}) \quad \delta(x,y)=\lim_{N \rightarrow \infty} N J_1(2\pi Nx^2+y^2)x^2+y^2.$$

$$\begin{aligned}\delta(x, y) &= \lim_{N \rightarrow \infty} \frac{N^2}{\pi} \operatorname{circ}\left(N\sqrt{x^2 + y^2}\right) \quad (\text{A-4}) \\ \delta(x, y) &= \lim_{N \rightarrow \infty} N \frac{J_1(2\pi N\sqrt{x^2 + y^2})}{\sqrt{x^2 + y^2}}.\end{aligned}$$

(A-5)

In some applications one definition may be more convenient than others, and the definition best suited for the problem can be chosen.

A property of all two-dimensional delta functions that can be easily proved (see [Prob. 2.1\(a\)](#)) is

$$\delta(ax, by)=1|ab|\delta(x,y),$$

$$\delta(ax, by)=\frac{1}{|ab|} \delta(x, y),$$

(A-6)

which describes how such entities behave under scaling of coordinates. Again this statement has meaning only under integral signs.

A.2 Derivation of Fourier Transform Theorems

In this section, brief proofs of basic Fourier transform theorems are presented. For more complete derivations, see [37], [275], and [146].

- 1. Linearity theorem.** $\mathcal{F}\{\alpha g + \beta h\} = \alpha \mathcal{F}\{g\} + \beta \mathcal{F}\{h\}$

Proof: This theorem follows directly from the linearity of the integrals that define the Fourier transform.

- 2. Similarity theorem.** If $\mathcal{F}\{g(x,y)\} = G(f_X, f_Y)$, then

$$\mathcal{F}\{g(ax, by)\} = 1/|ab| G(f_X/a, f_Y/b).$$

$$\mathcal{F}\{g(ax, by)\} = \frac{1}{|ab|} G\left(\frac{f_X}{a}, \frac{f_Y}{b}\right).$$

Proof:

$$\mathcal{F}\{g(ax, by)\} = \iint_{-\infty}^{\infty} g(ax, by) e^{-j2\pi(f_X x + f_Y y)} dx dy = \iint_{-\infty}^{\infty} g(ax, by) e^{-j2\pi(f_X ax + f_Y by)} dx dy = 1/|ab| G(f_X/a, f_Y/b).$$

$$\begin{aligned} \mathcal{F}\{g(ax, by)\} &= \iint_{-\infty}^{\infty} g(ax, by) e^{-j2\pi(f_X x + f_Y y)} dx dy \\ &= \iint_{-\infty}^{\infty} g(ax, by) e^{-j2\pi\left(\frac{f_X}{a}ax + \frac{f_Y}{b}by\right)} \frac{d(ax)}{|a|} \frac{d(by)}{|b|} dx dy \\ &= \frac{1}{|ab|} G\left(\frac{f_X}{a}, \frac{f_Y}{b}\right). \end{aligned}$$

- 3. Shift theorem.** If $\mathcal{F}\{g(x,y)\} = G(f_X, f_Y)$, then

$$\mathcal{F}\{g(x-a, y-b)\} = G(f_X, f_Y) e^{-j2\pi(f_X a + f_Y b)}.$$

$$\mathcal{F}\{g(x-a, y-b)\} = G(f_X, f_Y) \exp[-j2\pi(f_X a + f_Y b)].$$

Proof:

$$\mathcal{F}\{g(x-a, y-b)\} = \iint_{-\infty}^{\infty} g(x-a, y-b) e^{-j2\pi(f_X x + f_Y y)} dx dy = \iint_{-\infty}^{\infty} g(x', y') e^{-j2\pi(f_X(x'+a) + f_Y(y'+b))} dx' dy' = G(f_X, f_Y) \exp[-j2\pi(f_X a + f_Y b)].$$

$$\begin{aligned} \mathcal{F}\{g(x-a, y-b)\} &= \iint_{-\infty}^{\infty} g(x-a, y-b) e^{-j2\pi(f_X x + f_Y y)} dx dy \\ &= \iint_{-\infty}^{\infty} g(x', y') e^{-j2\pi(f_X(x'+a) + f_Y(y'+b))} dx' dy' \\ &= G(f_X, f_Y) \exp[-j2\pi(f_X a + f_Y b)]. \end{aligned}$$

- 4. Rayleigh's (Parseval's) theorem.** If $\mathcal{F}\{g(x,y)\} = G(f_X, f_Y)$, then

$$\iint_{-\infty}^{\infty} g(x, y)^2 dx dy = \iint_{-\infty}^{\infty} G(f_X, f_Y)^2 df_X df_Y.$$

$$\int_{-\infty}^{\infty} \int |g(x, y)|^2 dx dy = \int_{-\infty}^{\infty} \int |G(f_X, f_Y)|^2 df_X df_Y.$$

Proof:

$$\begin{aligned} & \iint_{-\infty}^{\infty} |g(x, y)|^2 dx dy = \iint_{-\infty}^{\infty} g(x, y) g^*(x, y) dx dy = \iint_{-\infty}^{\infty} dx dy \iint_{-\infty}^{\infty} \\ & d\xi d\eta G(\xi, \eta) \exp[j2\pi(x\xi + y\eta)] \times \iint_{-\infty}^{\infty} d\alpha d\beta G^*(\alpha, \beta) \exp[-j2\pi(x\alpha + y\beta)] = \iint_{-\infty}^{\infty} \\ & d\xi d\eta G(\xi, \eta) \iint_{-\infty}^{\infty} d\alpha d\beta G^*(\alpha, \beta) \times \iint_{-\infty}^{\infty} \exp\{j2\pi[x(\xi - \alpha) + y(\eta - \beta)]\} dx dy = \iint_{-\infty}^{\infty} \\ & d\xi d\eta G(\xi, \eta) \iint_{-\infty}^{\infty} d\alpha d\beta G^*(\alpha, \beta) \delta(\xi - \alpha, \eta - \beta) = \iint_{-\infty}^{\infty} |G(\xi, \eta)|^2 d\xi d\eta. \end{aligned}$$

$$\begin{aligned} \int_{-\infty}^{\infty} \int |g(x, y)|^2 dx dy &= \int_{-\infty}^{\infty} \int g(x, y) g^*(x, y) dx dy \\ &= \int_{-\infty}^{\infty} dx dy \left[\int_{-\infty}^{\infty} d\xi d\eta G(\xi, \eta) \exp[j2\pi(x\xi + y\eta)] \right] \\ &\quad \times \left[\int_{-\infty}^{\infty} d\alpha d\beta G^*(\alpha, \beta) \exp[-j2\pi(x\alpha + y\beta)] \right] \\ &= \int_{-\infty}^{\infty} d\xi d\eta G(\xi, \eta) \int_{-\infty}^{\infty} d\alpha d\beta G^*(\alpha, \beta) \\ &\quad \times \left[\int_{-\infty}^{\infty} \exp\{j2\pi[x(\xi - \alpha) + y(\eta - \beta)]\} dx dy \right] \\ &= \int_{-\infty}^{\infty} d\xi d\eta G(\xi, \eta) \int_{-\infty}^{\infty} d\alpha d\beta G^*(\alpha, \beta) \delta(\xi - \alpha, \eta - \beta) \\ &= \int_{-\infty}^{\infty} \int |G(\xi, \eta)|^2 d\xi d\eta. \end{aligned}$$

5. Convolution theorem. If $\mathcal{F}\{g(x, y)\} = G(f_X, f_Y)$ and $\mathcal{F}\{h(x, y)\} = H(f_X, f_Y)$, then

$$\mathcal{F} \iint_{-\infty}^{\infty} g(\xi, \eta) h(x - \xi, y - \eta) d\xi d\eta = G(f_X, f_Y) H(f_X, f_Y).$$

$$\mathcal{F} \left\{ \iint_{-\infty}^{\infty} g(\xi, \eta) h(x - \xi, y - \eta) d\xi d\eta \right\} = G(f_X, f_Y) H(f_X, f_Y).$$

Proof:

$$\mathcal{F} \iint_{-\infty}^{\infty} g(\xi, \eta) h(x - \xi, y - \eta) d\xi d\eta = \iint_{-\infty}^{\infty} g(\xi, \eta) \mathcal{F} h(x - \xi, y - \eta) d\xi d\eta = \iint_{-\infty}^{\infty} g(\xi, \eta) \exp[-j2\pi(f_X \xi + f_Y \eta)] d\xi d\eta H(f_X, f_Y) = G(f_X, f_Y) H(f_X, f_Y).$$

$$\begin{aligned} \mathcal{F} \left\{ \iint_{-\infty}^{\infty} g(\xi, \eta) h(x - \xi, y - \eta) d\xi d\eta \right\} &= \iint_{-\infty}^{\infty} g(\xi, \eta) \mathcal{F} \{h(x - \xi, y - \eta)\} d\xi d\eta \\ &= \iint_{-\infty}^{\infty} g(\xi, \eta) \exp[-j2\pi(f_X \xi + f_Y \eta)] d\xi d\eta H(f_X, f_Y) = G(f_X, f_Y) H(f_X, f_Y). \end{aligned}$$

6. Autocorrelation theorem. If $\mathcal{F}\{g(x, y)\} = G(f_X, f_Y)$, then

$$\mathcal{F} \int_{-\infty}^{\infty} g(\xi, \eta) g^*(\xi-x, \eta-y) d\xi d\eta = G(f_X, f_Y) 2.$$

$$\mathcal{F} \left\{ \int_{-\infty}^{\infty} \int g(\xi, \eta) g^*(\xi-x, \eta-y) d\xi d\eta \right\} = |G(f_X, f_Y)|^2.$$

Proof:

$$\begin{aligned} \mathcal{F} \int_{-\infty}^{\infty} g(\xi, \eta) g^*(\xi-x, \eta-y) d\xi d\eta &= \mathcal{F} \int_{-\infty}^{\infty} g(\xi+x, \eta+y) g^*(\xi', \eta') d\xi' d\eta' = \int_{-\infty}^{\infty} \int \\ d\xi' d\eta' g^*(\xi', \eta') \mathcal{F}\{g(\xi'+x, \eta'+y)\} &= \int_{-\infty}^{\infty} \int d\xi' d\eta' g^*(\xi', \eta') \exp[j2\pi(f_X \xi' + f_Y \eta')] G(f_X, f_Y) = G^*(f_X, f_Y) G(f_X, f_Y) = |G(f_X, f_Y)|^2. \end{aligned}$$

$$\begin{aligned} &\mathcal{F} \left\{ \int_{-\infty}^{\infty} \int g(\xi, \eta) g^*(\xi-x, \eta-y) d\xi d\eta \right\} \\ &= \mathcal{F} \left\{ \int_{-\infty}^{\infty} \int g(\xi'+x, \eta'+y) g^*(\xi', \eta') d\xi' d\eta' \right\} \\ &= \int_{-\infty}^{\infty} \int d\xi' d\eta' g^*(\xi', \eta') \mathcal{F}\{g(\xi'+x, \eta'+y)\} \\ &= \int_{-\infty}^{\infty} \int d\xi' d\eta' g^*(\xi', \eta') \exp[j2\pi(f_X \xi' + f_Y \eta')] G(f_X, f_Y) \\ &= G^*(f_X, f_Y) G(f_X, f_Y) = |G(f_X, f_Y)|^2. \end{aligned}$$

7. Rotation Theorem. $\mathcal{F}\{g(r, \theta - \theta_0)\} = G(\rho, \phi - \theta_0)$. If the Fourier transform is performed in a polar coordinate system, we have (cf. [Eq.\(2-30\)](#))

$$\begin{aligned} \mathcal{F}\{g(r, \theta)\} &= G(\rho, \phi) = \int_0^\infty \int_0^{2\pi} r g(r, \theta) \exp[-j2\pi r \rho (\cos \theta \cos \phi \\ &\quad + \sin \theta \sin \phi)] d\theta dr = \int_0^\infty \int_0^{2\pi} r g(r, \theta) \exp[-j2\pi r \rho \cos(\theta - \phi)] d\theta dr. \end{aligned}$$

$$\begin{aligned} \mathcal{F}\{g(r, \theta)\} = G(\rho, \phi) &= \int_0^\infty \int_0^{2\pi} r g(r, \theta) \exp[-j2\pi r \rho (\cos \theta \cos \phi \\ &\quad + \sin \theta \sin \phi)] d\theta dr \\ &= \int_0^\infty \int_0^{2\pi} r g(r, \theta) \exp[-j2\pi r \rho \cos(\theta - \phi)] d\theta dr. \end{aligned}$$

Here (r, θ) are polar coordinates in the space domain, while (ρ, ϕ) are polar coordinates in the frequency domain. It follows that if we rotate the function g by angle θ_0 about the origin, we obtain a new Fourier transform $G(\rho, \phi)$ given by

$$G(\rho, \phi) = \int_0^\infty \int_0^{2\pi} r g(r, \theta - \theta_0) \exp[-j2\pi r \rho \cos(\theta - \phi)] d\theta dr.$$

$$\tilde{G}(\rho, \phi) = \int_0^\infty \int_0^{2\pi} r g(r, \theta') \exp[-j2\pi r \rho \cos(\theta' + \theta_0 - \phi)] d\theta' dr.$$

Now change variables of integration from θ' to $\theta' = \theta - \theta_0$, with the limits of integration remaining over any period of 2π radians, yielding

$$G(\rho, \phi) = \int_0^\infty \int_0^{2\pi} r g(r, \theta') \exp[-j2\pi r \rho \cos(\theta' + \theta_0 - \phi)] d\theta' dr = G(\rho, \phi - \theta_0).$$

$$\begin{aligned} \tilde{G}(\rho, \phi) &= \int_0^\infty \int_0^{2\pi} r g(r, \theta') \exp[-j2\pi r \rho \cos(\theta' + \theta_0 - \phi)] d\theta' dr \\ &= G(\rho, \phi - \theta_0). \end{aligned}$$

Thus rotation of $g(r, \theta)$ by angle θ_0 has rotated its Fourier transform $G(\rho, \phi)$ by θ_0 .

8. Shear theorem. Assume a horizontal shear. Then to prove $\mathcal{F}\{g(x+by, y)\} = G(f_X, f_Y - bf_X)$. The Fourier transform of $g(x+by, y)$ can be written

$$\mathcal{F}\{g(x+by, y)\} = \int_{-\infty}^{\infty} dy e^{-j2\pi f_Y y} \int_{-\infty}^{\infty} dx g(x+by, y) e^{-j2\pi f_X x} = \int_{-\infty}^{\infty} dy e^{-j2\pi f_Y y} G(f_X, y) e^{j2\pi b y f_X},$$

$$\begin{aligned} \mathcal{F}\{g(x+by, y)\} &= \int_{-\infty}^{\infty} dy e^{-j2\pi f_Y y} \int_{-\infty}^{\infty} dx g(x+by, y) e^{-j2\pi f_X x} \\ &= \int_{-\infty}^{\infty} dy e^{-j2\pi f_Y y} \tilde{G}(f_X, y) e^{j2\pi b y f_X}, \end{aligned}$$

where the shift theorem has been used, and \tilde{G} represents the partial transform of g , for which the integral has been performed only with respect to x . It follows that

$$\mathcal{F}\{g(x+by, y)\} = \int_{-\infty}^{\infty} dy \tilde{G}(f_X, y) e^{-j2\pi(f_Y - bf_X)y} = G(f_X, f_Y - bf_X).$$

$$\mathcal{F}\{g(x+by, y)\} = \int_{-\infty}^{\infty} dy \tilde{G}(f_X, y) e^{-j2\pi(f_Y - bf_X)y} = G(f_X, f_Y - bf_X).$$

The proof for a vertical shear follows the same outline of steps.

9. Fourier integral theorem. At each point of continuity of g ,

$$\mathcal{F}^{-1}\{g(x, y)\} = \mathcal{F}^{-1}\mathcal{F}\{g(x, y)\} = g(x, y).$$

$$\mathcal{F}^{-1}\{g(x, y)\} = \mathcal{F}^{-1}\mathcal{F}\{g(x, y)\} = g(x, y).$$

At each point of discontinuity of g , the two successive transformations yield the angular average of the value of g in a small neighborhood of that point.

Proof: Let the function $g_R(x, y)$ be defined by

$$g_R(x, y) = \iint_A R G(f_X, f_Y) \exp[j2\pi(f_X x + f_Y y)] df_X df_Y,$$

$$g_R(x, y) = \iint_{A_R} G(f_X, f_Y) \exp[j2\pi(f_X x + f_Y y)] df_X df_Y,$$

where A_R is a circle of radius R , centered at the origin of the (f_X, f_Y) plane. To prove the theorem, it suffices to show that, at each point of continuity of g ,

$$\lim_{R \rightarrow \infty} g_R(x, y) = g(x, y),$$

$$\lim_{R \rightarrow \infty} g_R(x, y) = g(x, y),$$

and that, at each point of discontinuity of g^S ,

$$\lim_{R \rightarrow \infty} g_R(x, y) = 12\pi \int_0^{2\pi} g_o(\theta) d\theta,$$

$$\lim_{R \rightarrow \infty} g_R(x, y) = \frac{1}{2\pi} \int_0^{2\pi} g_o(\theta) d\theta,$$

where $g_o(\theta)$ is the angular dependence of g^S in a small neighborhood about the point in question.

Some initial straightforward manipulation can be performed as follows:

$$g_R(x, y) = \iint_{A_R} A R \iint_{-\infty}^{\infty} d\xi d\eta g(\xi, \eta) e^{-j2\pi(f_X \xi + f_Y \eta)} e^{j2\pi(f_X x + f_Y y)} df_X df_Y = \iint_{-\infty}^{\infty} d\xi d\eta g(\xi, \eta) \iint_{A_R} df_X df_Y e^{j2\pi(f_X(x - \xi) + f_Y(y - \eta))}.$$

$$\begin{aligned} g_R(x, y) &= \iint_{A_R} \left\{ \int_{-\infty}^{\infty} d\xi d\eta g(\xi, \eta) e^{-j2\pi(f_X \xi + f_Y \eta)} \right\} e^{j2\pi(f_X(x - \xi) + f_Y(y - \eta))} df_X df_Y \\ &= \int_{-\infty}^{\infty} d\xi d\eta g(\xi, \eta) \iint_{A_R} df_X df_Y \exp\{j2\pi[f_X(x - \xi) + f_Y(y - \eta)]\}. \end{aligned}$$

Noting that

$$\iint_{A_R} df_X df_Y \exp\{j2\pi[f_X(x - \xi) + f_Y(y - \eta)]\} = RJ1(2\pi Rr),$$

$$\int_{A_R} df_X df_Y \exp\{j2\pi[f_X(x - \xi) + f_Y(y - \eta)]\} = R \left[\frac{J_1(2\pi Rr)}{r} \right],$$

where $r = \sqrt{(x - \xi)^2 + (y - \eta)^2}$, we have

$$g_R(x, y) = \iint_{-\infty}^{\infty} d\xi d\eta g(\xi, \eta) RJ1(2\pi Rr).$$

$$g_R(x, y) = \int_{-\infty}^{\infty} d\xi d\eta g(\xi, \eta) R \left[\frac{J_1(2\pi Rr)}{r} \right].$$

Suppose initially that (x, y) is a point of continuity of g^S . Then

$$\lim_{R \rightarrow \infty} g_R(x, y) = \iint_{-\infty}^{\infty} d\xi d\eta g(\xi, \eta) \lim_{R \rightarrow \infty} RJ1(2\pi Rr) = \iint_{-\infty}^{\infty} d\xi d\eta g(\xi, \eta) \delta(x - \xi, y - \eta) = g(x, y),$$

$$\begin{aligned} \lim_{R \rightarrow \infty} g_R(x, y) &= \int_{-\infty}^{\infty} d\xi d\eta g(\xi, \eta) \lim_{R \rightarrow \infty} R \left[\frac{J_1(2\pi Rr)}{r} \right] \\ &= \int_{-\infty}^{\infty} d\xi d\eta g(\xi, \eta) \delta(x - \xi, y - \eta) = g(x, y), \end{aligned}$$

where (A-5) has been used in the second step. Thus the first part of the theorem has been proved.

Consider next a point of discontinuity of g^S . Without loss of generality that point can be taken to be the origin. Thus we write

$$gR(0,0) = \iint_{-\infty}^{\infty} d\xi d\eta g(\xi, \eta) R J_1(2\pi Rr),$$

$$g_R(0,0) = \int_{-\infty}^{\infty} d\xi d\eta g(\xi, \eta) R \left[\frac{J_1(2\pi Rr)}{r} \right],$$

where $r = \sqrt{\xi^2 + \eta^2}$. But for sufficiently large R , the quantity following $g(\xi, \eta)$ in the integrand acts as a delta function at the origin (cf. [Eq. \(A-5\)](#)). In addition, in this small neighborhood the function g^S depends (approximately) only on the angle θ about that point, and therefore

$$gR(0,0) \approx \int_0^{2\pi} g_o(\theta) d\theta \int_0^{\infty} r R \left[\frac{J_1(2\pi Rr)}{r} \right] dr$$

$$g_R(0,0) \approx \int_0^{2\pi} g_o(\theta) d\theta \int_0^{\infty} r R \left[\frac{J_1(2\pi Rr)}{r} \right] dr$$

where $g_o(\theta)$ represents the θ dependence of g^S about the origin. Finally, noting that

$$\int_0^{\infty} r R \left[\frac{J_1(2\pi Rr)}{r} \right] dr = 12\pi,$$

$$\int_0^{\infty} r R \left[\frac{J_1(2\pi Rr)}{r} \right] dr = \frac{1}{2\pi},$$

we conclude that

$$\lim_{R \rightarrow \infty} gR(0,0) = 12\pi \int_0^{2\pi} g_o(\theta) d\theta,$$

$$\lim_{R \rightarrow \infty} g_R(0,0) = \frac{1}{2\pi} \int_0^{2\pi} g_o(\theta) d\theta,$$

and the proof is thus complete.

B Introduction to Paraxial Geometrical Optics

B.1 The Domain of Geometrical Optics

If the wavelength of light is imagined to become vanishingly small, we enter a domain in which the concepts of geometrical optics suffice to analyze optical systems. While the actual wavelength of light is always finite, nonetheless provided all variations or changes of the amplitude and phase of a wavefield take place on spatial scales that are very large compared with a wavelength, the predictions of geometrical optics will be accurate. Examples of situations for which geometrical optics does not yield accurate predictions occur when we insert a sharp edge or a sharply defined aperture in a beam of light, or when we change the phase of a wave by a significant fraction of 2π radians over spatial scales that are comparable with a wavelength.

Thus if we imagine a periodic phase grating for which a “smooth” change of phase by 2π radians takes place only over a distance of many wavelengths, the predictions of geometrical optics for the amplitude distribution behind the grating will be reasonably accurate. On the other hand, if the changes of 2π radians take place in only a few wavelengths, or take place very abruptly, then diffraction effects can not be ignored, and a full wave-optics (or “physical-optics”) treatment of the problem is needed.

This appendix is not a complete introduction to the subject of geometrical optics. Rather, we have selected several topics that will help the reader better understand the relationship between geometrical optics and physical optics. In addition, several geometrical concepts that are needed in formulating the physical-optics description of imaging and spatial filtering systems are introduced.

The Concept of a Ray

Consider a monochromatic disturbance traveling in a medium with refractive index that varies slowly on the scale of an optical wavelength. Such a disturbance can be described by an amplitude and phase distribution

$$U(\vec{r}) = A(\vec{r}) \exp[jk_o S(\vec{r})],$$
$$U(\vec{r}) = A(\vec{r}) \exp[jk_o S(\vec{r})],$$

(B-1)

where $A(\vec{r})$ is the amplitude and $k_o S(\vec{r})$ is the phase of the wave. Here k_o is the free-space wavenumber $2\pi/\lambda_o$; the refractive index n of the medium is contained in the definition of S . $S(\vec{r})$ is called the *Eikonal* function. We follow the argument presented in [305] (p. 52) in finding the equation that must be satisfied by the Eikonal function.

Surfaces defined by

$$S(r \rightarrow) = \text{constant}$$

$$S(\vec{r}) = \text{constant}$$

are called *wavefronts* of the disturbance. The direction of power flow and the direction of the wave vector $\vec{k} \rightarrow$ are both normal to the wavefronts at each point $r \rightarrow \vec{r}$ in an isotropic medium. A *ray* is defined as a trajectory or a path through space that starts at any particular point on a wavefront and moves through space with the wave, always remaining perpendicular to the wavefront at every point on the trajectory. Thus a ray traces out the path of power flow in an isotropic medium. Substitution of (B-1) in the Helmholtz equation of Eq.(3-13) yields the following equation that must be satisfied by both $A(r \rightarrow)$ and $S(r \rightarrow)$:

$$k_0^2 n^2 - |\nabla S|^2 A + \nabla^2 A + j k_0 [2 \nabla S \cdot \nabla A + A \nabla^2 S] = 0.$$

$$k_0^2 [n^2 - |\nabla S|^2] A + \nabla^2 A + j k_0 [2 \nabla S \cdot \nabla A + A \nabla^2 S] = 0.$$

The real and imaginary parts of this equation must vanish independently. For the real part to vanish, we require

$$\nabla^2 S = n^2 + \left(\frac{\lambda_0}{2\pi} \right)^2 \frac{\nabla^2 A}{A}.$$

(B-2)

Using the artifice of allowing the wavelength to approach zero to recover the geometrical-optics limit of this equation, the last term is seen to vanish, leaving the so-called *Eikonal equation*, which is perhaps the most fundamental description of the behavior of light under the approximations of geometrical optics,

$$|\nabla S(r \rightarrow)|^2 = n^2(r \rightarrow).$$

$$|\nabla S(\vec{r})|^2 = n^2(\vec{r}).$$

(B-3)

This equation serves to define the wavefront S^S . Once the wavefronts are known, the trajectories defining rays can be determined.

Rays and Local Spatial Frequency

Consider a monochromatic wave propagating in three-dimensional space defined by an (x, y, z) coordinate system, with propagation being in the positive $z \rightarrow$ direction. At each point on a plane of constant $z \rightarrow$, there is a well-defined direction of the ray through that point, a direction that coincides with the direction of the wave vector $\vec{k} \rightarrow$ at that point.

We have seen previously that an arbitrary distribution of a complex field across a plane can be decomposed by means of a Fourier transform into a collection of plane-wave components traveling in different directions. Each such plane-wave component has a unique wave vector with direction cosines (α, β, γ) defined in [Fig. 3.9](#), and can be regarded as one spatial frequency associated with the wave.

The spatial frequencies defined through the Fourier decomposition exist everywhere in space and cannot be regarded as being localized. However, for complex functions with a phase that does not vary too rapidly, the concept of a local spatial frequency can be introduced, as was done in

[Section 2.2.1](#). The definitions of the local spatial frequencies $(f_X(\ell), f_Y(\ell))$ given there can also be viewed as defining the *local* direction cosines $(\alpha_l, \beta_l, \gamma_l)$ of the wavefront through the relations

$$\alpha_l = \lambda f_X(\ell) \quad \beta_l = \lambda f_Y(\ell) \quad \gamma_l = 1 - \alpha_l^2 - \beta_l^2.$$

$$\alpha_l = \lambda f_X^{(\ell)} \quad \beta_l = \lambda f_Y^{(\ell)} \quad \gamma_l = \sqrt{1 - \alpha_l^2 - \beta_l^2}.$$

(B-4)

These local direction cosines are in fact the direction cosines of the ray through the (x, y) plane at each point. This leads us to the following important observation:

The description of the local spatial frequencies of a wavefront is identical with the description of that wavefront in terms of the rays of geometrical optics. Ray direction cosines are found from local spatial frequencies simply by multiplication by the wavelength.

B.2 Refraction, Snell's Law, and the Paraxial Approximation

Rays traveling in a medium with constant index of refraction always travel in straight lines, as can be derived from the Eikonal equation. However, when the wave travels through a medium having an index of refraction that changes in space (i.e. an inhomogeneous medium), the ray directions will undergo changes that depend on the changes of refractive index. When the changes of refractive index are gradual, the ray trajectories will be smoothly changing curves in space. Such bending of the rays is called *refraction*.

However, when a wave encounters an abrupt boundary between two media having different refractive indices, the ray directions are changed suddenly as they pass through the interface. The angles of incidence θ_1 and refraction θ_2 , as shown in Fig. 3.1, are related by *Snell's law*,

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

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

(B-5)

where n_1 and n_2 are the refractive indices of the first and second media, respectively. In the problems of interest here, the changes of refractive index, as encountered, for example, on passage through a lens, will always be abrupt, so Snell's law will form the basis for our analyses.

A further simplifying approximation can be made if we restrict attention to rays that are traveling close to the optical axis and at small angles to that axis, the geometrical optics version of the *paraxial approximation*. In such a case, Snell's law reduces to a simple linear relationship between the angle of incidence and the angle of refraction,

$$n_1 \theta_1 = n_2 \theta_2,$$

$$n_1 \theta_1 = n_2 \theta_2,$$

(B-6)

and in addition the cosines of these angles can be replaced by unity.

The product $\hat{\theta} = n\theta$ of the refractive index n and an angle θ within that medium is called a *reduced angle*. Thus the paraxial version of Snell's law states that the reduced angle remains constant as light passes through a sharp interface between media of different refractive indices,

$$\hat{\theta}_1 = \hat{\theta}_2.$$

$$\hat{\theta}_1 = \hat{\theta}_2.$$

(B-7)

B.3 The Ray-Transfer Matrix

Under paraxial conditions, the properties of rays in optical systems can be treated with an elegant matrix formalism. Additional references for this material are [305], [196], and [316]. To apply this methodology, it is necessary to consider only *meridional rays*, which are rays traveling in paths that are completely contained in a single plane containing the z axis. We call the transverse axis in this plane the y axis, and therefore the plane of interest is the (y, z) plane.

[Figure B.1](#) shows the typical kind of ray propagation problem that must be solved in order to understand the effects of an optical system. On the left, at axial coordinate z_1 , is an input plane of an optical system, and on the right, at axial coordinate z_2 , is an output plane. A ray with transverse coordinate y_1 enters the system at angle θ_1 , and the same ray, now with transverse coordinate y_2 , leaves the system with angle θ_2 . The goal is to determine the position y_2 and angle θ_2 of the output ray for every possible y_1 and θ_1 associated with an input ray.

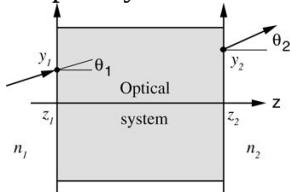


Figure B.1
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure B.1 Input and output of an optical system.

The illustration shows a rightward pointing horizontal axis z cutting through two parallel upward pointing rays, representing the input plane (on the left) and the output plane (on the right). The rays are separated by a shaded area labeled “Optical system,” whose top and bottom limits are marked by lines parallel to the z axis, one at each extreme of the vertical rays. The refractive indices n_1 , to the left of the input plane, and n_2 , to the right of the output plane, are marked. The z axis intersects the input plane at z_1 and the output plane at z_2 . Above the z axis on the input plane, a rightward upward sloping ray is incident at y_1 such that its extension makes an angle θ_1 with the perpendicular at the point of incidence from within the optical system. Near the upper end of the output plane, a rightward upward sloping ray exits at y_2 , making an angle θ_2 with the perpendicular at the point of exit from the outside.

Under the paraxial condition, the relationships between (y_2, θ_2) and (y_1, θ_1) are linear and can be written explicitly as

$$y_2 = Ay_1 + B\theta_1, \quad \hat{\theta}_2 = Cy_1 + D\hat{\theta}_1,$$

$$\begin{aligned} y_2 &= Ay_1 + B\hat{\theta}_1 \\ \hat{\theta}_2 &= Cy_1 + D\hat{\theta}_1, \end{aligned}$$

where for reasons that will become evident, we use reduced angles rather than just angles. The above equation can be expressed more compactly in matrix notation,

$$y_2 \theta^2 = ABCD y_1 \theta^1.$$

$$\begin{bmatrix} y_2 \\ \hat{\theta}_2 \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} y_1 \\ \hat{\theta}_1 \end{bmatrix}.$$

(B-8)

The matrix

$$M = ABCD$$

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

(B-27)

is called the *ray-transfer matrix* or the *ABCD* matrix.

The ray-transfer matrix has an interesting interpretation in terms of local spatial frequencies. In the (y, z) plane under paraxial conditions, the reduced ray angle θ^1 with respect to the z axis is related to local spatial frequency $f(\ell)$ through

$$f(\ell) = \theta^1 = \lambda \theta^1.$$

$$f(\ell) = \frac{\theta}{\lambda} = \frac{\hat{\theta}}{\lambda_0}.$$

(B-9)

Therefore the ray-transfer matrix can be regarded as specifying a transformation between the spatial distribution of local spatial frequency at the input and the corresponding distribution at the output.

Elementary Ray-Transfer Matrices

Certain simple structures are commonly encountered in ray tracing problems. Here we specify the ray-transfer matrices for the most important of these structures. They are all illustrated in [Fig. B.2](#).

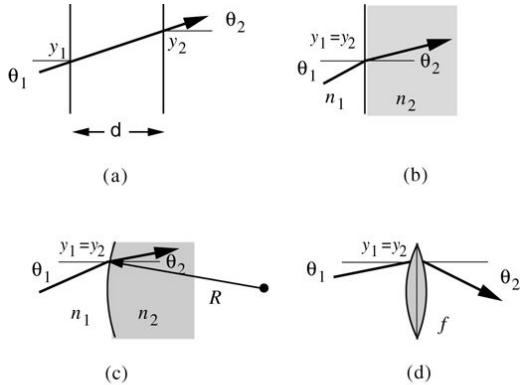


Figure B.2
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure B.2 Elementary structures for ray-transfer matrix calculations. (a) Free space, (b) a planar interface, (c) a spherical interface, and (d) a thin lens.

Illustration a shows two vertical parallel lines separated by a distance of d . An upward sloping rightward arrow intersects the two lines at points y_1 and y_2 , making angles θ_1 and θ_2 , respectively, with the perpendiculars dropped at the intersections from the outside.

Illustration b shows a vertical line separating two media of indexes n_1 and n_2 , n_1 being on the left side. It is intersected by a horizontal line at $y_1 = y_2$, where an upward sloping ray from the bottom left corner makes angle θ_1 with the horizontal line in the clockwise direction. Once the ray enters the index n_2 side, in the top right corner, it bends slightly downward, thus making a smaller angle, θ_2 , with the horizontal line, in the clockwise direction.

Illustration c shows a vertically oriented curve bowed out to the left separating two media of indexes n_1 and n_2 , n_1 being on the left side. The curve is intersected by a horizontal line at $y_1 = y_2$, where an upward sloping ray from the bottom left corner makes angle θ_1 with the horizontal line in the clockwise direction. Once the ray enters the index n_2 side, in the top right corner, it bends slightly downward, thus making a smaller angle, θ_2 , with the horizontal line in the clockwise direction.

Illustration d shows a biconvex lens. A horizontal line passes through a point near the upper end of the lens at $y_1 = y_2$, where an upward sloping ray enters from the bottom left corner, making an angle θ_1 with the horizontal line in the clockwise direction. The ray exits the lens where the horizontal line cuts the right side curvature and bends downward, in the bottom right corner, thus making a larger angle, θ_2 , with the horizontal line in the counterclockwise direction.

1. **Propagation through free space of index n** . Geometrical rays travel in straight lines in a medium with constant refractive index. Therefore the effect of propagation through free space is to translate the location of the ray in proportion to the angle at which it travels and to leave the angle of the ray unchanged. The ray-transfer matrix describing propagation over distance d is therefore

$$M = 1/d \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$M = \begin{bmatrix} 1 & d/n \\ 0 & 1 \end{bmatrix}$$

(B-10)

2. Refraction at a planar interface. At a planar interface the position of the ray is unchanged but the angle of the ray is transformed according to Snell's law; the reduced angle remains unchanged. Therefore the ray-transfer matrix for a planar interface between a medium of refractive index n_1 and a medium of refractive index n_2 is

$$M=1001.$$

$$M = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

(B-11)

3. Refraction at a spherical interface. At a spherical interface between an initial medium with refractive index n_1 and a final medium with refractive index n_2 , the position of a ray is again not changed, but the angle is changed. However, at a point on the interface at distance y from the optical axis, the normal to the interface is not parallel to the optical axis, but rather is inclined with respect to the optical axis by angle

$$\psi = \text{arcsin} y R \approx y R,$$

$$\psi = \arcsin \frac{y}{R} \approx \frac{y}{R},$$

where R is the radius of the spherical surface. Therefore if θ_1 and θ_2 are measured with respect to the optical axis, Snell's law at transverse coordinate y becomes

$$n_1 \theta_1 + n_1 y R = n_2 \theta_2 + n_2 y R,$$

$$n_1 \theta_1 + n_1 \frac{y}{R} = n_2 \theta_2 + n_2 \frac{y}{R},$$

or, using reduced angles,

$$\hat{\theta}_1 + n_1 \frac{y}{R} = \hat{\theta}_2 + n_2 \frac{y}{R}.$$

$$\hat{\theta}_1 + n_1 \frac{y}{R} = \hat{\theta}_2 + n_2 \frac{y}{R}.$$

Solving for $\hat{\theta}_2$ yields

$$\hat{\theta}_2 = \hat{\theta}_1 + \frac{n_1 - n_2}{R} y.$$

$$\hat{\theta}_2 = \hat{\theta}_1 + \frac{n_1 - n_2}{R} y.$$

$$\hat{\theta}_2 = \hat{\theta}_1 + \frac{n_1 - n_2}{R} y.$$

The ray-transfer matrix for a spherical interface can now be written as

$$M=10n1-n2R1.$$

$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ \frac{n_1 - n_2}{R} & 1 \end{bmatrix}$$

(B-12)

Note that a positive value for R signifies a convex surface encountered from left to right, while a negative value for R signifies a concave surface.

- 4. Passage through a thin lens.** A thin lens (index n_2 embedded in a medium of index n_1) can be treated by cascading two spherical interfaces. The roles of n_1 and n_2 are interchanged for the two surfaces. Representing the ray-transfer matrices of the surfaces on the left and the right by M_1 and M_2 , respectively, the ray-transfer matrix for the sequence of two surfaces is

$$M = M_2 M_1 = 10n_2 - n_1 R_2 110n_1 - n_2 R_1 = 10 - (n_2 - n_1) R_1 - 1 R_2.$$

$$\begin{aligned} M &= M_2 M_1 \\ &= \begin{bmatrix} 1 & 0 \\ \frac{n_2 - n_1}{R_2} & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \frac{n_1 - n_2}{R_1} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -(n_2 - n_1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right) & 1 \end{bmatrix} \end{aligned}$$

(B-32)

We define the focal length of the lens by

$$1/f = n_2 - n_1 R_1 - 1 R_2,$$

$$\frac{1}{f} = \frac{n_2 - n_1}{n_1} \left(\frac{1}{R_1} - \frac{1}{R_2} \right),$$

(B-13)

in which case the ray-transfer matrix for a thin lens becomes

$$M = 10 - n_1 f 1.$$

$$\mathbf{M} = \begin{bmatrix} 1 & 0 \\ -\frac{n_1}{f} & 1 \end{bmatrix}$$

(B-14)

The most useful elementary ray-transfer matrices have now been presented. Propagation through a system consisting of regions of free space separated by thin lenses can be treated with these matrices. Note that the ray-transfer matrices should be applied in the sequence in which the structures are encountered. If light propagates first through a structure with ray-transfer matrix M_1

\mathbf{M}_1 , then through a structure with ray-transfer matrix $\mathbf{M}_2 \mathbf{M}_2$, etc., with a final structure having ray-transfer matrix $\mathbf{M}_n \mathbf{M}_n$, then the overall ray-transfer matrix for the entire system is

$$\mathbf{M} = \mathbf{M}_n \cdots \mathbf{M}_2 \mathbf{M}_1.$$

$$\mathbf{M} = \mathbf{M}_n \cdots \mathbf{M}_2 \mathbf{M}_1.$$

(B-15)

We note also that, because we have chosen to use *reduced* angles, rather than the angles themselves in the definition of the ray-transfer matrix, all of the elementary matrices presented have a determinant that is unity.

B.4 Conjugate Planes, Focal Planes, and Principal Planes

There exist certain planes within an optical system that play important conceptual and practical roles. In this section we explain the three most important of these types of planes.

Conjugate Planes

Two planes within an optical system are said to be *conjugate planes* if the intensity distribution across one plane is an image (generally magnified or demagnified) of the intensity distribution across the other plane. Likewise, two points are said to be conjugate points if one is the image of the other.

The properties that must be satisfied by the ray-transfer matrix between two conjugate planes can be deduced by considering the relation between two conjugate points y_1 and y_2 . The position of the point y_2 that is conjugate to y_1 should be independent of the reduced angle θ_1 of a ray through y_1 , implying that the matrix element B should be zero. The position y_2 should be related to the position y_1 only through the transverse magnification m_t , which is the scale factor between coordinates in the two planes. We conclude that the matrix element A must equal m_t . In addition, the angles of the rays passing through y_2 will generally be magnified or demagnified with respect to the angles of the same rays passing through y_1 . The magnification for reduced angles is represented by m_α , and we conclude that the matrix element D must satisfy $D=m_\alpha$. There is no general restriction on the matrix element C , so the ray-transfer matrix between conjugate planes takes the general form

$$M=mt0Cm\alpha.$$

$$M = \begin{bmatrix} m_t & 0 \\ C & m_\alpha \end{bmatrix}.$$

Recalling that angles and positions are conjugate Fourier variables, the scaling theorem of Fourier analysis implies that the transverse magnification and the angular magnification must be related in a reciprocal fashion. The magnifications m_t and m_α are in fact related by

$$mtm\alpha=1.$$

$$m_t m_\alpha = 1.$$

(B-16)

Thus the form of the ray-transfer matrix for conjugate planes is

$$M=mt0Cmt^{-1}.$$

$$M = \begin{bmatrix} m_t & 0 \\ C & m_t^{-1} \end{bmatrix}.$$

Note that both m_t and m_α can be positive or negative (signifying image inversion), but they must be of the same sign.

The paraxial relation (B-16) has a more general nonparaxial form, known as the *sine condition*, which states that for conjugate points y_1 and y_2 the following equation must be satisfied:

$$\begin{aligned} n_1 y_1 \sin\theta_1 &= n_2 y_2 \sin\theta_2. \\ n_1 y_1 \sin\theta_1 &= n_2 y_2 \sin\theta_2. \end{aligned} \tag{B-17}$$

Focal Planes

Consider a parallel bundle of rays traveling parallel to the optical axis and entering a lens. Whether that lens is thick or thin, for paraxial rays there will exist a point on the optical axis toward which that ray bundle will converge (positive lens) or from which it will appear to diverge (negative lens). See Fig. B.3 for an illustration. Considering a positive lens for the moment, the point behind the lens at which this originally parallel ray bundle crosses in a focused point is called the *rear focal point* or the *second focal point* of the lens. A plane constructed through that point perpendicular to the optical axis is called the *rear focal plane* or the *second focal plane*. It has the property that a paraxial parallel bundle of rays traveling into the lens at any angle with respect to the optical axis will be brought to a focus at a point in the focal plane that depends on the initial angle of the bundle.

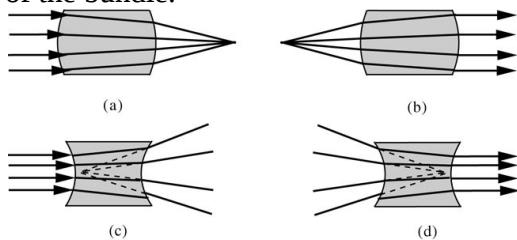


Figure B.3
Goodman, *Introduction to Fourier Optics*, 4e, © 2017 W. H. Freeman and Company

Figure B.3 Definition of focal points. (a) Rear focal point of a positive lens, (b) front focal point of a positive lens, (c) front focal point of a negative lens, and (d) rear focal point of a negative lens.

Illustration a shows four rightward parallel rays entering a biconvex lens and tending to converge as they move rightward within the lens. On exiting the right side curvature, the rays converge to a single point.

Illustration b shows four rightward rays diverging from a single point and entering a biconvex lens and continuing to diverge within the lens. On exiting the right side curvature, the rays run mutually parallel and horizontal.

Illustration c shows four rightward rays running parallel towards a biconcave lens, though which they diverge slightly. On exiting the right side curvature, the rays diverge significantly more. The divergent rays outside the lens, when extended leftward by dotted lines, meet at a point inside the lens close to the left curvature.

Illustration d shows four rightward rays tending to converge as they enter a biconcave lens, inside which they tend to converge more. On exiting the right side curvature, the rays run mutually

parallel and horizontal. The initial convergent rays outside the lens, when extended rightward by dotted lines, meet at a point inside the lens close to the right curvature.

In a similar fashion, consider a point source on the optical axis in front of a positive lens, thick or thin. The particular point in front of the lens for which the diverging bundle of rays is made to emerge as a parallel bundle traveling parallel to the optical axis behind the lens is called the *front focal point* (or the *first focal point*) of the lens. A plane erected through the front focal point normal to the optical axis is called the *front focal plane* (or the *first focal plane*) of the lens.

For a negative lens, the roles of the front and rear focal points and planes are reversed. The front focal point is now the point from which a bundle of rays, originally parallel to the optical axis, appears to be diverging when viewed from the exit side of the lens. The rear focal point is defined by the point of convergence of an incident bundle of rays that emerges parallel or collimated after passage through the lens.

The mapping from the front focal plane to the rear focal plane is one that maps angles into positions, and positions into angles. If f^f is the focal length of the lens, then the ray-transfer matrix between focal planes takes the form

$$M = 0fn_1 - n_1 f_0,$$

$$M = \begin{bmatrix} 0 & \frac{f}{n_1} \\ -\frac{n_1}{f} & 0 \end{bmatrix},$$

as can be readily verified by multiplying together three matrices representing propagation over distance f^f , passage through a thin lens with focal length f^f , and propagation over a second distance f^f .

Principal Planes

By the definition of a thin lens, a ray incident at input coordinate y_1 exits that lens at the same coordinate $y_2 = y_1$. For a thick lens this simple idealization is no longer valid. A ray entering the first spherical surface at coordinate y_1 will in general leave the second spherical surface at a different coordinate $y_2 \neq y_1$, as can be seen in [Fig. B.3](#).

Much of the simplicity of a thin lens can be retained for a thick lens by introducing the concept of *principal planes*. Principal planes are planes where the focusing power of the lens can be imagined to be concentrated.

To find the first principal plane of a lens, trace a ray from the front focal point to the first lens surface, as shown in [Fig. B.4](#). By definition of the focal point, that ray will exit the second surface of the lens parallel to the optical axis, i.e. in a *collimated beam*. If we project the incident ray forwards and the exiting ray backwards into the lens, retaining their original angles, they will intersect at a point. A plane through this point normal to the optical axis defines the *first principal plane*. For this geometry it is possible to imagine that all the refraction associated with the lens takes place in this principal plane.

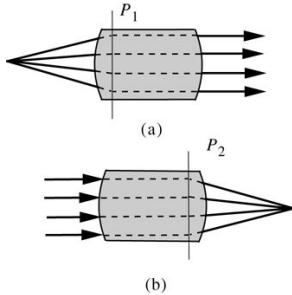


Figure B.4
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure B.4 Definitions of principal planes. (a) First principal plane P_1 , (b) second principal plane P_2 .

Illustration a shows four rightward rays diverging from a point and entering a biconvex lens, within which dotted lines mark their horizontal, mutually parallel path, which is continued on their exit through the right side curvature. A vertical line on the left extreme of the lens where the rays begin to go parallel is labeled P_1 .

Illustration b shows four rightward mutually parallel rays entering a biconvex lens, within which dotted lines mark their continued horizontal path. On exiting the right side curvature, the rays converge to a point. A vertical line on the right extreme of the lens where the rays begin to converge is labeled P_2 .

In the most general case, different rays diverging from the front focal point might define different planes, which would be an indication that the principal plane is not a plane at all, but rather is a curved surface. Such can be the case for lenses with very large aperture or for special lenses such as wide-angle lenses, but for the lenses of interest to us in this book the principal planes are indeed flat to an excellent approximation.

The second principal plane is found by starting with a ray that is parallel to the optical axis, and tracing it through the rear focal point of the lens. The extension of the incident ray and the exiting ray intersect in a point, which in turn defines the *second principal plane* of the lens, again normal to the optical axis. For this geometry it is possible to imagine that all of the power of the lens is concentrated in the second principal plane.

For more general geometries, ray bending can be imagined to take place in both of the principal planes. As will be seen shortly, the two planes are in fact conjugate to one another with unit magnification. A ray incident at particular transverse coordinates on the first principal plane will exit from the second principal plane at those same coordinates, but in general with a change of angle.

In general, the first and second principal planes are separate planes. However, the definition of a thin lens implies that for such a lens the distinguishing characteristic is that the first and second principal planes coincide, and all the focusing power can be imagined to be concentrated in a single plane.

The relationship between the principal planes can be more fully understood if we derive the ray-transfer matrix that holds for propagation between the two principal planes. The derivation is based on the two geometries already introduced, namely that of a point source at the front focal point that yields a collimated ray bundle leaving the second principal plane, and that of a collimated bundle incident on the first principal plane that yields a ray bundle converging from the second principal plane toward a focus at the rear focal point. Considering the case of collimated input light passing through the rear focal point, we find that the matrix element A must be

unity, and the matrix element C must be $-n_1/f - n_1/f$. Consideration of the case of input rays diverging from the front focal point shows that $B=0$ and $D=1$. Thus the ray-transfer matrix for the passage between principal planes is

$$M=10-n_1f.$$

$$M = \begin{bmatrix} 1 & 0 \\ -\frac{n_1}{f} & 1 \end{bmatrix}.$$

This matrix is identical with the ray-transfer matrix describing passage through a thin lens. Thus by constructing the principal planes, and by tracing rays only up to the first principal plane and away from the second principal plane, we are able to treat a complex lens system as if it were a simpler thin lens. Note that the ray-transfer matrix above implies that the two principal planes are conjugate to one another, and the magnification between them is unity.

The *focal length* of a lens is by definition the distance of a principal plane from the corresponding focal point that was used in its definition. Assuming that the refractive indices of the media in front of and behind the lens are the same, the distance of the front focal plane from the first principal plane is identical with the distance of the rear focal point from the second principal plane. That is, the two focal lengths of the lens are the same. Note that for some lenses the second principal plane may lie to the left of the first principal plane. Such an occurrence does not change the definition of the focal length. It can also be shown that the distances z_1 and z_2 in the lens law

$$1/z_1 + 1/z_2 = 1/f$$

$$\frac{1}{z_1} + \frac{1}{z_2} = \frac{1}{f}$$

are measured from the first and second principal planes. These various relations are illustrated in [Fig. B.5](#).

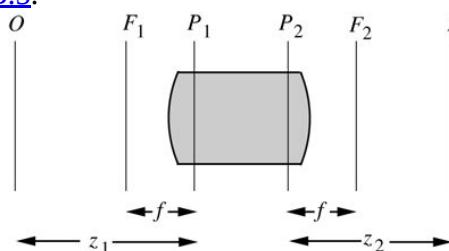


Figure B.5
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure B.5 Relations between principal planes, focal lengths, and object/image distances.

The illustration shows a biconvex lens with two vertical lines cutting through it, P_1 close to the left curvature and P_2 close to the right curvature. To the left of the lens, at distance of f from P_1 , is vertical line F_1 . To the right of the lens, at distance of f from P_2 , is vertical line F_2 . To the left of the lens, at distance of z_1 from P_1 , is vertical line O . To the left of the lens, at distance of z_2 from P_2 , is vertical line I ; z_1 is greater than f , z_2 is greater than f .

B.5 Entrance and Exit Pupils

Until now, we have not considered the effects of pupils (i.e. finite apertures) in optical systems. Apertures, of course, give rise to diffraction effects. The concepts of entrance and exit apertures are of great importance in calculations of the effects of diffraction on optical systems.

A system of lenses may contain several or many different apertures, but one such aperture always provides the severest limitation to the extent of the optical wavefront captured at the input of the system, and to the extent of the optical wavefront leaving the system. That aperture may lie deep within the system of lenses, but the single aperture that most severely restricts the bundle of rays passing through the system is in effect the aperture that limits the extent of the wavefront at both the input and at the output.

The *entrance pupil* of the optical system is defined as the *image of the most severely limiting aperture*, when viewed from the object space, looking through any optical elements that may precede the physical aperture. The *exit pupil* of the system is also defined as the image of the physical aperture, but this time looking from the image space through any optical elements that may lie between that aperture and the image plane.

[Figure B.6](#) illustrates the entrance and exit pupils for a very simple system consisting of a single lens, for three cases: a limiting pupil (1) in the plane of the lens, (2) following the lens, and (3) preceding the lens. In the first case, the entrance and exit apertures coincide with the real physical aperture in the plane of the lens. In the second case, the exit pupil coincides with the physical pupil (which is assumed to limit the angle of the bundle of rays more severely than does the lens aperture), and the entrance pupil is a virtual image of the physical aperture, lying to the right of the lens. In the third case, the entrance pupil is the real physical aperture lying to the left of the lens. In this case, the exit pupil is a virtual image of the physical aperture, lying in a plane to the left of the lens.

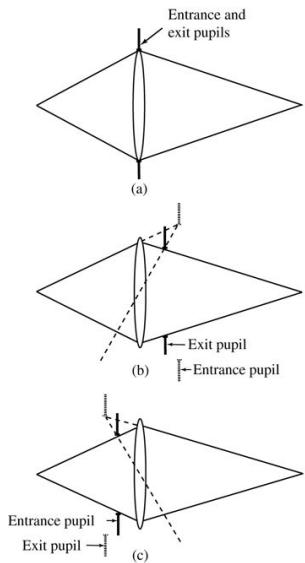


Figure B.6
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure B.6 Entrance and exit pupils. (a) Entrance and exit pupils coincide with the physical pupil, (b) the exit pupil coincides with the physical pupil, and (c) the entrance pupil coincides with the physical pupil.

All three illustrations show a vertical biconvex lens. From a point on the left, rays diverge to the lens; beyond the lens, the rays converge to a point on the right that is at a greater distance from the lens than the point on the left. In illustration a, a vertical marker is shown at both extremes of the lens, top and bottom, labeled "Entrance and exit pupils." In illustration b, a vertical marker is located near the lens on the ray connecting the top extreme of the lens to the point of convergence on the right. The marker is reflected on the ray at the other extreme below. It is labeled "Exit pupil." The ray on the left side connecting to the top extreme of the lens is extended upward in a dotted line, which is intersected by another dotted line connecting the center of the lens and the exit pupil's location on the ray. At the point of intersection a different vertical marker is located and also reflected on the opposite side below. It is labeled "Entrance pupil." In illustration c, a vertical marker is located near the lens on the ray connecting the top extreme of the lens to the point of divergence on the left. The marker is reflected on the ray at the other extreme below. It is labeled "Entrance pupil." The ray on the right side connecting to the top extreme of the lens is extended upward in a dotted line, which is intersected by another dotted line connecting the center of the lens and the entrance pupil's location on the ray. At the point of intersection a different vertical marker is located and also reflected on the opposite side below. It is labeled "Exit pupil."

In a more complex optical system, containing many lenses and many apertures, it is in general necessary to trace rays through the entire system in order to determine which aperture constitutes the most severe restriction on the ray bundles and therefore which aperture must be imaged to find the entrance and exit pupils.

Once the location and extent of the exit pupil are known, the effects of diffraction on the image of a point-source object can be calculated. For an object point source, a converging bundle of rays fills the exit pupil on its way to a geometrical image. If the optical system has no aberrations, the geometrical image is an ideal point and the converging bundle defines a perfect spherical wave. The exit pupil limits the angular extent of the converging bundle. The Fraunhofer diffraction formula can now be applied at the exit pupil, using the distance from that pupil to the image as the distance appearing in the formula.

C Polarization and Jones Matrices

Birefringent media play an important role in the analysis of spatial light modulators of various kinds, as described in [Chapter 9](#). In this appendix we introduce a tool for analyzing polarization-based devices, the so-called Jones calculus, first introduced by R.C. Jones. For an alternative discussion, together with references, see [\[135\]](#), Section 4.3.1.

For simplicity, we restrict attention here to monochromatic light, since the problems of interest here arise primarily in coherent optical systems. However, the theory is more general, and can be extended to both narrowband and broadband optical signals with appropriate modifications.

C.1 Definition of the Jones Matrix

Consider a monochromatic light wave, polarized in the (x, y) plane, but with an arbitrary state of polarization in that plane. Let the polarization state be defined by a vector $\vec{U} \rightarrow$ formed from the complex amplitudes (phasor amplitudes) of the x and y components of polarization, U_X and U_Y , as follows:

$$\vec{U} \rightarrow = U_X \hat{x} + U_Y \hat{y}$$

$$\vec{U} \rightarrow = \begin{bmatrix} U_X \\ U_Y \end{bmatrix}$$

(C-1)

We will refer to $\vec{U} \rightarrow$ as the *polarization vector* of the light. Some examples of unit-length polarization vectors describing light with different states of polarization are as follows:

Linearly polarized in the x direction: $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$, Linearly polarized in the y direction: $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$, Linearly polarized at $+45$ degrees: $\begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$, Right-hand circularly polarized: $\begin{bmatrix} 1 \\ -j \end{bmatrix}$, Left-hand circularly polarized: $\begin{bmatrix} 1 \\ j \end{bmatrix}$. (C-2)

$$\text{Linearly polarized in the } x \text{ direction: } \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

$$\text{Linearly polarized in the } y \text{ direction: } \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

$$\text{Linearly polarized at } +45 \text{ degrees: } \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad (\text{C-2})$$

$$\text{Right-hand circularly polarized: } \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -j \end{bmatrix},$$

$$\text{Left-hand circularly polarized: } \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ j \end{bmatrix}.$$

As an aside, the convention adopted in optics is to define left-hand and right-hand circular polarization as follows. The observer always looks “head-on” into the wave as it approaches, i.e. towards the source of the light. If from such a perspective the polarization vector is rotating (with a period equal to the optical period and without change of length) in the *clockwise* sense, then the wave is said to be *right-hand circularly polarized*. This is because if you point the thumb of your right hand towards the source, the direction your fingers curl is clockwise, which in this case is the direction of rotation of the polarization vector. If, on the other hand, the direction of rotation is *countrerclockwise*, then for reasons that are probably now obvious we call this wave *left-hand circularly polarized*.

Left-hand and right-hand elliptical polarizations are similar to circular polarizations except that the length of the polarization vector changes periodically as the vector rotates.

When light passes through a polarization-sensitive device, the state of polarization of the wave will in general change, and it is of interest to find a simple representation of the new state of polarization, described by the vector \vec{U}' , in terms of the initial state of polarization \vec{U} . All of the polarization devices of interest here are *linear*, and for such devices the initial and final polarization vectors can be related through a 2×2 matrix \mathbf{L} , known as the *Jones matrix*,

$$\vec{U}' = \mathbf{L} \vec{U}.$$

$$\vec{U}' = \mathbf{L} \vec{U} = \begin{bmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \end{bmatrix} \vec{U}.$$

(C-3)

The four elements of the Jones matrix fully describe the effects of a linear device on the state of polarization of the wave.

When light passes through a sequence of linear polarization devices, the Jones matrices of the various transformations can be chained together, defining a single new Jones matrix for the sequence of devices through the relation

$$\mathbf{L} = \mathbf{L}_N \cdots \mathbf{L}_2 \mathbf{L}_1,$$

$$\mathbf{L} = \mathbf{L}_N \cdots \mathbf{L}_2 \mathbf{L}_1,$$

(C-4)

where \mathbf{L}_1 is the Jones matrix of the first device encountered, \mathbf{L}_2 that of the second device, etc.

C.2 Examples of Simple Polarization Transformations

Perhaps the simplest transformation of the state of polarization of a wave is that defined by a rotation of the coordinate system within which the wave is described (the wave itself does not change under such a rotation, only our mathematical description of it). If the (x, y) coordinate system is rotated by angle θ in the counterclockwise direction (as illustrated in Fig. C.1), simple geometry shows that

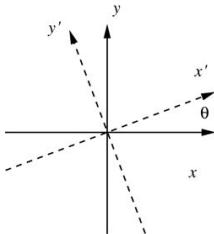


Figure C.1
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure C.1 Coordinate rotation. The direction of wave propagation is out of the page.

A coordinate plane shows horizontal axis x and vertical axis y . A dotted line copy of the axes, namely, x dash and y dash, shows a slight counterclockwise rotation of the axes x and y such that y dash cuts through the second and fourth quadrants, while x dash cuts through the first and the third quadrants. Axis x dash makes an angle measuring θ with axis x in the clockwise direction.

$$U_X' = \cos\theta U_X + \sin\theta U_Y, \quad U_Y' = -\sin\theta U_X + \cos\theta U_Y,$$

$$\begin{aligned} U'_X &= \cos\theta U_X + \sin\theta U_Y \\ U'_Y &= -\sin\theta U_X + \cos\theta U_Y, \end{aligned}$$

and therefore that the Jones matrix for a coordinate rotation is given by

$$L_{\text{rotate}}(\theta) = \cos\theta \sin\theta \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}.$$

(C-5)

Closely related to the Jones matrix of a coordinate rotation is the Jones matrix of a polarization device that transforms the polarization of a linearly polarized wave, initially polarized in direction θ_1 with respect to the x axis, into a linearly polarized wave with new polarization direction $\theta_2 = \theta_1 + \theta$. Such a device is called a *polarization rotator*. Since the polarization vectors before and after rotation are given, respectively, by $\begin{bmatrix} \cos\theta_1 \\ \sin\theta_1 \end{bmatrix}$

and $\cos\theta \begin{bmatrix} \cos\theta_2 \\ \sin\theta_2 \end{bmatrix}$, the Jones matrix of a device that rotates the polarization counterclockwise by angle θ must be given by

$$LR(\theta) = L_{\text{rotate}}(-\theta) = \cos\theta - \sin\theta \sin\theta \cos\theta.$$

$$L_R(\theta) = L_{\text{rotate}}(-\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

(C-6)

A second simple case is one in which the X and Y components of the wave undergo different phase delays. A device introducing such a polarization transformation is called a *wave retarder*. For example, a transparent birefringent plate of thickness d having refractive index n_X for the polarization component in the x direction and refractive index n_Y for the polarization component in the y direction, will introduce phase delays $\phi_X = 2\pi n_X d / \lambda_0$ and $\phi_Y = 2\pi n_Y d / \lambda_0$, respectively, in those two components. The Jones matrix for such a transformation can be written

$$L_{\text{retard}}(\Delta) = 100e^{-j\Delta},$$

$$L_{\text{retard}}(\Delta) = \begin{bmatrix} 1 & 0 \\ 0 & e^{-j\Delta} \end{bmatrix},$$

(C-7)

where λ_0 is the vacuum wavelength of light, a common phase delay suffered by both components has been dropped, and the *relative* phase shift Δ is given by

$$\Delta = 2\pi(n_X - n_Y)d/\lambda_0.$$

$$\Delta = \frac{2\pi (n_X - n_Y) d}{\lambda_0}.$$

(C-8)

A wave retarder of special interest is a *quarter wave plate*, for which $\Delta = \pi/2$. The Jones matrix for such a device is

$$L_{\text{retard}}(\pi/2) = 100-j.$$

$$L_{\text{retard}}(\pi/2) = \begin{bmatrix} 1 & 0 \\ 0 & -j \end{bmatrix}.$$

(C-9)

The device is easily seen to convert linearly polarized light with polarization direction at 45° to the x axis, described by the polarization vector $1211 \begin{bmatrix} 1 \\ \frac{1}{\sqrt{2}} \\ 1 \end{bmatrix}$, into right-hand circularly polarized light described by polarization vector $121-j \begin{bmatrix} 1 \\ \frac{1}{\sqrt{2}} \\ -j \end{bmatrix}$. Equivalently, this device converts left-hand circularly polarized light $121j \begin{bmatrix} 1 \\ \frac{1}{\sqrt{2}} \\ j \end{bmatrix}$ into linearly polarized light $1211 \begin{bmatrix} 1 \\ \frac{1}{\sqrt{2}} \\ 1 \end{bmatrix}$.

Another wave retarder of special interest is a *half-wave plate*, for which $\Delta=\pi$ and

$$L_{\text{retard}}(\pi) = 100-1.$$

$$\mathbf{L}_{\text{retard}}(\pi) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

(C-10)

Comparison of the Jones matrix for such a device with [Eq.\(C-6\)](#) shows that a half-wave plate is a device that rotates the polarization of a wave, initially linearly polarized at 45° to the x axis, by 90° .

As a final example of a polarization device we consider a *polarizer* (or equivalently a *polarization analyzer*) which passes only the wave component that is linearly polarized at angle α to the x axis. With a small amount of work it can be shown that the Jones matrix for such a device is given by

$$L(\alpha) = \cos 2\alpha \begin{bmatrix} \sin \alpha & \cos \alpha \\ \sin \alpha & \cos \alpha \end{bmatrix} \begin{bmatrix} \cos \alpha & \sin \alpha \\ \sin \alpha & \cos \alpha \end{bmatrix} = \begin{bmatrix} \cos^2 \alpha & \sin \alpha \cos \alpha \\ \sin \alpha \cos \alpha & \sin^2 \alpha \end{bmatrix}.$$

(C-11)

C.3 Reflective Polarization Devices

Until this point we have considered only polarization devices used in transmission. Since many spatial light modulators operate in a reflective mode, we turn attention to such a geometry.

Consider a reflective polarization device as illustrated in [Fig. C.2](#). Light enters the device from the left, with normal incidence assumed. It passes through a polarization element having Jones matrix L , is normally incident on a lossless mirror, reflects from the mirror, and passes a second time through the same polarization element. We wish to specify the Jones matrix for an equivalent *transmissive* device that will function in the same way as this reflective device.

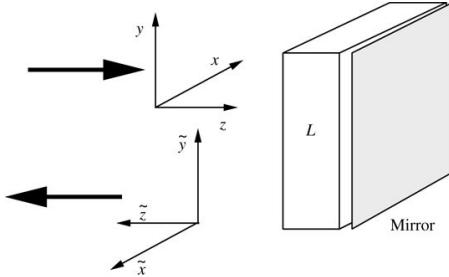


Figure C.2
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure C.2 Reflective polarization device.

In the illustration, a rightward arrow points at a 3 dimensional plane with rightward horizontal axis z and upward vertical axis y , and third axis x oriented away from the viewer. Below it is a leftward arrow pointing away from another 3 dimensional plane with leftward horizontal axis z tilde, upward vertical axis y tilde, and the third axis x tilde oriented toward the viewer. On the right side is cuboid block labeled L standing on its thin rectangular base. A mirror is standing flush with the right face of the cuboid.

An important point to consider at the start is that we will consider only *reciprocal* polarization elements before the mirror. For a reciprocal element, the coupling from, say, the x x component of polarization to the y y component of polarization on the forward pass through the device must equal the coupling from the y y component back to the x x component on the reverse pass. In addition the forward coupling from the y y component to the x x component must be the same as the backward coupling from x x to y y . For a reciprocal element, the Jones matrix for backward passage of light is exactly equal to the transpose of the Jones matrix for forward passage of light. Most polarization elements are reciprocal, the most important exceptions being devices that operate by the Faraday effect in the presence of a magnetic field. For such devices the dependence on the direction of the magnetic field destroys reciprocity, and the Jones matrix for reverse propagation is identical with the Jones matrix for forward propagation.

It is also important to note several geometrical factors at the start. First, we specify the polarization vectors of waves by examining the polarization state from a “heads-on” geometry, looking towards the source and using x x and y y axes that form a right-hand coordinate system, with the z z axis pointing in the direction of propagation. This is a convention that we must consistently apply. Note that for a transmissive device, the coordinate system both before and after

passage through the device is right-handed. We attempt to retain this convention even with the reflective device.

As shown in [Fig.C.2](#), the z^z axis is taken to reverse direction after reflection to become $z^{\tilde{z}}$. We have also shown the x^x axis reversed to obtain a right-hand system, with x^x being changed to $x^{\tilde{x}}$. However, for the time being, we allow the coordinate system to be left-handed, converting to a right-handed system shown only at the very end.

Consider now the progress of a wave as it travels through the reflective device. It begins with a polarization state described by a vector $\vec{U} \rightarrow$. This polarization state is modified by passage through the polarization element, yielding a polarization state

$$\vec{U} \rightarrow' = L \vec{U} \rightarrow.$$

$$\vec{U}' = L \vec{U}.$$

Next the light reflects from the mirror. Since the tangential components of the electric field must be zero at a perfectly conducting boundary, the electric field components after reflection are the negative of their values before reflection. However, we regularly drop constant phase factors, and a negation of the two components of the electric field is just a common phase factor of 180°

180° that we drop. So with this understanding, after reflection, the field components UX^{U_X} and UY^{U_Y} are the same as they were before reflection, when measured in the original (x,y) coordinate system.

The wave now proceeds back through the polarization element. As argued above, for a reciprocal device, the Jones matrix under reverse propagation is $L^t t^t$, where the superscript t indicates the ordinary matrix transpose operation.

Finally, if we wish to specify the polarization vector leaving the element in a *right-hand* coordinate system, rather than a left-hand system, we must reverse either the direction of the x^x axis or the direction of the y^y axis. We choose to reverse the direction of the x^x axis. Such a reversal is accounted for by a Jones matrix of the form

$$R=-1001.$$

$$R = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Thus the transmission equivalent of the reflective device has a Jones matrix of the form

$$L_{\text{reflect}} = R L^t t^t L.$$

$$L_{\text{reflect}} = R L^t t^t L.$$

(C-12)

As an example, consider a polarization device that consists of a simple coordinate rotation by angle $+θ +θ$, followed by reflection from a mirror. On passage through the coordinate rotation the second time, in the backwards direction, the coordinate system is once again rotated, but this time

back to its original orientation. Utilizing (C-12), the Jones matrix for the entire device, expressed in a right-hand coordinate system at the output, is

$$L = -1001 \cos\theta - \sin\theta \sin\theta \cos\theta \cos\theta \sin\theta - \sin\theta \cos\theta = -1001.$$

$$L = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Thus the only effect of passage through this simple device is a reversal of the direction of the x^x axis, a reversal we intentionally introduced to ensure a right-handed system. Note that in this case the transpose operation was critical to obtaining the right result.

D The Grating Equation

Gratings of one kind or another play important roles in many of the chapters of this book. In this appendix we present a very brief derivation of the so-called *grating equation*, which describes the most fundamental behavior of a grating of any kind.

In [Fig. D.1](#) we illustrate the simplest case of a transmission grating. For simplicity the grating is assumed to consist of a periodic array of slits or pinholes, only two of which are shown. While a particular type of grating is shown, the arguments and results apply equally well to any periodic grating, regardless of the amplitude and/or phase transmission of a grating period.

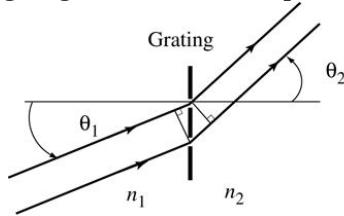


Figure D.1
Goodman, *Introduction to Fourier Optics*, 4e,
© 2017 W. H. Freeman and Company

Figure D.1 Transmission grating geometry.

The illustration shows a vertical grating with two holes; a horizontal line passes through the upper hole. To the left of the grating is medium index n₁ and to the right is n₂. From the bottom left corner two mutually parallel upward rays run rightward, one reaches the upper hole while the other, the lower hole. The upper ray makes angle theta 1 with the horizontal line in the counterclockwise direction. On the other side of the grating, both rays bend equally upward, thus continuing to run parallel to each other. The lower ray makes angle theta 2 with the horizontal line in the counterclockwise direction. A perpendicular is dropped to the lower ray from where the upper ray bends. And similarly, a perpendicular is dropped to the upper ray from where the lower ray bends.

For the purposes of this discussion, sign conventions are adopted. Angles of rays are positive in the counterclockwise direction from the normal to the grating, and negative in the clockwise direction. Thus, in the figure above, both θ_1 and θ_2 are positive.

There are two incoming rays shown incident at angle θ_1 with respect to a normal to the plane of the grating, each passing through an adjacent period of the grating, and two exiting rays leaving the grating at angle θ_2 to the normal. The refractive index on the left of the grating is n_1 and on the right it is n_2 . A grating order leaves the grating in a direction such that the path length difference between the upper and lower paths is an integer multiple of a wavelength. The extra physical distance traveled by the lower ray with respect to the upper ray is $\Lambda \sin \theta_2 - \Lambda \sin \theta_1$, where Λ is the period of the grating. The extra “optical” distance (taking account of the refractive indices) is $n_2 \Lambda \sin \theta_2 - n_1 \Lambda \sin \theta_1$. The m

m th grating order will be observed whenever this extra distance is an integer multiple $m\lambda$ of the free-space optical wavelength λ , i.e.

$$n_2 \Lambda \sin \theta_2 - n_1 \Lambda \sin \theta_1 = m\lambda,$$

$$n_2 \Lambda \sin \theta_2 - n_1 \Lambda \sin \theta_1 = m\lambda,$$

or equivalently

$$n_2 \sin \theta_2 = n_1 \sin \theta_1 + m \frac{\lambda}{\Lambda}.$$

(D-1)

This result is the grating equation for a transmission grating.

A “positive” diffraction order ($m > 0$) corresponds to a (signed) angle $\theta_2 > \theta_1$ that is greater than θ_1 ; for such an order the path length travelled by a ray increases by $m\lambda$ as we move down the grating by one period. A “negative” diffraction order ($m < 0$) corresponds to $\theta_2 < \theta_1$; for such an order the path length travelled by a ray decreases by $-m\lambda$ as we move down the grating by one period. For the “zero” order, $\theta_2 = \theta_1$ and the path lengths travelled by all rays are identical. For a reflection grating, the same results hold [163] with the exception that, since both incident and diffracted rays are on the same side of the grating, $n_1 = n_2 = n$.

Chapter 2 Notes

¹ When a single limit of integration appears above or below a double integral, then that limit applies to *both* integrations.

² For a more detailed discussion of the delta function, including definitions, see [Appendix A](#).

³ Note the subscript in G_o is added simply because the functional form of the expression for the transform in polar coordinates is in general different than the functional form for the same transform in rectangular coordinates.

⁴ For a tutorial discussion of the importance of quadratic-phase functions in various fields of optics, see [\[278\]](#).

⁵ The name “chirp function”, without the “finite” qualifier, will be used for the infinite-length quadratic-phase exponential, $\exp[j\pi\beta(x^2 + y^2)]$.

⁶ From the definition [\(2-38\)](#) the dimensions of $f_X(\ell) f_X^{(\ell)}$ and $f_Y(\ell) f_Y^{(\ell)}$ are both *cycles per meter*. The dimensions of β are meters⁻² meters⁻².

⁷ We are taking some liberties with notation here, since the function h^h on the left is a function of four variables, while the function h^h on the right is a function of only two variables. However, it will be clear from the number of arguments of the function whether we are representing a space-invariant impulse response (two variables) or a space-variant impulse response (four variables).

⁸ The complex exponentials are not the only eigenfunctions of linear invariant systems. See, for example, [Prob. 2-11](#).

⁹ For simplicity we assume that this rectangle is centered on the origin. If this is not the case, the arguments can be modified in a straightforward manner to yield a somewhat more efficient sampling theorem.

¹⁰ It is possible to see that this is the case with a simple example. Consider the infinite-length one-dimensional function $\cos(2\pi f_0 x) \cos(2\pi f_0 x)$. Its spectrum consists of two delta functions, one located at $-f_0$ and the other at $+f_0$. If we truncate this function by multiplying it by $\text{rect}(x/L) \text{rect}(x/L)$, the delta functions in the spectral domain will be convolved with a sinc function, the Fourier transform of the rect function. Thus the bandlimited pair of delta functions has been changed to a pair of sinc functions. But the sinc function in the spectral domain has tails of infinite extent, so the spectrum is no longer bandlimited. The same is true of *any* bandlimited function that is subsequently space limited—the spectrum will no longer have finite extent.

¹¹ For a periodic sequence, such as results from sampling in the frequency domain, a circular shift represents a shift that simply translates the periodic array through a portion of its primary period. For example, if we translate the sequence by one unit to the right, the element that moved out of the primary period to the right enters the new periodic sequence from the left.

Chapter 3 Notes

- ¹ For a more detailed discussion of these inconsistencies, see [Section 3.5](#).
- ² The reader may wish to verify that, for our choice of clockwise rotation of phasors, the description of an expanding wave should have a + sign in the exponential.
- ³ As we shall see, objections to the use of the Kirchhoff boundary conditions arise, not because of the fringing effects, but rather because of certain internal inconsistencies.
- ⁴ The fact that one theory is consistent and the other is not does not necessarily mean that the former is *more accurate* than the latter.
- ⁵ It is perhaps worth noting that the exact first Sommerfeld solution, [\(3-35\)](#), satisfies the Helmholtz equation, but the approximate solution, [Eq.\(3-41\)](#) does not.
- ⁶ A delta function impulse response is required if the boundary conditions are to be reproduced.
- ⁷ Hereafter we drop the subscript on the first Rayleigh-Sommerfeld solution, since it will be the solution we use exclusively.
- ⁸ Note that evanescent waves are predicted only under the very same conditions for which the use of the scalar theory is suspect. Nonetheless, they are a real phenomenon, although perhaps more accurately treated in a full vectorial theory.
- ⁹ We can usually assume that the distance z is larger than a few wavelengths, allowing us to completely drop the evanescent components of the spectrum.

Chapter 4 Notes

- ¹ The reader may wonder why the generation of both an electron and a hole does not lead to a charge $2q$ rather than q in this equation. For an answer, see [305], p. 650.
- ² An interesting relation between the Fresnel diffraction formula and an entity known as the “fractional Fourier transform” exists. The interested reader can consult [273] and the references contained therein.
- ³ In the past it has been customary to introduce a graphical aid known as “Cornu’s spiral” as a tool for estimating values of Fresnel integrals. Modern computer software packages that contain the Fresnel integrals have made this graphical aid largely obsolete, so we have omitted it here.
- ⁴ Babinet’s principle states that when transmitting and non-transmitting portions of a binary aperture are interchanged (complemented), the diffracted fields are complemented, that is, dark becomes light, and visa versa ([34], p. 424).
- ⁵ This problem was inspired by Prof. James Fienup.

Chapter 5 Notes

¹ This bandwidth is strictly valid only when the Fresnel number $\tilde{N}_F = (L/2)^2 / (\lambda z)$

$\tilde{N}_F = (L/2)^2 / (\lambda z)$ of the finite-length quadratic-phase exponential (not to be confused with the Fresnel number N_F of the aperture function) is greater than 0.25. However, as the last paragraph in this subsection argues, the length L must be increased as N_F decreases below 0.25, with the result that \tilde{N}_F stays greater than 0.25 for any N_F of the aperture function.

² When a conclusion is stated as being based on simulations, it means that calculations of the discrete convolution were performed with various M . The parameter K was set equal to

$M^2 / (4N_F)$, and an allowable level of normalized intensity at the fold-over frequency was chosen. The results were compared with the known analytical results. A subjective judgement was then made, based on logarithmic plots of the intensity distribution predicted by both analytic and discrete results, as to when the simulation results were satisfactorily close to the analytical result. Of necessity, subjective judgements are made.

³ Strictly speaking, the FFT requires a number $KN \log_2 N$ of complex multiplications, where k is a constant that varies in size, depending on the particular FFT algorithm used.

However, the constant k is of the order of unity, and we will use $k=1$ in our calculations of computational complexity. There do exist algorithms that will perform an FFT on a complex-valued array in a slightly smaller number of operations than $N \log_2 N$ [184]. However, we will be comparing the computational complexities of the various methods discussed here, and we will consistently use the complexity $N \log_2 N$ for all FFTs in the comparison.

⁴ Assuming that the primary sources of aliasing at the edges of the diffraction pattern arise from the directly adjacent periods of the DFT, the amplitude of the calculated diffraction pattern will be at most a factor of 2 too large, and the intensity at most a factor of 4 too large. By specifying a level for the intensity at the edge of the diffraction pattern, we are in effect requiring that the intensity aliasing associated with any one period of the periodic spectrum be no more than a factor of 4 below this level.

⁵ Note that if $N_F < 0.25$ we are in the Fraunhofer region, in which case the quadratic-phase exponential reduces to unity. In this case the factor M in Eq. (5-32), included to account for multiplying the two arrays, vanishes. However, the complexity is dominated by $N \log_2 N$ in any case.

⁶ Note that the Fourier transforms here and in the section to follow do not have the scaling factor of $1/(\lambda z)$ in frequency space.

⁷ If we were only interested in a rectangular aperture of known width, we could precompute the DFT of such an aperture and avoid repeated DFTs of the aperture function. However, in more general cases than a simple rectangular aperture, the DFT of the aperture function would have to be calculated.

⁸ A portion of the efficiency of this method comes from the fact that complex sums are much faster to perform than complex multiplies. The $2N^2\log_2 N + M^2$ complex multiply-and-adds for the two-dimensional Fresnel transform approach are replaced by M^2 complex multiplies of the two arrays, M^2 sums in the projection operation, and $N\log_2 N$ complex multiply-and-adds in the one-dimensional DFT.

⁹ Using *Mathematica*, it is possible to create a three-dimensional display of a circularly symmetric function from its discretized central slice using two commands: 1) `ListInterpolate` changes the discrete array into a continuous interpolated function, and 2) `RevolutionPlot` creates a three-dimensional display of the circularly symmetric continuous function.

Chapter 6 Notes

¹ The astute reader may wonder whether there should be an additive constant in the expression for the phase distribution obtained by integrating instantaneous frequency. In this case, we know that the ray incident along the optical axis leaves the lens along the optical axis, implying that $\theta_2 = \theta_1$ when $y=0$. This fact establishes that the additive constant should be zero.

² This assumption will not necessarily be true if the imaging system has significant aberrations.

Chapter 7 Notes

¹ In general it is not necessary that the entrance pupil lie to the left of the exit pupil as shown in [Fig. 7.1](#). However the conceptual idea of a system mapping the light incident on the entrance pupil to the light leaving the exit pupil remains valid, regardless of the order of the two pupils.

² We reserve the symbols z_1 and z_2 for the distances from the object to the first principal plane and the distance from the second principal plane to the image, respectively.

³ We have retained the assumption of monochromatic illumination but will remove it in the section to follow.

⁴ The impulse response must have dimensions $1/m^2$, so the expression below has an extra $1/\lambda z_i$ factor compared to earlier equations for the Fraunhofer diffraction pattern. See [\(6-44\)](#).

⁵ Often advantages are gained by using much more complex changes of coordinates, particularly when the analysis is nonparaxial. We have chosen to remain with the simplest coordinate system consistent with a paraxially space-invariant system. For discussions of other coordinate mappings (many of which are due to H.H. Hopkins) and their advantages, see [\[373\]](#).

⁶ This is a sufficient but not necessary condition for complete coherence. For example, when monochromatic light from a point source is passed through a stationary diffuser, the relative phases of the light at any two points behind the diffuser remain correlated. Therefore the transmitted light is still spatially coherent, even though it no longer appears to originate from a point source. Note, however, that before impinging on the diffuser it did originate from a point source.

⁷ As mentioned previously, this assumption may not be true for a system with significant aberrations.

⁸ Here and throughout, we shall retain the subscripts X and Y on frequency variables, even though the space variables to which they correspond may have different symbols.

⁹ Note that this conclusion has been drawn only for a system free from aberrations. As we shall see in [Section 7.4](#), a system that has aberrations is not free from phase distortion within its passband.

¹⁰ This should not be taken to imply that the incoherent system has twice the resolving power of the coherent system. See [Section 7.5](#).

Chapter 8 Notes

¹ For the special case of an imaging system focused at infinity, the *hyperfocal distance* is defined as the shortest distance in object space beyond which all objects are in focus. If the object is at the hyperfocal distance in such a case, (8-5) holds for objects that are closer to the lens than the hyperfocal distance, but the depth of field beyond the hyperfocal distance is infinite. The expression for Δz_o in (8-5) is most useful in microscopy and lithography.

² A full analysis of the Lyot coronagraph should include the effects of the telescope mirrors that proceed it in the optical train, which can lead to polarization effects not included here [40].

³ In practice, it is the ratio of the intensity of the image of the planet to the intensity of the *sidelobes* of the star point-spread function that determines detectability of the planet.

⁴ In a microscope, there is a tradeoff between the resolution obtained and the field of view over which that resolution holds. High NA systems have a small field of view and high resolution, while low NA systems have a larger field of view and a lower resolution.

⁵ Upsampling refers to the process of taking an original set of M samples, interpolating with sinc functions as required by the sampling theorem, and resampling with a larger number N of samples. Thus the length of the discrete sequence is increased from M to N ($N > M$) samples, where in this case, M refers to the number of samples required when a low-resolution image is sampled at the Nyquist rate, and N corresponds to the number of samples required when the high-resolution image is sampled at the Nyquist rate. Upsampling in the space domain can be accomplished by zero-padding in the spectral domain followed by an inverse DFT.

⁶ The approach taken here is very similar to a modulated wideband converter designed to reduce the sampling rate in analog-to-digital conversion. See [254] and [255].

⁷ There is a close relationship between the properties of light fields and Wigner distributions. See [383] for a discussion of the relationship.

Chapter 9 Notes

- ¹ Mylar base should be avoided when coherent light is used, due to the fact that it is birefringent and causes unwanted variations of the polarization and phase of the transmitted light.
- ² The threshold is actually not a fixed number, but a statistical one. The assumption that the threshold is four atoms is an approximation.
- ³ These ideal and quantized gratings may also be considered to be local approximations to more general gratings for which the local period, and therefore the angle of deflection, change across the grating.
- ⁴ The liquid crystal cell is filled with material at an elevated temperature, where the phase of the liquid crystal is smectic-A. Such a phase has no tilt angle, and therefore the molecules align with their long direction parallel to the alignment grooves. When the material cools, it is transformed to the smectic-C* state, which has the tilt mentioned above.
- ⁵ In the real device, operation is complicated by the fact that the photosensor and the light-blocking layer together form an electrical diode with asymmetric I-V $I - V$ properties.

Chapter 10 Notes

¹ For a discussion of the history of the phase contrast technique, as well as the scientific life of Frits Zernike, see [110].

² In practice, phase-contrast microscopes usually have a source that is a circular ring and a phase-shifting structure that is also a circular ring, placed over the image of the source in the focal plane. However, the explanation based on the assumption of point-source illumination is somewhat simpler to understand.

³ An optical system is called anamorphic if the focusing powers of the system in two orthogonal directions are unequal.

⁴ Historically, this type of filter had been preceded by a related but less general technique, known as the *hard-clipped filter*, which was a filter generated by computer and is the first example of what now might be called a *phase-only* filter. While the hard-clipped filter was used in radar signal processing as early as 1961, due to classification it did not appear in the open literature until 1965 [213]. The fundamental idea that an interferometric recording of the Fourier transform of the impulse response could realize a complex filter with a desired transfer function or its conjugate is attributable to C. Palermo (private communication, E.N. Leith).

⁵ Here and frequently in what follows, we drop a multiplicative factor $1/j$ associated with the optical Fourier transform, with the justification that we can always change the phase reference for convenience.

⁶ Strictly speaking, the function should also be conjugated, but in practice the functions g^S and h^H are usually real.

⁷ For a detailed discussion of the concept of power spectral density, see [135], Section 3.3.2.

⁸ For simplicity, we have assumed that the transfer function S^S has been normalized to unity at the origin.

⁹ We retain the convention that the $+1 + 1$ order is the order deflected counter-clockwise while the $-1 - 1$ order is the order deflected clockwise.

¹⁰ Strictly speaking, only $(M-1) \times (N-1)$ additions are needed, but we count generation of the first component of a sum as addition with zero.

Chapter 11 Notes

¹ The reason for calling this angle 2Θ rather than Θ will become evident when we consider fringe orientation through the depth of a thick emulsion.

² Spatial filtering operations are seldom used in practice to separate the twin images. If the reference angle satisfies the requirements to be derived here, the images will separate of their own accord due to the different directions of propagation of the respective wave components (cf. Fig. 11.7). However, spatial-filtering arguments do provide a conceptually simple way of finding sufficient conditions for separation.

³ Note that diverging spherical waves have a negative sign in the exponent, whereas in the past they have had a positive sign. This is because the values of z^z being used here are negative, whereas in previous cases they were positive. It remains true that, if the overall sign in the exponent is positive, the spherical wave is diverging, and if it is negative, the wave is converging.

⁴ Note also that the optical wavelength in the emulsion is smaller than the vacuum wavelength by a factor $1/n$, which for $n \approx 1.5$ is a factor of $2/3$.

⁵ Because this is a pseudoscopic image of a pseudoscopic image, it is orthoscopic from the perspective of the viewer.

⁶ The path-length difference could be any integer number of wavelengths. We consider only a single wavelength, which corresponds to the *first order* diffracted wave.

⁷ We use the subscript i^i on the k^k -vector of diffracted wave because in most circumstances it is an “image” wave. The i^i does *not* stand for “incident.”

⁸ Remember that the angle between two waves within the recording medium is different than the angle between them external to that medium, due to the generally higher refractive index of the recording medium.

⁹ An exception is found for nonlinear crystals, which may have sizes that are comparable in all three dimensions.

¹⁰ Note that the angle of incidence θ^θ should not be confused with the angle Θ^Θ , which is the half-angle between the two recording beams.

¹¹ Note that the definition of α^α used here is the reciprocal of the propagation distance within which the *field* drops to $1/e^{1/e}$ of its original value. The intensity drops to $1/e^2^{1/e^2}$ in this same distance.

¹² In expressions involving both λ^λ and θ^θ , it is possible to take them to have their values outside the emulsion or inside the emulsion, as long as the same condition holds for both (cf. Prob. 9-7(a)).

¹³ When evaluating this equation under the condition $\chi > \Phi$, use must be made of the fact that $\sinh u = i \sin u$.

Chapter 12 Notes

1 As frequency changes, the portion of the propagating mode that extends into the cladding changes slightly, leading to a change in propagation constant of the mode, or *waveguide dispersion*.

2 This final step is made clearer by noticing that $\omega_2 - \omega_1 = \Delta\omega = 2\pi c(1/\lambda_2 - 1/\lambda_1) = -2\pi c \lambda \Delta\lambda$

$\omega_2 - \omega_1 = \Delta\omega = 2\pi c(\frac{1}{\lambda_2} - \frac{1}{\lambda_1}) = -\frac{2\pi c}{\lambda^2} \Delta\lambda$ where $\Delta\lambda = \lambda_2 - \lambda_1$, and it has been assumed that $\Delta\lambda \ll \lambda_1$ $\Delta\lambda < < \lambda_1$ and $\lambda_2 \lambda_2$.

3 We continue to use λ to represent free-space wavelengths.

4 A thin transmission or reflection grating is referred to as a *blazed* grating if the profile of a period of the grating favors a particular diffraction order over others. For an example, see the discussion surrounding [Fig. 9.12](#).

5 In the event that the small angle approximation for $\sin\theta_2 \approx \theta_2$ is not valid, the more complex

$$x = \frac{f(\sin\theta_1 - \lambda/\Lambda)}{\sqrt{1 - (\sin\theta_1 - \lambda/\Lambda)^2}}$$

mapping $x = f(\sin\theta_1 - \lambda/\Lambda) - (\sin\theta_1 - \lambda/\Lambda)^2$ can be used.

6 Since we are interested in values of x that are less than x_0 , the value of v remains positive, as it should.

7 Both commercial two-layer liquid crystal arrays and acousto-optic modulators can achieve complex control when such control is desired.

8 These complex spectral amplitudes have important differences with respect to the usual complex amplitudes we have used many times throughout this book. At each point in the spectrum, the temporal frequency v is different. This means that the spectral components $R(v_1)$ and $S(v_2)$ can not interfere with one another when $v_1 \neq v_2$. Interference is possible only between components at the same frequency. The recording geometry ensures that $R(v)$ and $S(v)$ are properly aligned so that frequency-by-frequency interference can take place.

9 The first two terms (in brackets) are considered a single wave component, since they are propagating in the same direction.

10 β_{eff} can alternatively be expressed as $2\pi n_{\text{eff}}/\lambda$ where n_{eff} is an *effective refractive index* and λ is the free-space wavelength.

11 In this discussion, it is important to keep track of + + and - - signs. Angles are positive in the counter-clockwise direction, negative in the clockwise direction. For the geometry shown in the figure, θ_2 is negative, θ_1 is positive, and the diffraction order m is negative.

12 We continue to denote all wavelengths in waveguides as $\tilde{\lambda}$, while wavelengths λ are values in free space.

Appendix Notes

¹ We continue to use the word *function* due to its common use, even though it is not strictly correct.

Bibliography

1. E. Abbe. Beitrage zur Theorie des Mikroskops und der Mikroskopischen wahrnehmung. *Archiv. Microskopische Anat.*, 9:413–468, 1873.
2. E. H. Adelson and J. R. Bergen. The plenoptic function and elements of early vision. In M. Landy and J. A. Movshon, editors, *Computational Models of Visual Processing*, pages 3–20. MIT Press, 1991.
3. C. Aime. Apodized apertures for solar coronography. *Astronomy & Astrophysics*, 467:317–325, 2007.
4. M. A. Alonso. Wigner functions in optics: describing beams as ray bundles and pulses as particles. *Advances in Optics and Photonics*, 3:272–365, 2011.
5. J. J. Amodei. Analysis of transport processes during hologram recording in crystals. *RCA Rev.*, 32:185–198, 1971.
6. J. J. Amodei. Electron diffusion effects during hologram recording in crystals. *Appl. Phys. Lett.*, 18:22–24, 1971.
7. D. Z. Anderson. Competitive and cooperative dynamics in nonlinear optical circuits. In S. F. Zornetzer, J.L. Davis, and C. Lau, editors, *An Introduction to Neural and Electronic Networks*. Academic Press, 1990.
8. L. K. Anderson. Holographic optical memory for bulk data storage. *Bell Lab. Record*, 46:318–325, 1968.
9. M. Arm, L. Lambert, and I. Weissman. Optical correlation techniques for radar pulse compression. *Proc. I.E.E.E.*, 52:842, 1964.
10. E. H. Armstrong. A method for reducing disturbances in radio signaling by a system of frequency modulation. *Proc. IRE*, 24:689–740, 1936.
11. H. H. Arsenault. Distortion-invariant pattern recognition using circular harmonic matched filters. In H. H. Arsenault, T. Szoplik, and B. Macukow, editors, *Optical Processing and Computing*. Academic Press, San Diego, CA, 1989.
12. J. M. Artigas, M. J. Buades, and A. Filipe. Contrast sensitivity of the visual system in speckle imaging. *J. Opt. Soc. Am. A*, 11:2345–2349, 1994.
13. B. B. Baker and E. T. Copson. *The Mathematical Theory of Huygen's Principle*. Clarendon Press, Oxford, second edition, 1949.
14. P. R. Barbier, L. Wang, and G. Moddel. Thin-film photosensor design for liquid crystal spatial light modulators. *Opt. Engin.*, 33:1322–1329, 1994.
15. C. W. Barnes. Object restoration in a diffraction-limited imaging system. *J. Opt. Soc. Am.*, 56:575, 1966.
16. G. Barton. *Elements of Green's Functions and Propagation*. Oxford University Press, New York, NY, 1989.

17. M. Bass, editor. *Handbook of Optics*. McGraw-Hill, Inc., New York, NY, third edition, 2010.
18. M. J. Bastiaans. The Wigner distribution function applied to optical signals and systems. *Opt. Comm.*, 25:26–30, 1978.
19. M. J. Bastiaans. Wigner distribution function and its application to first-order optics. *J. Opt. Soc. Am.*, 69:1710–1716, 1979.
20. M. J. Bastiaans. Wigner distribution in optics. In M. Testorf, B. Hennelly, and J. Ojeda-Castaneda, editors, *Phase Space Optics*, chapter 1, pages 1–44. McGraw-Hill, 2010.
21. L. Beiser. *Holographic Scanning*. J. Wiley & Sons, New York, NY, 1988.
22. C. V. Bennett and B. H. Kolner. Upconversion time microscope demonstrating 103x magnification of femtosecond waveforms. *Opt. Lett.*, 24:783–785, 1999.
23. S. A. Benton. On a method for reducing the information content of holograms. *J. Opt. Soc. Am.*, 59:1545, 1969.
24. S. A. Benton and Jr. V. M. Bove. *Holographic Imaging*. John Wiley & Sons, Hoboken, NJ, 2008.
25. M. J. Beran and G. B. Parrent, Jr. *Theory of Partial Coherence*. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1964.
26. N. J. Berg and J. N. Lee. *Acousto-Optic Signal Processing: Theory & Applications*. Marcel Dekker, 1983.
27. E. Betzig. Proposed method for molecular optical imaging. *Optics Lett.*, 20:237–239, 1995.
28. E. Betzig, G. H. Patterson, R. Sougrat, O. W. Lindwasser, S. Olenych, J. S. Bonifacino, M. W. Davidson, J. Lippincott-Schwartz, and H. F. Hess. Imaging intracellular fluorescent proteins at nanometer resolution. *Science*, 15:1642–1645, 2006.
29. H. I. Bjelkhagen. *Silver-Halide Recording Materials*. Springer-Verlag, Berlin, 1993.
30. G. Bonnet. Introduction to metaxial optics. I. *Ann. des Télècom.*, 33:143–165, 1978.
31. G. Bonnet. Introduction to metaxial optics. II. *Ann. des Télècom.*, 33:225–243, 1978.
32. B. L. Booth. Photopolymer material for holography. *Appl. Opt.*, 11:2994–2995, 1972.
33. B. L. Booth. Photopolymer material for holography. *Appl. Opt.*, 14:593–601, 1975.
34. M. Born and E. Wolf. *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction*. Cambridge University Press, Cambridge, UK, seventh expanded edition, 1999.
35. C. J. Bouwkamp. Diffraction theory. In A. C. Strickland, editor, *Reports on Progress in Physics*, volume XVII. The Physical Society, London, 1954.
36. R. N. Bracewell. Two-dimensional aerial smoothing in radio astronomy. *Australia J. Phys.*, 9:297, 1956.
37. R. N. Bracewell. *The Fourier Transform and Its Applications*. McGraw-Hill Book Company, Inc., New York, second revised edition, 1986.
38. R. N. Bracewell. *Two Dimensional Imaging*. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1994.
39. W. L. Bragg. The X-ray microscope. *Nature*, 149:470, 1942.
40. J. B. Breckinridge and R. A. Chipman. Telescope polarization and image quality: Lyot coronagraph performance. In *Proc. S.P.I.E., Astronomical Telescopes & Instruments*, volume 9904, pages 042–057, Bellingham, WA, 2016. S.P.I.E.

41. E. O. Brigham. *The Fast Fourier Transform and its Applications*. Prentice-Hall, Upper Saddle River, NJ, 1988.
42. R. E. Brooks, L. O. Heflinger, and R. F. Wuerker. Interferometry with a holographically reconstructed reference beam. *Appl. Phys. Lett.*, 7:248, 1965.
43. R. E. Brooks, L. O. Heflinger, and R. F. Wuerker. Pulsed laser holograms. *I.E.E.E. J. Quant. Electr.*, QE-2:275–279, 1966.
44. B. R. Brown and A. W. Lohmann. Complex spatial filter. *Appl. Opt.*, 5:967, 1966.
45. B. R. Brown and A. W. Lohmann. Computer-generated binary holograms. *IBM J. Res. Dev.*, 13:160–168, 1969.
46. J. H. Bruning, D. R. Herriott, J. E. Gallagher, D. P. Rosenfeld, A. D. White, and D. J. Brangaccio. Digital wavefront measuring interferometer for testing optical surfaces and lenses. *Appl. Opt.*, 13:2693–2703, 1974.
47. O. Bryngdahl and A. Lohmann. Nonlinear effects in holography. *J. Opt. Soc. Am.*, 58:1325–1334, 1968.
48. C. B. Burckhardt. A simplification of Lee’s method of generating holograms by computer. *Appl. Opt.*, 9:1949, 1970.
49. G. W. Burr, F. Mok, and D. Psaltis. Large-scale holographic memory: experimental results. *Proc. S.P.I.E.*, 2026:630–641, 1993.
50. S. A. Campbell. *The Science and Engineering of Microelectronic Fabrication*. Oxford University Press, Oxford, UK, second edition, 2001.
51. A. Carlotti, R. Vanderbei, and N. J. Kasdin. Optimal pupil apodization of arbitrary apertures for high contrast imaging. *Optics Express*, 19:26796–26809, 2011.
52. P. Carré. Installation et utilisation du comparateur photoelectrique et interferentiel du Bureau International de Poids et Measures. *Metrologia*, 2:13–23, 1966.
53. D. Casasent, editor. *Optical Data Processing—Applications*. Springer-Verlag, Berlin, 1978.
54. D. Casasent and W.-T. Chang. Correlation synthetic discriminant functions. *Appl. Opt.*, 25:2343–2350, 1986.
55. D. Casasent and D. Psaltis. New optical transforms for pattern recognition. *Proc. I.E.E.E.*, 65:77–84, 1977.
56. S. K. Case and R. Alferness. Index modulation and spatial harmonic generation in dichromated gelatin films. *Appl. Phys.*, 10:41–51, 1976.
57. W. T. Cathey, B. R. Frieden, W. T. Rhodes, and C. K. Rushforth. Image gathering and processing for enhanced resolution. *J. Opt. Soc. Am. A*, 1:241–250, 1984.
58. H. J. Caulfield, editor. *Handbook of Holography*. Academic Press, New York, NY, 1979.
59. C.-C. Chang, H. P. Sardesai, and A. M. Weiner. Dispersion-free fiber transmission for femtosecond pulses by use of a dispersion-compensating fiber and a programmable pulse shaper. *Optics Lett.*, 23:283–285, 1998.
60. M. Chang. Dichromated gelatin of improved optical quality. *Appl. Opt.*, 10:2550–2551, 1971.
61. F. S. Chen, J. T. LaMacchia, and D. B. Fraser. Holographic storage in lithium niobate. *Appl. Phys. Lett.*, 13:223–225, 1968.
62. D. C. Chu, J. R. Fienup, and J. W. Goodman. Multi-emulsion, on-axis, computer generated hologram. *Appl. Opt.*, 12:1386–1388, 1973.

63. K. Chu, N. George, and W. Chi. Extending the depth of field through unbalanced optical path difference. *Appl. Opt.*, 47:6895–6903, 2008.
64. I. Cindrich, editor. *Holographic Optics: Design and Application*, volume 883 of *Proceedings of the S.P.I.E.*, 1988.
65. I. Cindrich and S. H. Lee, editors. *Holographic Optics: Optically and Computer Generated*, volume 1052 of *Proceedings of the S.P.I.E.*, 1989.
66. I. Cindrich and S. H. Lee, editors. *Computer and Optically Formed Holographic Optics*, volume 1211 of *Proceedings of the S.P.I.E.*, 1990.
67. I. Cindrich and S. H. Lee, editors. *Computer and Optically Generated Holographic Optics*, volume 1555 of *Proceedings of the S.P.I.E.*, 1990.
68. I. Cindrich and S. H. Lee, editors. *Diffractive and Holographic Optics Technology*, volume 2152 of *Proceedings of the S.P.I.E.*, 1994.
69. N. A. Clark, M. A. Handschy, and S. T. Lagerwall. Ferroelectric liquid crystal electro-optics using the surface-stabilized structure. *Mol. Cryst. Liq. Cryst.*, 94:213–234, 1983.
70. N. A. Clark and S. T. Lagerwall. Submicrosecond bistable electrooptic switching in liquid crystals. *Appl. Phys. Lett.*, 36:899–901, 1980.
71. D. H. Close. Holographic optical elements. *Opt. Engin.*, 14:408–419, 1975.
72. D. H. Close, A. D. Jacobson, J. D. Margerum, R. G. Brault, and F. J. McClung. Hologram recording on photopolymer materials. *Appl. Phys. Lett.*, 14:159–160, 1969.
73. G. Cochran. New method of making Fresnel transforms with incoherent light. *J. Opt. Soc. Am.*, 56:1513, 1966.
74. W. S. Colburn and K. A. Haines. Volume hologram formation in photopolymer materials. *Appl. Opt.*, 10:1636–1641, 1971.
75. R. J. Collier, C. B. Burckhardt, and L. H. Lin. *Optical Holography*. Academic Press, New York, NY, 1971.
76. S. A. Collins. Lens-system diffraction integral written in terms of matrix optics. *J. Opt. Soc. Am.*, pages 1168–1177, 1970.
77. P. S. Considine. Effects of coherence on imaging systems. *J. Opt. Soc. Am.*, 56:1001–1009, 1966.
78. T. R. Corle and G. S. Kino. *Confocal Scanning Optical Microscopy and Related Imaging Systems*. Elsevier, Inc., Amsterdam, 1996.
79. K. Creath. Phase-shifting speckle interferometry. *Appl. Opt.*, 24:3053–3058, 1985.
80. Lloyd Cross. Multiplex holography. Presented at the S. P. I. E. Seminar on Three Dimensional Imaging, but unpublished, August 1977.
81. L. J. Cutrona, E. N. Leith, C. Palermo, and L. J. Porcello. Optical data processing and filtering systems. *IRE Trans. Inform. Theory*, IT-6:386–400, 1960.
82. L. J. Cutrona and others. On the application of coherent optical processing techniques to synthetic-aperture radar. *Proc. I.E.E.E.*, 54:1026–1032, 1966.
83. J. C. Dainty, editor. *Laser Speckle and Related Phenomena*. Springer-Verlag, New York, second edition, 1984.
84. J.C Dainty and R. Shaw. *Image Science*. Academic Press, London, 1974.
85. W. J. Dallas. Phase quantization—a compact derivation. *Appl. Opt.*, 10:673–674, 1971.
86. W. J. Dallas. Phase quantization in holograms—a few illustrations. *Appl. Opt.*, 10:674–676, 1971.

87. H. Dammann. Spectral characteristics of stepped-phase gratings. *Optik*, 53:409–417, 1979.
88. D. J. De Bitteto. Holographic panoramic stereograms synthesized from white light recordings. *Appl. Opt.*, 8:1740–1741, 1970.
89. Y. N. Denisyuk. Photographic reconstruction of the optical properties of an object in its own scattered radiation field. *Sov. Phys. - Dokl.*, 7:543, 1962.
90. J. B. Develis and G. O. Reynolds. *Theory and Applications of Holography*. Addison-Wesley Publishing Company, Reading, MA, 1967.
91. A. R. Dias, R. F. Kalman, J. W. Goodman, and A. A. Sawchuk. Fiber-optic crossbar switch with broadcast capability. *Opt. Engin.*, 27:955–960, 1988.
92. E. R. Dowski, Jr. and W. T. Cathey. Extended depth of field through wave-front coding. *Appl. Opt.*, 34:1859–1866, 1995.
93. C. Dragone. Efficient N x N star couplers using Fourier optics. *J. Lightwave Techn.*, 7:479–489, 1989.
94. C. Dragone. Integrated optics N x N multiplexer on silicon. *IEEE Photon. Technol. Lett.*, 3:896–899, 1991.
95. C. Dragone. An N x N optical multiplexer using a planar arrangement of two star couplers. *IEEE Photon. Technol. Lett.*, 3:812–815, 1991.
96. D. E. Dudgeon and R. M. Mersereau. *Multidimensional Digital Signal Processing*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1984.
97. P. M. Duffieux. *L'Integral de Fourier et ses Applications à l'Optique*. Rennes, Société Anonyme des Imprimeries Oberthur, 1946.
98. P. M. Duffieux. *The Fourier Transform and Its Applications to Optics*. John Wiley & Sons, New York, NY, second edition, 1983.
99. U. Efron, editor. *Spatial Light Modulators and Applications I*, volume 465 of *Proceedings of the S.P.I.E.*, 1984.
100. U. Efron, editor. *Spatial Light Modulators and Applications II*, volume 825 of *Proceedings of the S.P.I.E.*, 1988.
101. U. Efron, editor. *Spatial Light Modulators and Applications III*, volume 1150 of *Proceedings of the S.P.I.E.*, 1990.
102. U. Efron, editor. *Spatial Light Modulator Technology*. Marcel Dekker, New York, NY, 1994.
103. H. M. A. El-Sum. *Reconstructed wavefront microscopy*. PhD thesis, Stanford University, Dept. of Physics, 1952.
104. L. H. Enloe, J. A. Murphy, and C. B. Rubinstein. Hologram transmission via television. *Bell Syst. Tech. J.*, 45:335–339, 1966.
105. W. F. Fagan, editor. *Holographic Optical Security Systems*, volume 1509, Bellingham, WA, 1991. S.P.I.E.
106. N. H. Farhat, D. Psaltis, A. Prata, and E. Paek. Optical implementation of the Hopfield model. *Appl. Opt.*, 24:1469–1475, 1985.
107. M. W. Farn. Binary optics. In R. Stern, editor, *Handbook of Photonics*. CRC Press, Boca Raton, FL, 1995.
108. M. W. Farn and J. W. Goodman. Diffractive doublets corrected at two wavelengths. *J. Opt. Soc. Am. A*, 8:860–867, 1991.
109. D. Feitelson. *Optical Computing*. MIT Press, Cambridge, MA, 1988.

110. H. A. Ferwerda. Frits Zernike—life and achievements. *Opt. Engin.*, 32:3176–3181, 1993.
111. J. R. Fienup. Reconstruction of an object from the modulus of its Fourier transform. *Opt. Lett.*, 3:27–29, 1978.
112. J. R. Fienup. Phase retrieval algorithms: a comparison. *Appl. Opt.*, 21:2758–2769, 1982.
113. J. R. Fienup. Phase retrieval algorithms: a personal tour. *Appl. Opt.*, 52:45–56, 2013.
114. J. R. Fienup and A. E. Tippie. Gigapixel synthetic-aperture digital holography. In *Tribute to Joseph W. Goodman*, volume 8122, page 812202. S.P.I.E., 2011.
115. G. Foo, D. M. Palacios, and G. A. Swartzlander. Optical vortex coronagraph. *Opt. Lett.*, 30:3308–3310, 2005.
116. R. L. Fork, B. I. Greene, and C. V. Shank. Generation of optical pulses shorter than 0.1 psec by colliding pulse mode locking. *Appl. Phys. Lett.*, 38:671–672, 1981.
117. M. Françon. *Modern Applications of Physical Optics*. John Wiley & Sons, New York, NY, 1963.
118. A. A. Friesem and J. S. Zelenka. Effects of film nonlinearities in holography. *Appl. Opt.*, 6:1755–1759, 1967.
119. C. Froehly, B. Colombeau, and M. Vampouille. Shaping and analysis of picosecond light pulses. In E. Wolf, editor, *Progress in Optics*, volume 20, pages 65–153. North Holland, Amsterdam, 1983.
120. D. Gabor. A new microscope principle. *Nature*, 161:777–778, 1948.
121. D. Gabor. Microscopy by reconstructed wavefronts. *Proc. Roy. Soc.*, A197:454–486, 1949.
122. D. Gabor. Microscopy by reconstructed wavefronts II. *Proc. Phys. Soc.*, B64:449–469, 1951.
123. D. Gabor. Associative holographic memories. *IBM J. Res. Dev.*, 13:156–159, 1969.
124. D. Gabor and W. P. Goss. Interference microscope with total wavefront reconstruction. *J. Opt. Soc. Am.*, 56:849–858, 1966.
125. D. Gabor, G. W. Stroke, R. Restrick, and A. Funkhouser. Optical image synthesis (complex addition and subtraction) by holographic Fourier transformation. *Phys. Lett.*, 18:116–118, 1965.
126. J. D. Gaskill. Imaging through a randomly inhomogeneous medium by wavefront reconstruction. *J. Opt. Soc. Am.*, 58:600–608, 1968.
127. J. D. Gaskill. Atmospheric degradation of holographic images. *J. Opt. Soc. Am.*, 59:308–318, 1969.
128. T. K. Gaylord and M. G. Moharam. Analysis and applications of optical diffraction by gratings. *Proc. I.E.E.E.*, 73:894–937, 1985.
129. S. A. Gerasimova and V. M. Zakharchenko. Holographic processor for associative information retrieval. *Soviet J. Opt. Techn.*, 48:404–406, 1981.
130. R. W. Gerchberg. Super-resolution through error energy reduction. *Optica Acta*, 21:709–720, 1974.
131. R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Addison-Wesley Publishing Company, Reading, MA, 1992.
132. J. W. Goodman. Some effects of target-induced scintillation on optical radar performance. *Proc. I.E.E.E.*, 53:1688–1700, 1965.

133. J. W. Goodman. Film grain noise in wavefront reconstruction imaging. *J. Opt. Soc. Am.*, 57:493–502, 1967.
134. J. W. Goodman. Noise in coherent optical processing. In Y. E. Nesterikhin, G. W. Stroke, and W. E. Kock, editors, *Optical Information Processing*. Plenum Press, New York, NY, 1976.
135. J. W. Goodman. *Statistical Optics*. John Wiley & Sons, New York, NY, second edition, 2015.
136. J. W. Goodman, A. R. Dias, and L. M. Woody. Fully parallel, high-speed incoherent optical method for performing discrete Fourier transforms. *Opt. Letter.*, 2:1–3, 1978.
137. J. W. Goodman, W. H. Huntley, Jr., D. W. Jackson, and M. Lehmann. Wavefront-reconstruction imaging through random media. *Appl. Phys. Lett.*, 8:311, 1966.
138. J. W. Goodman, D. W. Jackson, M. Lehmann, and J. Knotts. Experiments in long-distance holographic imagery. *Appl. Opt.*, 8:1581, 1969.
139. J. W. Goodman and G. R. Knight. Effects of film nonlinearities on wavefront-reconstruction images of diffuse objects. *J. Opt. Soc. Am.*, 58:1276–1283, 1968.
140. J. W. Goodman and R. W. Lawrence. Digital image formation from electronically detected holograms. *App. Phys. Lett.*, 11:77–79, 1967.
141. J. W. Goodman, R. B. Miles, and R. B. Kimball. Comparative noise performance of photographic emulsions in conventional and holographic imagery. *J. Opt. Soc. Am.*, 58:609–614, 1968.
142. J. W. Goodman and A. M. Silvestri. Some effects of Fourier domain phase quantization. *IBM J. Res. and Dev.*, 14:478–484, 1970.
143. J. W. Goodman and L. M. Woody. Method for performing complex-valued linear operations on complex-valued data using incoherent light. *Appl. Opt.*, 16:2611–2612, 1977.
144. F. Gori. Fresnel transform and sampling theorem. *Optics Commun.*, 39:293–297, 1981.
145. F. Gori, G. Guattari, and C. Padovani. Bessel-Gauss beams. *Optics Commun.*, 64:491–495, 1987.
146. R. M. Gray and J. W. Goodman. *Fourier Transforms: An Introduction for Engineers*. Kluwer Academic Publishers, Norwell, MA, 1995.
147. A. Greengard, Y. Y. Schechner, and R. Piestun. Depth from diffracted rotation. *Opt. Let.*, 31:181–183, 2006.
148. D. A. Gregory. Real-time pattern recognition using a modified liquid crystal television in a coherent optical correlator. *Appl. Opt.*, 25:467–469, 1986.
149. J. Grinberg, A. Jacobson, W. Bleha, L. Lewis, L. Fraas, D. Boswell, and G. Myer. A new real-time non-coherent to coherent light image converter: the hybrid field effect liquid crystal light valve. *Opt. Engin.*, 14:217–225, 1975.
150. S. Guha and G. D. Gillen. Description of light propagation through a circular aperture using nonparaxial vector diffraction theory. *Optics Express*, 13:1424–1447, 2005.
151. E. A. Guillemin. *The Mathematics of Circuit Analysis*. Principles of Electrical Engineering. John Wiley & Sons, Inc., New York, NY, 1965.
152. M. Guizar-Sicairos and J. R. Fienup. Understanding the twin-image problem in phase retrieval. *J. Opt. Soc. Am. A*, 29:2367–2375, 2012.
153. P. Günter and J.-P. Huignard, editors. *Photorefractive Materials and Their Applications I*. Springer-Verlag, Berlin, 1988.

154. P. Günter and J.-P. Huignard, editors. *Photorefractive Materials and Their Applications II*. Springer-Verlag, Berlin, 1989.
155. J. Guo, M. R. Gleeson, and J. T. Sheridan. A review of the optimisation of photopolymer materials for holographic data storage. *Phys. Research International*, 2012:1–16, 2012.
156. M. G. L. Gustafsson. Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy. *J. of Microscopy*, 198:82–87, 2000.
157. M. G. L. Gustafsson. Nonlinear structured-illumination microscopy: wide-field fluorescence imaging with theoretically unlimited resolution. *Proc. Nat. Acad. Sci.*, 102:13081–13086, 2005.
158. M. G. L. Gustafsson, J. W. Sedat, and D. A. Agard. Method and apparatus for three-dimensional microscopy with enhanced depth resolution, 1997. US Patent 5,671,085.
159. O. Guyon, C. Roddier, J. E. Graves, F. Roddier, S. Cuevas, C. Espejo, S. Gonzalez, A. Martinez, G. Bisiacchi, and V. Vuntesmeri. The nulling stellar coronagraph: laboratory tests and performance evaluation. *Pubs. Astron. Soc. Pacific*, 111:1321–1330, 1999.
160. P. Hariharan. *Optical Holography: Principles, Techniques & Applications*. Cambridge University Press, Cambridge, UK, second edition, 1999.
161. J. L. Harris. Diffraction and resolving power. *J. Opt. Soc. Am.*, 54:931, 1964.
162. J. F. Heane, M. C. Bashaw, and L. Hesselink. Volume holographic storage and retrieval of digital data. *Science*, 265:749–752, 1994.
163. Eugene Hecht. *Optics*. Addison-Wesley Publishing Company, Reading, MA, fourth edition, 2001.
164. R. Hecht-Nelson. *Neurocomputing*. Addison-Wesley, Reading, MA, 1980.
165. R. Heintzmann and C. Cremer. Laterally modulated excitation microscopy: improvement of resolution by using a diffraction grating. In *Proc. EUROPTO Conf. on Optical Microscopy*, volume 3568, pages 185–196, Bellingham, WA, 1998. S.P.I.E.
166. S. W. Hell and J. Wichmann. Breaking the diffraction resolution limit by stimulated emission: Stimulated-emission-depletion fluorescence microscopy. *Optics Lett.*, 19:780–782, 1994.
167. G. T. Herman. *Fundamentals of Computerized Tomography*. Springer-Verlag London, 2009.
168. J. C. Heurtley. Scalar Rayleigh-Sommerfeld and Kirchhoff diffraction integrals: a comparison of exact evaluations for axial points. *J. Opt. Soc. Am.*, 63:1003, 1973.
169. B. P. Hildebrand and K. A. Haines. Multiple-wavelength and multiple-source holography applied to contour generation. *J. Opt. Soc. Am.*, 57:155–162, 1967.
170. K. O. Hill, Y. Fujii, D. C. Johnson, and B. S. Kawasaki. Photosensitivity in optical fiber waveguides: application to reflection filter fabrication. *Appl. Phys. Lett.*, 32:647–649, 1978.
171. C. W. Hillegas, J. X. Tull, D. Goswami, D. Strickland, and W. S. Warren. Femtosecond laser pulse shaping by use of microsecond radio-frequency pulses. *Opt. Lett.*, 19:737–739, 1994.
172. H. A. Hoenl, A. W. Maue, and K. Westpfahl. Theorie der Beugung. In S. Fluegge, editor, *Handbuch der Physik*, volume 25. Springer-Verlag, Berlin, 1961.
173. H. H. Hopkins. *Wave Theory of Aberrations*. Oxford University Press, Oxford, 1950.
174. L. J. Hornbeck. Deformable-mirror spatial light modulators. *Proc. S.P.I.E.*, 1150:86–102, 1990.
175. J. Horner, editor. *Optical Signal Processing*. Academic Press, Inc., Orlando, FL, 1988.

176. J. R. Horner and J. R. Leger. Pattern recognition with binary phase-only filters. *Appl. Opt.*, 24:609–611, 1985.
177. R. G. Hunsperger. *Integrated Optics: Theory and Technology*. Springer-Verlag, Heidelberg, fifth edition, 2002.
178. A. L. Ingalls. The effects of film thickness variations on coherent light. *Phot. Sci. Eng.*, 4:135, 1960.
179. P. L. Jackson. Diffractive processing of geophysical data. *Appl. Opt.*, 4:419, 1965.
180. J. Jahns and S. H. Lee. *Optical Computing Hardware*. Academic Press, San Diego, CA, 1993.
181. B. Javidi, J. Ruiz, and C. Ruiz. Image enhancement by nonlinear signal processing. *Appl. Opt.*, 29:4812–4818, 1990.
182. B. M. Javidi. Nonlinear joint transform correlators. In B. Javidi and J. L. Horner, editors, *Real-Time Optical Information Processing*. Academic Press, San Diego, CA, 1994.
183. K. M. Johnson, D. J. McKnight, and I. Underwood. Smart spatial light modulators using liquid crystals on silicon. *I.E.E.E. J. Quant. Electr.*, 29:699–714, 1993.
184. S. G. Johnson and M. Frigo. A modified split-radix FFT with fewer arithmetic operations. *IEEE Trans. Signal Processing*, 55:111–119, 2007.
185. G. A. Swartzlander Jr. The optical vortex coronagraph. *J. Opt. A: Pure and Appl. Opt.*, 11:1–9, 2009.
186. T. Kailath. Channel characterization: Time varying dispersive channels. In E. J. Baghdady, editor, *Lectures on Communications System Theory*. McGraw-Hill Book Company, New York, NY, 1960.
187. I. Kaminow and T. Li. *Optical Fiber Telecommunications IVA: Components*. Academic Press, New York, NY, fourth edition, 2002.
188. I. Kaminow and T. Li. *Optical Fiber Telecommunications IVB: Systems and Impairments*. Academic Press, New York, NY, fourth edition, 2002.
189. E. Kaneko. *Liquid Crystal TV Displays: Principles and Applications of Liquid Crystal Displays*. KTK Scientific Publishers, Tokyo, Japan, 1987.
190. M. A. Karim and A. A. S. Awwal. *Optical Computing: an Introduction*. John Wiley & Sons, Inc., New York, NY, 1992.
191. J. B. Keller. Geometrical theory of diffraction. *J. Opt. Soc. Am.*, 52:116–130, 1962.
192. D. P. Kelly. Numerical calculation of the Fresnel transform. *J. Opt. Soc. Am. A*, 31:755–764, 2014.
193. S. N. Khonina, V. V. Kotlyar, V. A. Soifer, and M. Honkanan. Generation of rotating Gauss-Laguerre modes with binary-phase diffractive optics. *J. Mod. Opt.*, 46:227–238, 1999.
194. G. Kirchhoff. Zur Theorie der Lichtstrahlen. *Weidemann Ann. (2)*, 18:663, 1883.
195. T. A. Klar and S. W. Hell. Subdiffraction resolution in far-field fluorescence microscopy. *Optics Lett.*, 24:954–956, 1999.
196. M. V. Klein and T. E. Furtak. *Optics*. John Wiley & Sons, New York, NY, second edition, 1986.
197. G. Knight. Page-oriented associative holographic memory. *Appl. Opt.*, 13:904–912, 1974.
198. G. Knight. Holographic associative memory and processor. *Appl. Opt.*, 14:1088–1092, 1975.

199. G. R. Knight. Holographic memories. *Opt. Engin.*, 14:453–459, 1975.
200. C. Knox. Holographic microscopy as a technique for recording dynamic microscopic subjects. *Science*, 153:989, 1966.
201. H. Kogelnik. Holographic image projection through inhomogeneous media. *Bell Syst. Tech. J.*, 44:2451, 1965.
202. H. Kogelnik. Reconstructing response and efficiency of hologram gratings. In J. Fox, editor, *Proc. Symp. Modern Opt.*, pages 605–617. Polytechnic Press, Brooklyn, NY, 1967.
203. H. Kogelnik. Coupled wave theory for thick hologram gratings. *Bell Syst. Tech. J.*, 48:2909–2947, 1969.
204. B. Kolner. Generalization of the concepts of focal length and f-number to space and time. *J. Opt. Soc. Am. A*, 11:3229–3234, 1994.
205. A. Korpel. *Acousto-Optics*. Marcel Dekker, New York, NY, 1988.
206. Y. Kotlyar, V. A. Soifer, and S. N. Khonina. An algorithm for the generation of laser beams with longitudinal periodicity: rotating images. *J. Mod. Opt.*, 44:1409–1416, 1997.
207. F. Kottler. Electromagnetische Theorie der Beugung an schwarzen Schirmen. *Ann. Physik*, (4) 71:457, 1923.
208. F. Kottler. Zur Theorie der Beugung an schwarzen Schirmen. *Ann. Physik*, (4) 70:405, 1923.
209. F. Kottler. Diffraction at a black screen. In E. Wolf, editor, *Progress in Optics*, volume IV. North Holland Publishing Company, Amsterdam, 1965.
210. A. Kozma. Photographic recording of spatially modulated coherent light. *J. Opt. Soc. Am.*, 56:428, 1966.
211. A. Kozma. Effects of film grain noise in holography. *J. Opt. Soc. Am.*, 58:436–438, 1968.
212. A. Kozma, G. W. Jull, and K. O. Hill. An analytical and experimental study of nonlinearities in hologram recording. *Appl. Opt.*, 9:721–731, 1970.
213. A. Kozma and D. L. Kelly. Spatial filtering for detection of signals submerged in noise. *Appl. Opt.*, 4:387, 1965.
214. C. J. Kramer. Holographic laser scanners for nonimpact printing. *Laser Focus*, 17:70–82, 1981.
215. H. J. Landau and H. O. Pollak. Prolate spheroidal wave functions, Fourier analysis and uncertainty – II. *Bell Syst. Tech. J.*, 40:65–84, 1961.
216. S. H. Lee, editor. *Optical Information Processing—Fundamentals*. Springer-Verlag, Berlin, 1981.
217. W. H. Lee. Sampled Fourier transform hologram generated by computer. *Appl. Opt.*, 9:639–643, 1970.
218. W. H. Lee. Binary synthetic holograms. *Appl. Opt.*, 13:1677–1682, 1974.
219. W. H. Lee. Computer-generated holograms: techniques and applications. In E. Wolf, editor, *Progress in Optics*, volume 16, pages 121–232. North-Holland, Amsterdam, 1978.
220. W. H. Lee. Binary computer-generated holograms. *Appl. Opt.*, 18:3661–3669, 1979.
221. E. N. Leith. Photographic film as an element of a coherent optical system. *Phot. Sci. Eng.*, 6:75–80, 1962.

222. E. N. Leith and J. Upatnieks. Reconstructed wavefronts and communication theory. *J. Opt. Soc. Am.*, 52:1123–1130, 1962.
223. E. N. Leith and J. Upatnieks. Wavefront reconstruction with continuous-tone objects. *J. Opt. Soc. Am.*, 53:1377–1381, 1963.
224. E. N. Leith and J. Upatnieks. Wavefront reconstruction with diffused illumination and three-dimensional objects. *J. Opt. Soc. Am.*, 54:1295–1301, 1964.
225. E. N. Leith and J. Upatnieks. Holograms: their properties and uses. *S.P.I.E. J.*, 4:3–6, 1965.
226. L. B. Lesem, P. M. Hirsch, and J. A. Jordan, Jr. The kinoform: a new wavefront reconstruction device. *IBM J. Res. and Dev.*, 13:150–155, 1969.
227. M. J. Lighthill. *Introduction to Fourier Analysis and Generalized Functions*. Cambridge University Press, NY, 1960.
228. L. H. Lin. Hologram formation in hardened dichromated gelatin films. *Appl. Opt.*, 8:963–966, 1969.
229. E. H. Linfoot. *Fourier Methods in Optical Image Evaluation*. Focal Press, Ltd., London, 1964.
230. A. W. Lohmann. Optical single-sideband transmission applied to the Gabor microscope. *Opt. Acta*, 3:97, 1956.
231. A. W. Lohmann. Wavefront reconstruction for incoherent objects. *J. Opt. Soc. Am.*, 55:1555–1556, 1965.
232. A. W. Lohmann and D. P. Paris. Space-variant image formation. *J. Opt. Soc. Am.*, 55:1007–1013, 1965.
233. A. W. Lohmann and D. P. Paris. Binary Fraunhofer holograms generated by computer. *Appl. Opt.*, 6:1739–1748, 1967.
234. X. J. Lu, F. T. S. Yu, and D. A. Gregory. Comparison of Vander Lugt and joint transform correlators. *Appl. Phys. B*, 51:153–164, 1990.
235. W. Lukosz. Optical systems with resolving power exceeding the classical limit. *J. Opt. Soc. Am.*, 56:1463–1472, 1966.
236. W. Lukosz. Optical systems with resolving power exceeding the classical limit II. *J. Opt. Soc. Am.*, 57:932–941, 1967.
237. A. Macovski. Hologram information capacity. *J. Opt. Soc. Am.*, 60:21–29, 1970.
238. G. A. Maggi. Sulla propagazione libera e perturbata della onde luminose in un mezzo isotropo. *Ann. Mathematica*, 16:21–48, 1888.
239. A. N. Mahajan. *Optical Imaging and Aberrations (Part II): Wave Diffraction Optics*. S.P.I.E. Press, Bellingham, WA, 2001.
240. J. N. Mait and D. W. Prather, editors. *Selected Papers on Subwavelength Diffractive Optics*, volume 166. S.P.I.E. Press, 2001.
241. D. Mas, J. Garcia, C. Ferreira, L. Bernado, and F. Marinho. Fast algorithms for free-space diffraction patterns calculation. *Optics Commun.*, 164:233–245, 1999.
242. L. R. McAdams and A. M. Gerrish. High-speed lossless optical crossbar switch based on semiconductor optical amplifiers. In *Multigigabit Fiber Communications Systems*, volume 2024, pages 295–302, Bellingham, WA, 1993. S.P.I.E. Press.
243. A. D. McAulay. *Optical Computer Architectures*. John Wiley & Sons, Inc., New York, NY, 1991.

244. J. T. McCrickerd and N. George. Holographic stereogram from sequential component photographs. *Appl. Phys. Lett.*, 12:10–12, 1968.
245. D. J. McKnight, K. M. Johnson, and R. A. Serati. 256 x 256 liquid-crystal-on-silicon spatial light modulator. *Appl. Opt.*, 33:2775–2784, 1994.
246. I. McNutty et al. Experimental demonstration of high-resolution three-dimensional X-ray holography. *Proc. S.P.I.E.*, 1741:78–84, 1993.
247. C. E. K. Mees and T. H. James, editors. *The Theory of the Photographic Process*. The Macmillan Company, New York, NY, third edition, 1966.
248. R. W. Meier. Magnification and third-order aberrations in holography. *J. Opt. Soc. Am.*, 55:987–992, 1965.
249. A. J. Mendez, R. M. Gagliardi, V. J. Hernandez, C. V. Bennett, and W. J. Lennon. Design and performance analysis of wavelength/time (w/t) matrix codes for optical CDMA. *J. Lightwave Techn.*, 21:2524–2533, 2003.
250. J. Mertz. *Introduction to Optical Microscopy*. Roberts & Company, Greenwood Village, CO, 2010.
251. L. Mertz and N. O. Young. Fresnel transformations of images. In K. J. Habell, editor, *Proc. Conf. Optical Instruments and Techniques*, page 305, New York, NY, 1963. John Wiley and Sons.
252. D. Meyerhofer. Phase holograms in dichromated gelatin. *RCA Rev.*, 33:110–130, 1972.
253. M. Minsky. Memoir on inventing the confocal scanning microscope. *Scanning*, 10:128–138, 1988.
254. M. Mishali and Y. C. Eldar. From theory to practice: sub-Nyquist sampling of sparse wideband analog signals. *I.E.E.E. J. Selected Topics on Signal Processing*, 4:375–391, 2010.
255. M. Mishali, Y. C. Eldar, O. Dounaevsky, and E. Shoshan. Xampling: analog to digital at sub-Nyquist rates. *IET Circuits, Devices & Systems*, 5:8–20, 2011.
256. W. E. Moerner. Microscopy beyond the diffraction limit using actively controlled single molecules. *J. Microscopy*, 246:213–220, 2012.
257. W. E. Moerner and L. Kador. Optical detection and spectroscopy of single molecules in a solid. *Phys. Rev. Lett.*, 62:2535–2538, 1989.
258. F. H. Mok and H. M. Stoll. Holographic inner-product processor for pattern recognition. *Proc. SPIE*, 1701:312–322, 1992.
259. R. M. Montgomery. Acousto-optical signal processing system, 1972. U. S. Patent 3,634,749.
260. I. C. Moore and M. Cada. Prolate spheroidal wave functions, an introduction to the Slepian series and its properties. *Appl. Comput. Harmon. Anal.*, 16:208–230, 2004.
261. G. M. Morris and D. L. Zweig. White light Fourier transformations. In J. L. Horner, editor, *Optical Signal Processing*. Academic Press, Orlando, FL, 1987.
262. M. Murdocca. *A Digital Design Methodology for Optical Computing*. MIT Press, Cambridge, MA, 1990.
263. J. A. Neff, R. A. Athale, and S. H. Lee. Two-dimensional spatial light modulators: a tutorial. *Proc. I.E.E.E.*, 78:826–855, 1990.
264. R. Ng. *Digital Light Field Photography*. PhD thesis, Stanford University, Dept. of Computer Science, 2006.
265. R. J. Ober, S. Ram, and E. S. Ward. Localization accuracy in single-molecule microscopy. *Biophys. J.*, 86:1185–1200, 2004.

266. E. Ochoa, J. W. Goodman, and L. Hesselink. Real-time enhancement of defects in a periodic mask using photorefractive $\text{Bi}_{12}\text{SiO}_{20}\text{Bi}_{12}\text{SiO}_{20}$. *Opt. Lett.*, 10:430–432, 1985.
267. K. Okamoto. Silica waveguide devices. In E. J. Murphy, editor, *Integrated Optical Circuits and Components*. Marcel Dekker, New York, NY, 1999.
268. B. M. Oliver. Sparkling spots and random diffraction. *Proc. I.E.E.E.*, 51:220, 1963.
269. E. L. O'Neill. *Introduction to Statistical Optics*. Addison-Wesley Publishing Company, Inc., Reading, MA, 1963.
270. L. Onural. Sampling of the diffraction field. *Appl. Opt.*, 39:5929–5935, 2000.
271. D. C. O'Shea, T. J. Suleski, A. D. Kathman, and D. W. Prather. *Diffractive Optics: Design, Fabrication, and Test*, volume TT62. S.P.I.E., Bellingham, WA, 2003.
272. Y. Owechko and B. H. Soffer. Holographic neural networks based on multi-grating processes. In B. Javidi and J. L. Horner, editors, *Real-Time Optical Information Processing*. Academic Press, San Diego, CA, 1994.
273. H. M. Ozaktas and D. Mendlovic. Fractional Fourier optics. *J. Opt. Soc. Am. A*, 12:743–751, 1995.
274. H. M. Ozaktas, Z. Zalevsky, and M. A. Kutay. *The Fractional Fourier Transform*. John Wiley & Sons, 2001.
275. A. Papoulis. *The Fourier Integral and Its Applications*. McGraw-Hill Book Company, Inc., NY, 1962.
276. A. Papoulis. *Systems and Transforms with Applications to Optics*. McGraw-Hill Book Company, New York, NY, 1968.
277. A. Papoulis. A new algorithm in spectral analysis and band-limited extrapolation. *I.E.E.E. Trans. on Circuits and Systems*, CAS-22:735–742, 1975.
278. A. Papoulis. Pulse compression, fiber communications, and diffraction: a unified approach. *J. Opt. Soc. Am. A*, 11:3–13, 1994.
279. S. R. P. Pavani and R. Piestun. High-efficiency rotating point spread functions. *Optics Express*, 16:3484–3489, 2008.
280. S. R. P. Pavani, M. A. Thompson, S. Biteen, S. J. Lord, N. Liu, R. J. Twieg, R. Piestun, and W. E. Moerner. Three-dimensional single-molecule fluorescence imaging beyond the diffraction limit using a double-helix point spread function. *Proc. Nat. Acad. Sci.*, 106:2995–2999, 2009.
281. D. P. Peterson and D. Middleton. Sampling and reconstruction of wave-number-limited functions in N-dimensional space. *Information and Control*, 5:279–323, 1962.
282. M. Petran and M. Hadravsky. Tandem-scanning reflected-light microscope. *J. Opt. Soc. Am.*, 58:661–664, 1968.
283. R. Piestun, Y. Y. Schechner, and J. Shamir. Propagation-invariant wave fields with finite energy. *J. Opt. Soc. Am. A*, 17:294–303, 2000.
284. D. K. Pollack, C. J. Koester, and J. T. Tippett, editors. *Optical Processing of Information*. Spartan Books, Inc., Baltimore, MD, 1963.
285. D. A. Pommet, M. G. Moharam, and E. B. Grann. Limits of scalar diffraction theory for diffractive phase elements. *J. Opt. Soc. Am. A*, 11:1827–1834, 1994.
286. A. B. Porter. On the diffraction theory of microscope vision. *Phil. Mag. (6)*, 11:154–166, 1906.
287. R. L. Powell and K. A. Stetson. Interferometric vibration analysis by wavefront reconstruction. *J. Opt. Soc. Am.*, 55:1593–1598, 1965.

288. K. Preston, Jr. Use of the Fourier transformable properties of lenses for signal spectrum analysis. In J. T. Tippett et al., editors, *Optical and Electro-Optical Information Processing*. M.I.T. Press, Cambridge, MA, 1965.
289. P. Prucnal, M. A. Santoro, and T. R. Fan. Spread spectrum fiberoptic local area network using optical processing. *J. Lightwave Techn.*, 4:547–554, 1986.
290. D. Psaltis and N. Farhat. Optical information processing based on an associative-memory model of neural nets with thresholding and feedback. *Optics Lett.*, 10:98–100, 1985.
291. D. Psaltis, X. Gu, and D. Brady. Holographic implementations of neural networks. In S. F. Zornetzer, J.L. Davis, and C. Lau, editors, *An Introduction to Neural and Electronic Networks*. Academic Press, 1990.
292. S. I. Ragnarsson. A new holographic method of generating a high efficiency, extended range spatial filter with application to restoration of defocussed images. *Physica Scripta*, 2:145–153, 1970.
293. B. Rappaz, B. Breton, E. Shaffer, and G. Turcatti. Digital holographic microscopy: a quantitative label-free microscopy technique for phenotypic screening. *Comb. Chem. Throughput Screen.*, 17:80–88, 2014.
294. J. A. Ratcliffe. Aspects of diffraction theory and their application to the ionosphere. In A. C. Strickland, editor, *Reports on Progress in Physics*, volume XIX. The Physical Society, London, 1956.
295. Lord Rayleigh. On the theory of optical images, with special references to the microscope. *Phil. Mag. (5)*, 42:167–195, 1896.
296. J. D. Redman, W. P. Wolton, and E. Shuttleworth. Use of holography to make truly three-dimensional X-ray images. *Nature*, 220:58–60, 1968.
297. J. Rhodes. Analysis and synthesis of optical images. *Am. J. Phys.*, 21:337–343, 1953.
298. G. L. Rogers. Gabor diffraction microscope: the hologram as a generalized zone plate. *Nature*, 166:237, 1950.
299. G. L. Rogers. *Noncoherent Optical Processing*. John Wiley & Sons, New York, NY, 1977.
300. D. Rouan, P. Riaud, A. Boccaletti, Y. Clénet, and A. Labeyrie. The four-quadrant phase-mask coronograph. I. Principle. *Pubs. Astron. Soc. Pacific*, 112:1479–1486, 2000.
301. A. Rubinowicz. Die Beugungswelle in der Kirchhoffschen Theorie der Beugungserscheinungen. *Ann. Physik*, 53:257, 1917.
302. A. Rubinowicz. The Miyamoto-Wolf diffraction wave. In E. Wolf, editor, *Progress in Optics*, volume IV. North Holland Publishing Company, Amsterdam, 1965.
303. D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing*. MIT Press, Cambridge, MA, 1986.
304. C. K. Rushforth and R. W. Harris. Restoration, resolution and noise. *J. Opt. Soc. Am.*, 58:539–545, 1968.
305. B. E. A. Saleh and M. C. Teich. *Fundamentals of Photonics*. John Wiley & Sons, Inc., New York, NY, second edition, 2007.
306. H. P. Sardesai, C.-C. Chang, and A. M. Weiner. A femtosecond code-division multiple-access communication system test bed. *J. Lightwave Techn.*, 16:1953–1964, 1998.
307. G. Saxby. *Practical Holography*. Institute of Physics Publishing, Bristol, U.K., third edition, 2003.

308. O. H. Schade. Electro-optical characteristics of television systems. *RCA Review*, IX:5 (Part I), 245 (Part II), 490 (Part III), 653 (Part IV), 1948.
309. Y. Y. Schechner, R. Piestun, and J. Shamir. Wave propagation and rotating intensity distributions. *Phys. Rev. E*, 54:R50–R53, 1996.
310. J. Schmit and K. Creath. Extended averaging technique for derivation of error-compensating algorithms in phase-shifting interferometry. *Appl. Opt.*, 34:3610–3619, 1995.
311. W. Schumann. *Holography and Deformation Analysis*. Springer-Verlag, Berlin, 1985.
312. P. J. Sementilli, B. R. Hunt, and M. S. Nadar. Analysis of the limit to superresolution in incoherent imaging. *J. Opt. Soc. Am. A*, 10:2265–2276, 1993.
313. T. A. Shankoff. Phase holograms in dichromated gelatin. *Appl. Opt.*, 7:2101–2105, 1968.
314. C. E. Shannon. Communication in the presence of noise. *Proc. IRE*, 37:10, 1949.
315. G. C. Sherman. Application of the convolution theorem to Rayleigh’s integral formulas. *J. Opt. Soc. Am.*, 57:546, 1967.
316. A. E. Siegman. *Lasers*. University Science Books, Mill Valley, CA, 1986.
317. S. Silver. Microwave aperture antennas and diffraction theory. *J. Opt. Soc. Am.*, 52:131–139, 1962.
318. T. J. Skinner. Surface texture effects in coherent imaging. *J. Opt. Soc. Am.*, 53:1350A, 1963.
319. D. Slepian. Prolate spheroidal wave functions, Fourier analysis and uncertainty—IV. *Bell Syst. Tech. J.*, 43:3009–3057, 1964.
320. D. Slepian and H. O. Pollak. Prolate spheroidal wave functions, Fourier analysis and uncertainty—I. *Bell Syst. Tech. J.*, 40:43–63, 1961.
321. L. Slobodin. Optical correlation techniques. *Proc. I.E.E.E.*, 51:1782, 1963.
322. H. M. Smith. *Principles of Holography*. John Wiley & Sons, New York, NY, second edition, 1975.
323. H. M. Smith, editor. *Holographic Recording Materials*. Springer-Verlag, Berlin, 1977.
324. L. Solymar and D. J. Cook. *Volume Holography and Volume Gratings*. Academic Press, New York, NY, 1981.
325. M. Somayaji and M. P. Christensen. Enhancing form factor and light collection of multiplex imaging systems by using a cubic phase mask. *Appl. Opt.*, pages 2911–2923, 2006.
326. A. Sommerfeld. Mathematische Theorie der Diffraction. *Math. Ann.*, 47:317, 1896.
327. A. Sommerfeld. Die Greensche Funktion der Schwingungsgleichung. *Jahresber. Deut. Math. Ver.*, 21:309, 1912.
328. A. Sommerfeld. *Optics*, volume IV of *Lectures on Theoretical Physics*. Academic Press, Inc., NY, 1954.
329. W. H. Southwell. Validity of the Fresnel approximation in the near field. *J. Opt. Soc. Am.*, 71:7–14, 1981.
330. R. A. Sprague and C. L. Koliopoulos. Time integrating acousto-optic correlator. *Appl. Opt.*, 15:89–92, 1975.
331. T. Stone and N. George. Hybrid diffractive-refractive lenses and achromats. *Appl. Opt.*, 27:2960–2971, 1988.
332. T. A. Strasser and T. Erdogan. Fiber grating devices in high-performance optical communications systems. In I. Kaminow and T. Li, editors, *Optical Fiber*

- Telecommunications IVA: Components*. Academic Press, New York, NY, fourth edition, 2002.
- 333. G. W. Stroke. Image deblurring and aperture synthesis using a posteriori processing by Fourier-transform holography. *Optica Acta*, 16:401–422, 1969.
 - 334. G. W. Stroke and A. T. Funkhouser. Fourier-transform spectroscopy using holographic imaging without computing and with stationary interferometers. *Phys. Lett.*, 16:272–274, 1965.
 - 335. G. W. Stroke and R. C. Restrick III. Holography with spatially incoherent light. *Appl. Phys. Lett.*, 7:229–231, 1965.
 - 336. K. J. Strozewski, C-Y. Wang, Jr. G. C. Wetsel, R. M. Boysel, and J. M. Florence. Characterization of a micromechanical spatial light modulator. *J. Appl. Phys.*, 73:7125–7128, 1993.
 - 337. Royal Swedish Academy of Sciences. Super-resolved fluorescence microscopy. *Scientific Background on the Nobel Prize in Chemistry*, 2014.
 - 338. M. R. Taghizadeh and J. Turunen. Synthetic diffractive elements for optical interconnection. *Opt. Comp. and Proc.*, 2:221–242, 1992.
 - 339. H. Takahashi, S. Suzuki, K. Kato, and I. Nishi. Arrayed-waveguide grating for wavelength division multi/demultiplexer with nanometer resolution. *Electron. Lett.*, 26:87–88, 1990.
 - 340. H. F. Talbot. Facts relating to optical science, no. IV. *Philos. Mag.*, 9:401–407, 1836.
 - 341. B. J. Thompson and J. H. Ward and W. R. Zinky. Application of hologram techniques for particle size analysis. *Appl. Opt.*, 6:519–526, 1967.
 - 342. D. A. Tichenor and J. W. Goodman. Coherent transfer function. *J. Opt. Soc. Am.*, 62:293–295, 1972.
 - 343. D. A. Tichenor and J. W. Goodman. Restored impulse response of finite-range image deblurring filter. *Appl. Optics*, 14:1059–1060, 1975.
 - 344. J. T. Tippett et al., editors. *Optical and Electro-optical Information Processing*. MIT Press, Cambridge, MA, 1965.
 - 345. V. Toal. *Introduction to Holography*. CRC Press, Boca Raton, FL, 2012.
 - 346. A. Tonomura. *Electron Holography*. Springer-Verlag, Berlin, 1994.
 - 347. G. Toraldo di Francia. Super-gain antennas and optical resolving power. *Nuovo Cimento*, IX:426–438, 1952.
 - 348. G. Toraldo di Francia. Degrees of freedom of an image. *J. Opt. Soc. Am.*, 59:799–804, 1969.
 - 349. W. A. Traub and B. R. Oppenheimer. Direct imaging of exoplanets. In Sara Seager, editor, *Exoplanets*, Space Science Series, pages 111–156. University of Arizona Press, 2010.
 - 350. G. L. Turin. An introduction to matched filters. *IRE Trans. Info. Theory*, IT-6:311–329, 1960.
 - 351. J. Upatnieks, A. Vander Lugt, and E. Leith. Correction of lens aberrations by means of holograms. *Appl. Opt.*, 5:589–593, 1966.
 - 352. R. F. van Ligten. Influence of photographic film on wavefront reconstruction. I: Plane wavefronts. *J. Opt. Soc. Am.*, 56:1–9, 1966.
 - 353. R. F. van Ligten. Influence of photographic film on wavefront reconstruction. II: ‘Cylindrical’ wavefronts. *J. Opt. Soc. Am.*, 56:1009–1014, 1966.

354. A. B. VanderLugt. Signal detection by complex spatial filtering. Technical report, Institute of Science and Technology, University of Michigan, Ann Arbor, MI, 1963.
355. A. B. VanderLugt. Signal detection by complex spatial filtering. *I.E.E.E. Trans. Info. Theory*, IT-10:139–145, 1964.
356. A. B. VanderLugt. *Optical Signal Processing*. John Wiley & Sons, Inc., New York, NY, 1992.
357. C. M. Vest. *Holographic Interferometry*. John Wiley & Sons, New York, NY, 1979.
358. D. Voelz. *Computational Fourier Optics: A Matlab Tutorial*. SPIE Press, 2011.
359. D. Voelz and M. Roggemann. Digital simulation of scalar optical diffraction: revisiting chirp function sampling criteria and consequences. *Appl. Opt.*, 48:6132–6142, 2009.
360. C. J. Weaver and J. W. Goodman. A technique for optically convolving two functions. *Appl. Opt.*, 5:1248–1249, 1966.
361. A. M. Weiner. Femtosecond Fourier optics: shaping and processing of ultrashort optical pulses. In T. Asakura, editor, *International Trends in Optics and Photonics—ICO IV*, pages 233–246. Springer-Verlag, Heidelberg, 1999.
362. A. M. Weiner. Femtosecond pulse shaping using spatial light modulators. *Review of Scientific Instruments*, 71:1929–1960, 2000.
363. A. M. Weiner and J. P. Heritage. Picosecond and femtosecond Fourier pulse shape synthesis. *Phys. Rev. Appl.*, 22:1619, 1987.
364. A. M. Weiner, D. E. Leaird, J. S. Patel, and J. R. Wullert. Programmable shaping of femtosecond optical pulses by use of a 128-element liquid crystal phase modulator. *IEEE J. Quant. Electron.*, 28:908–920, 1992.
365. A. M. Weiner, D. E. Leaird, D. H. Reitze, and E. G. Paek. Spectral holography of shaped femtosecond pulses. *Opt. Lett.*, 17:224–226, 1992.
366. W. T. Welford. *Aberrations of the Symmetrical Optical System*. Academic Press, Inc., New York, NY, 1974.
367. B. S. Wherrett and F. A. P. Tooley, editors. *Optical Computing*. Edinburgh University Press, Edinburgh, UK, 1989.
368. E. T. Whittaker. On the functions which are represented by the expansions of the interpolation theory. *Proc. Roy. Soc. Edinburgh, Sect. A*, 35:181–194, 1915.
369. B. Widrow and S. D. Stearns. *Adaptive Signal Processing*. Prentice-Hall, Englewood Cliffs, NJ, 1985.
370. E. Wigner. On the quantum correction for thermodynamic equilibrium. *Phys. Rev.*, 40:749–759, 1932.
371. J. P. Wilde, J. W. Goodman, Y. C. Eldar, and Y. Takashima. Grating-enhanced coherent imaging. In *Novel Techniques in Microscopy*, OSA Technical Digest, page NMA3. Optical Society of America, 2011.
372. J. P. Wilde, J. W. Goodman, Y. C. Eldar, and Y. Takashima. Grating-enhanced coherent imaging. In *9th International Conference on Sampling Theory (SampTA)*, page P0213, 2011.
373. C. S. Williams and O. A. Becklund. *Introduction to the Optical Transfer Function*. John Wiley & Sons, New York, NY, 1989.
374. T. Wilson and C. J. R. Sheppard. *Theory and Practice of Scanning Optical Microscopy*. Academic Press, London, 1984.
375. E. Wolf. Is a complete determination of the energy spectrum of light possible from measurements of the degree of coherence? *Proc. Phys. Soc.*, 80:1269–1272, 1962.

376. E. Wolf. *Introduction to the Theory of Coherence and Polarization of Light*. Cambridge University Press, Cambridge, UK, 2007.
377. E. Wolf and E. W. Marchand. Comparison of the Kirchhoff and Rayleigh-Sommerfeld theories of diffraction at an aperture. *J. Opt. Soc. Am.*, 54:587–594, 1964.
378. I. Yamaguchi and T. Zhang. Phase-shifting digital holography. *Opt. Lett.*, 22:1268–1270, 1997.
379. L. P. Yaroslavskii and N. S. Merzlyakov. *Methods of Digital Holography*. Consultants Bureau, Plenum Publishing Company, New York, NY, 1980.
380. L. H. Yeh, J. Dong, J. Zhong, L. Tian, M. Chen, G. Tang, M. Soltanolkotabi, and L. Waller. Experimental robustness of Fourier ptychography phase retrieval algorithms. *Optics Express*, 23:33214–33240, 2015.
381. F. Zernike. Das Phasenkontrastverfahren bei der Mikroskopischen beobachtung. *Z. Tech. Phys.*, 16:454, 1935.
382. T. Zhang and I. Yamaguchi. Three-dimensional microscopy with phase-shifting digital holography. *Opt. Lett.*, 23:1221–1223, 1998.
383. Z. Zhang and M. Levoy. Wigner distributions and how they relate to the light field. In *2009 IEEE Conference on Computational Photography*, pages 1–10. I.E.E.E., 2009.
384. G. Zheng. *Fourier Ptychographic Imaging: A MATLAB Tutorial*. IOP Concise Physics. Morgan & Claypool Publishers, San Raphael, CA, 2016.
385. G. Zheng, R. Horstmeyer, and C. Yang. Wide-field high-resolution Fourier ptychographic microscopy. *Nature Photonics*, 7:739–745, 2013.

Index

- Abbe, Ernst, [185](#), [187–188](#), [310](#)
- Abbe-Porter experiments, [310–312](#)
- Aberrations, [160](#), [187](#), [204](#)
 - effects on frequency response, [204–214](#)
 - effects on the amplitude transfer function, [206](#)
 - effects on the optical transfer function, [206–207](#)
 - example of simple, [207–211](#)
- Absorption, [321](#)
- Acoustic-wave propagation, [43](#)
- Acousto-optic cells, [269](#)
- Acousto-optic signal processing systems, [341–347](#)
 - Bragg cell spectrum analyzer, [342–343](#)
 - space-integrating correlator, [343–345](#)
 - time-integrating correlator in, [345–347](#)
- Acousto-optic spatial light modulators, [304–308](#)
- Add/drop multiplexers, narrowband filters for, [461–462](#)
- Adjacency effect, [278](#)
- Agfa-Gaevert, [408](#)
- Airy, G. B., [92](#)
- Airy pattern, [92](#)
- Aliasing, [33](#), [34](#), [124](#), [131](#)
- Amplitude distribution, [254](#)
- Amplitude transfer function, [194–195](#)
 - effects of aberrations on, [206](#)
 - examples of, [195–197](#)
- Amplitude transmittance, [71](#), [274](#), [354](#)
- Analog optical information processing, [2](#), [309–356](#)
 - acousto-optic signal processing systems in, [341–347](#)
 - application to character recognition, [329–334](#)
 - coherent optical information processing systems and, [316–321](#)
 - discrete analog optical processors in, [347–351](#)
 - historical background in, [310–316](#)
 - image restoration in, [335–341](#)
 - joint transform correlator in, [326–329](#)
 - VanderLugt filter in, [321–326](#)
- Analytic continuation, [249](#)

• Anamorphic processors, [319](#)

• Angular spectrum of plane waves, [2](#), [68–73](#)

- effects of diffracting aperture on, [71–72](#)
- Fresnel approximation and, [83–84](#)
- physical interpretation of, [68–69](#)
- propagation of the, [69–71](#)

• Anisotropic materials, [293](#)

• Apertures

- circular, [91–93](#)
 - Fresnel diffraction by, [102–104](#)
- circularly symmetric, [150–151](#)
- Fresnel diffraction by square, [99–102](#)
- rectangular, [90–91](#), [148–149](#)
- two-dimensional, separable in (x,y) coordinates, [149](#)

• Apodization, [211–212](#)

- effects on frequency response, [211–214](#)
- Nuttall, [266–267](#)
- for starlight suppression, [242](#), [244–248](#)

• Approximation by a stepped thickness function, [283–285](#)

• Arago, F., [46](#)

• Armstrong, E. H., [314](#)

• Arrayed waveguide gratings (AWGs), [474–484](#)

- applications of, [481–484](#)
- component parts of, [475–481](#)

• Art, holographic, [447](#)

• Artificial neural networks, holographic weights for, [443–447](#)

• Associated Laguerre polynomial, [112](#)

• Autocorrelations theorem, [10](#), [491](#)

• Axial magnifications, [378–379](#)

• Babinet's principle, [104n](#)

• Back focal plane, [161](#)

• Bandlimited functions, [28–29](#)

• Bandwidth extrapolation, [249–250](#)

• Beam optics, [108–114](#)

• Beam ratio, [340](#)

• Bell Laboratories, [410](#)

• Benton, S. A., [357](#)

• Besinc function, [16](#)

• Bessel beams, [113–114](#)

• Binary optics, [283](#)

• Birefringence, [293](#)

• Blazed grating, [465n](#)

• Bleaching, [274](#)

- of photographic emulsions, [279–280](#)

► **Boundaries**

- diffraction at, [67–68](#)
- reproduction of conditions, [62–63](#)

► **Boundary values**, [54](#)

► **Bove, V. M.**, [357](#)

► **Bracewell, R. N.**, [6](#), [7](#), [32](#)

► **Bragg, W. L.**, [357](#)

► **Bragg angle**, [305](#)

► **Bragg cells**, [341](#)

- spectrum analyzer, [342–343](#), [356](#)

► **Bragg condition**, [384](#), [393](#)

► **Bragg effect**, [305](#), [391](#), [397](#)

► **Bragg matching condition**, [400–401](#), [406](#)

► **Bragg regime**, [304](#), [306](#)

► **Broadening the impulse response**, [170](#)

► **Brooks, R. E.**, [433](#)

► **Bruning, J. H.**, [427](#)

► **Burckhardt, C. B.**, [357](#)

► **Carré, P.**, [427](#)

► **Carrier frequency**, [324–325](#)

► **Cathey, W. T.**, [233](#)

► **Cat's eye mask**, [245](#), [246](#)

► **Caulfield, H. J.**, [357](#)

► **Central dark ground method**, [313](#), [352](#)

► **Changeable polarization**, [294](#)

► **Characteristic impedance**, [76](#)

► **Character recognition**, [309](#)

- in analog optical information processing, [329–334](#)
- matched filter and, [329–330](#)
- optical synthesis of, [332–334](#)
- sensitivity to scale size and rotation, [334](#)

► **Charge coupled device detectors**, [269](#)

► **Chemical bleaching**, [279](#)

► **Chemical diffusion**, [276](#)

► **Cholesteric liquid crystals**, [289](#)

► **Circle function**, [15](#), [17](#)

► **Circuit theory**, [49](#)

► **Circular aperture**, [91–93](#)

- Fresnel diffraction by a, [102–104](#)

► **Circular shift**, [36n](#)

► **Circular symmetry**, [12–14](#)

- apertures and, [150–151](#)

- ◀ Classical diffraction limit, resolution beyond the, [248–263](#)
- ◀ CMOS detectors, [269](#)
- ◀ Cochran, G., [428, 429](#)
- ◀ Code division multiple access, [467–469](#)
- ◀ Coherent imaging, comparison with incoherent imaging, [214–221](#)
- ◀ Coherent optical information processing systems
 - coherent system architectures, [316–319](#)
 - constraints on filter realization, [319–321](#)
- ◀ Coherent optics, application to data processing, [316](#)
- ◀ Coherent spectral multiplexing, [2, 253–258](#)
- ◀ Colburn, W. S., [409](#)
- ◀ Collier, R. J., [357](#)
- ◀ Comb function, [15](#)
- ◀ Communication systems, [1](#)
- ◀ Compensating filter, [314](#)
- ◀ Complex amplitude, [193](#)
- ◀ Complex-exponential functions, [28](#)
- ◀ Complex-valued field amplitude, [6](#)
- ◀ Computational diffraction and propagation, [121–154](#)
 - approaches to, [121–122](#)
 - comparison of computational complexities in, [145–148](#)
 - convolution approach to, [121, 125–130](#)
 - exact transfer function approach in, [122, 140–144, 153](#)
 - extension to more complex apertures, [148–152](#)
 - Fresnel transfer approach in, [122, 130–140, 152–153](#)
 - sampling a space-limited quadratic-phase exponential in, [122–125](#)
- ◀ Computer-generated holograms, [412–421](#)
- ◀ Computer-generated holographic optical elements, [286–287](#)
- ◀ Confocal microscopy, [221–225](#)
- ◀ Confocal spherical surfaces, Fresnel diffraction between, [84–85](#)
- ◀ Conjugate planes, [501–502](#)
- ◀ Contour generation, [435–437](#)
- ◀ Contrast reversal, [106](#)
- ◀ Convex lens, [160, 161](#)
- ◀ Convolution, [27, 121–122, 125–130, 145](#)
 - bandwidth and sampling considerations, [125–127](#)
 - discrete convolution equations, [127–128](#)
 - by Fourier transforms, [129–130](#)
 - simulation results, [128–129](#)
 - theorem of, [27](#)
- ◀ Convolution integral, [66](#)
- ◀ Convolution property, [22](#)
- ◀ Convolution relation, [27](#)
- ◀ Convolution theorem, [10, 71, 491](#)
- ◀ Cornu's spiral, [100n](#)

•Corpuscular theory, of light propagation, [45](#)
•Coupled mode theory, [398–407](#)
•Creath, K., [427](#)
•Cubic phase mask, [233–237](#)
 ◦ for increased depth of field, [231–237](#)
•Cutrona, L. J., [316](#)
•CW drive voltage, [304](#)
•CW lasers, [411](#)

•Data processing, application of coherent optics to, [316](#)
•Data storage, holographic, [442–443](#)
•Decomposition, Fourier transform as a, [8–9](#)
•Deformable mirror devices, [301](#)
•Deformable mirror spatial light modulators, [301–303](#)
•Degradations of holographic images, [421–426](#)
•Delta functions, [62, 487–489](#)
•Demultiplexers, [481](#)
•Denisyuk, Y. N., [358](#)
•Depth of field, [233](#)
 ◦ cubic phase mask for increased, [231–237](#)
•Depth of focus of an imaging system, [231–233](#)
•Depth resolution, rotating point spread functions for, [237–240](#)
•Detour-phase holograms, [413–418](#)
•Develis, J. B., [357](#)
•Development speck, [271](#)
•Dichromated gelatin films, [308, 409](#)
•Diffraction aperture, effects of, on the angular spectrum, [71–72](#)
•Diffraction, [2, 43](#). See also [Computational diffraction and propagation](#)
 ◦ at boundaries, [67–68](#)
 ◦ defined, [44](#)
 ◦ effects of, on an image, [187–189](#)
 ◦ Fresnel-Kirchhoff formula of, [57–58](#)
 ◦ Kirchoff formulation of, by a planar screen, [54–58](#)
 ◦ Rayleigh-Sommerfeld formulation of, [58–63](#)
•Diffraction efficiency, [95, 398–407](#)
 ◦ general method for calculating for gratings, [98–99](#)
•Diffraction-limited coherent imaging, frequency response for, [194–197](#)
•Diffraction-limited imaging system, [187](#)
•Diffraction-limited optical transfer function, [201–204](#)
•Diffraction-pattern calculations, [2](#)
•Diffractive optics, [2, 280](#)
 ◦ types of, [287](#)
 ◦ wavefront modulation with, [280–287](#)
•Digital holography, [3, 426–428](#)

- Digitally computing diffraction patterns, [2](#)
- Dirac delta function, [7](#), [8](#)
- Discrete analog optical processors, [347–351](#)
 - discrete representation of signals and systems, [347–348](#)
 - handling of bipolar and complex data, [350–351](#)
 - parallel incoherent matrix-vector multilier in, [348–350](#)
- Discrete convolution equations, [127–128](#)
- Discrete diffraction formulas, [132–133](#), [137–138](#)
- Discrete Fourier transform, [35–37](#)
- Dispersion, [456](#)
- Displays, holographic, [447](#)
- Dowski, E. R. Jr., [233](#)
- Dragone, C., [474](#), [476](#)
- Driffield, V. C., [272](#)
- Dudgeon, D. E., [32](#)
- Duffieux, P. M., [185](#)
- Dynamic devices, distinguishing between fixed masks and, [269](#)

- Eigenfunction, [28](#)
- Eigenvalue, [28](#)
- Eikonal equation, [497](#)
- Eikonal function, [496](#)
- Electrically driven liquid crystal spatial light modulators, [297–298](#)
- Electric field, [411](#)
- Elementary ray-transfer matrices, [499–501](#)
- El-Sum, H. M. A., [358](#)
- Embossed holograms, [389–390](#)
- Energy spectrum, [164](#)
- Equivalent area, [40](#)
- Equivalent bandwidth, [40](#), [123](#)
- Evanescent waves, [70–71](#)
- Exact transfer function approach, [122](#), [140–144](#), [147–148](#), [153](#)
 - computational complexity of the, [144](#)
 - samping in the frequency domain, [140–141](#)
 - sampling in the space domain, [141–143](#)
 - simulation results in, [143](#)
- Existence conditions, [7](#)
- Exoplanet discovery, point-spread function engineering for, [241–248](#)
- Exposure, [272](#)

- Fabrication, [286–287](#)
- Fast Fourier tranform (FFT) algorithms, [36–37](#), [309](#)
- Ferroelectric liquid crystals, [290](#)

- optical properties of, [293–296](#)
 - spatial light modulators and, [301](#)
- Fiber Bragg gratings, [3, 454–463](#)
 - applications of, [461–462](#)
- Fiber dispersion compensation, [469](#)
- Fienup, J. R., [38, 39](#)
- Film-grain noise, effects of, [425](#)
- Film nonlinearities, effects of, [424](#)
- Filter realization, constraints on, [319–321](#)
- Finite chirp function, [19](#)
- Finite integral of the quadratic-phase exponential function, [81–83](#)
- Finite-length quadratic-phase exponential, [126n](#)
- First orders, [95](#)
- First Rayleigh-Sommerfeld solution, [59, 73](#)
- Fixed masks, distinguishing between dynamic devices and, [269](#)
- Fluorescence labelling, [260–261](#)
- Fluorescence microscopy, [2](#)
- F-number, [229](#)
- Focal length, [158](#)
- Focal planes, [502–503](#)
- Fourier analysis, vii
 - as decomposition, [8–9](#)
 - definition and existence conditions, [6–8](#)
 - Fourier-Bessel transforms in, [12–14](#)
 - Fourier transform pairs in, [14–17](#)
 - Fourier transform theorems in, [9–11](#)
 - separable functions in, [11–12](#)
- Fourier-Bessel transforms, [12–14](#)
- Fourier decomposition, [497](#)
- Fourier integral, [6–7](#)
 - theorem, [11, 492–494](#)
- Fourier magnitude, phase retrieval from, [38–39](#)
- Fourier optics, [3, 453–484](#)
 - arrayed waveguide gratings, [474–484](#)
 - Fiber Bragg gratings, [454–463](#)
 - spectral holography, [469–474](#)
 - ultrashort pulse shaping processing, [463–469](#)
- Fourier ptychography, [2, 252–253](#)
- Fourier spectrum, [6](#)
- Fourier synthetic-aperture approach, [254](#)
- Fourier theory, [6](#)
- Fourier transform algorithm, [412](#)
- Fourier transform holograms, [422–423](#)
- Fourier transforming properties of lenses, [161–167](#)
- Fourier transforms, [6, 23, 68, 79](#)

- convolution by, [129–130](#)
 - discrete, [35–37](#)
 - fractional, [79n](#)
 - generalized, [2](#)
 - inverse, [6, 9](#)
 - one-dimensional, [12](#)
 - pairs, [14–17](#)
 - circle function, [15, 17](#)
 - comb function, [15](#)
 - rectangle function, [14](#)
 - signum function, [14](#)
 - sine function, [14](#)
 - triangle function, [15](#)
 - sampling increments and, [130](#)
 - sampling ratio Q , [131–132](#)
 - theorems, [9–11, 489–494](#)
 - autocorrelations theorem, [10](#)
 - convolution theorem, [10](#)
 - Fourier integral theorem, [11](#)
 - linearity theorem, [9](#)
 - Rayleigh’s theorem, [10](#)
 - rotation theorem, [10–11](#)
 - shear theorem, [11](#)
 - shift theorem, [10](#)
 - similarity theorem, [10](#)
 - two-dimensional, [8](#)
- Francia, Toraldo di, [248](#)
- Fraunhofer approximation, [2, 88–89](#)
- Fraunhofer diffraction, [89, 163](#)
 - patterns in, [90–99](#)
- Free space, [51](#)
- Free-space propagation, [2](#)
- Free spectral range, [480](#)
- Frequencies, [6](#)
- Frequency analysis, [1, 2](#)
 - of optical imaging systems, [185–230](#)
 - aberrations and their effects on frequency response, [204–214](#)
 - comparison of coherent and incoherent imaging, [214–221](#)
 - confocal microscopy, [221–225](#)
 - frequency response for diffraction-limited coherent imaging, [194–197](#)
 - frequency response for diffraction-limited incoherent imaging, [197–204](#)
 - generalized treatment, [186–194](#)
- Frequency domain, [28](#)
- Frequency-plane masks, [319](#)
 - synthesis of, [321–324](#)

Frequency response

- o for diffraction-limited coherent imaging, [194–197](#)
- o for diffraction-limited incoherent imaging, [197–204](#)
- o effects of aberrations on, [204–214](#)
- o effects of apodization on, [211–214](#)

Frequency spectrum, [6](#)

- o of the image intensity, [215–216](#)

Fresnel, Augustin Jean, [45–46, 58](#)

Fresnel approximation, [2, 78–88](#)

- o accuracy of, [80–81](#)
- o angular spectrum and, [83–84](#)
- o finite integral of the quadratic-phase exponential function, [81–83](#)
- o Fresnel diffraction between confocal spherical surfaces and, [84–85](#)
- o positive vs. negative phrases, [80](#)

Fresnel diffraction, [99–108](#)

- o by a circular aperture, [102–104](#)
- o between confocal spherical surfaces, [84–85](#)
- o by a sinusoidal amplitude grating, [104–108](#)
- o by a square aperture, [99](#)
- o in terms of ray transfer matrices, [85–88](#)

Fresnel diffraction integral, [79](#)

- o Fourier-transform version of the, [85](#)

Fresnel integrals, [100, 235](#)

Fresnel-Kirchoff diffraction formula, [57–58](#)

Fresnel number, [19, 100, 126n](#)

Fresnel transfer approach, [122, 130–140, 145, 146–147, 152–153](#)

- o computational complexity of, [140](#)
- o steps using, [138–139](#)

Fresnel transforms

- o computational complexity of, [135](#)
- o steps using, [133–135](#)

Fresnel zone plate, [182](#)

Friesem, A. A., [424](#)

Fringe orientations for more complex recording geometries, [394–396](#)

Gabor, Dennis, [3, 357–358, 363, 424, 443](#)

Gabor hologram, [363–366](#)

- o limitations of, [365–366](#)

Gamma of the emulsion, [273](#)

Gating amplitude transmittance, [254](#)

Gaussian beams, [108–110](#)

Gaussian function, [23](#)

Gaussian illumination beam, [257](#)

Generalized Fourier transform, [7](#)

- ◀ Generalized Laguerre polynomial, [112](#)
- ◀ Generalized pupil function, [205](#)
- ◀ Geometrical optics, [495–497](#)
 - matrix theory of paraxial, [155](#)
- ◀ Geometrical theory of diffraction, [68](#)
- ◀ Goodman, J. W., [6](#), [326](#), [341](#), [424](#)
- ◀ Gouy phase, [109](#)
- ◀ Grating equation, [515–516](#)
- ◀ Gratings, general method for calculating diffraction efficiency of, [98–99](#)
- ◀ Gray, R. M., [6](#)
- ◀ Green's functions, alternative, [59–61](#)
- ◀ Green's theorem, [51](#), [52](#), [53](#)
- ◀ Grimaldi, [44](#)
- ◀ Gross fog, [273](#)
- ◀ Guoy phase, [111](#), [112](#)
- ◀ Gustafsson, M. G. I., [258](#)

- ◀ Hadamard matrix, [256](#)
- ◀ Haines, K. A., [409](#)
- ◀ Halftone process, [315](#)
- ◀ Hankel transforms, [12](#)
 - of zero order, [14](#)
- ◀ Hard clipped filter, [321n](#)
- ◀ Hariharan, P., [357](#)
- ◀ Hell, Stephan, [261](#)
- ◀ Helmholtz, H. von
 - integral theorem of, [51–54](#)
 - reciprocity theorem of, [58](#)
- ◀ Helmholtz equation, [50–51](#)
- ◀ Hermite-Gaussian beams, [111](#)
- ◀ Hermite-Gaussian functions, [111](#)
- ◀ Hermite-Gaussian mode amplitudes, [111](#)
- ◀ Hermitian symmetry, [199](#)
- ◀ Heurtley, J. C., [64](#)
- ◀ High-frequency carrier, [322](#)
- ◀ High-resolution volume imagery, [431](#)
 - microscopy and, [431–432](#)
- ◀ Hill, K. O., [454](#)
- ◀ Holograms, [357](#)
 - computer-generated, [412–421](#)
 - detour-phase, [413–418](#)
 - embossed, [389–390](#)
 - Fourier, [381](#)
 - Fourier transform, [422–423](#)

- Fraunhofer, [381](#)
- Fresnel, [381](#), [382](#)
- Gabor, [363](#)–[366](#)
- image, [381](#)
- Leith-Upatnieks, [366](#)–[375](#)
- lensless Fourier transform, [381](#)–[382](#), [422](#)–[423](#)
- multiplex, [387](#)–[389](#)
- rainbow, [385](#)–[387](#)
- referenceless on-axis complex, [419](#)
- reflection, [382](#)–[383](#), [394](#)
- for security applications, [447](#)
- thick, [390](#)–[407](#)
- thick reflection, [358](#)
- transmission, [382](#)

• Holographic art, [447](#)

• Holographic data storage, [442](#)–[443](#)

• Holographic displays, [447](#)

• Holographic images, degradations of, [421](#)–[426](#)

• Holographic optical elements, [447](#)

• Holographic stereograms, [384](#)–[385](#)

• Holographic weights

- for artificial neural networks, [443](#)–[447](#)
- optical neural networks based on volume, [445](#)–[447](#)
 - holograms for security applications, [447](#)
 - holographic display and holographic art, [447](#)
 - holographic optical elements, [447](#)
 - holographic data storage, [442](#)–[443](#)
 - holographic weights for artificial neural networks, [443](#)–[447](#)
 - imaging through distorting media, [438](#)–[442](#)
 - interferometry, [358](#)–[359](#), [432](#)–[438](#)
 - microscopy and high-resolution volume imagery, [431](#)–[432](#)

- cameras in, [257](#)

- computer-generated holograms in, [412](#)–[421](#)

- degradations of holographic images in, [421](#)–[426](#)

- different types of holograms in, [380](#)–[390](#)

- digital, [3](#), [426](#)–[428](#)

- Gabor hologram in, [363](#)–[366](#)

- historical introduction to, [357](#)–[358](#)

- image formation by, [205](#)

- image locations and magnification in, [375](#)–[380](#)

- Leith-Upatnieks homogram in, [366](#)–[375](#)

- linearity of, [361](#)

- offset reference-wave digital, [427](#)

- phase-shifting digital, [427–428](#)
- recording materials in, [408–412](#)
- with spatially incoherent light, [428–431](#)
- spectral, [3, 469–474](#)
- synthetic aperture Fourier, [251–252](#)
- thick holograms in, [390–407](#)
- of three-dimensional scenes, [371–374](#)
- volume, [443–447](#)
- wavefront reconstruction problem in, [358–363](#)

¶Homogeneous medium, [48](#)

¶Hopkins, H. H., [185, 189n](#)

¶Hughes Research Laboratories, [409](#)

¶Hurter, F., [272](#)

¶Hurter-Drifford curve, [273–277](#)

- shoulder of, [273](#)
- toe of, [273](#)

¶Huygens, Christian, [45, 46, 58](#)

¶Huygens-Fresnel principle, [46, 64–66, 67, 75](#)

- in rectangular coordinates, [77–78](#)

¶Hybrid-field-effect mode, [299](#)

¶Hybrid input-output algorithm, [39](#)

¶Hyperfocal distance, [233n](#)

¶ideal image, [189](#)

¶-D Fourier transform, [37](#)

¶Ilford, [408](#)

¶Illumination

- monochromatic, [168–174](#)
- polychromatic, [189–194](#)

¶Image

- effects of diffraction on the, [187–189](#)
- relation between object and, [172–174](#)

¶Image formation, [168–174](#)

- by holography, [205](#)

¶Image intensity, frequency spectrum of the, [215–216](#)

¶Image restoration

- in analog optical information processing, [335–341](#)
- inverse filter in, [335–336, 338–339](#)
- Wiener filter or least-mean-square-error filter in, [336–337](#)

¶Imaginary error function, [235](#)

¶Imaging systems, [1](#)

- depth of focus of, [231–233](#)
- diffraction-limited, [187](#)
- generalized treatment of, [186–194](#)

- ◀ Imaging through distorting media, [438–442](#)
- ◀ Impulse response, [26](#)
 - broadening, [170](#)
 - of a positive lenses, [168–169](#)
- ◀ Incoherent structured illumination imaging, [257, 258–260](#)
- ◀ Interference gain, [221](#)
- ◀ Information processing, [309](#)
- ◀ Instantaneous intensity, [77](#)
- ◀ Integral theory, [51–54](#)
 - application of, [55–56](#)
- ◀ Integrograph, [408](#)
- ◀ Integrated optics waveguides, [475–476](#)
- ◀ Integrated star couplers, [476–478](#)
- ◀ Intensity, [164](#)
 - instantaneous, [77](#)
 - of a wave field, [75–77](#)
- ◀ Intensity impulse response, [193](#)
- ◀ Intensity transmittance, [272](#)
- ◀ Interference, [45](#)
- ◀ Interferograms, phase contour, [420–421](#)
- ◀ Interferometry, [358–359, 432–438](#)
- ◀ Invariance, [1](#)
- ◀ Invariant linear systems, [26–28](#)
- ◀ Inverse filter in image restoration, [335–336, 338–339](#)
- ◀ Inverse Fourier transform, [6](#)
- ◀ Isoplanatic patches, [27](#)
- ◀ Isotropic medium, [48](#)

- ◀ Jinc function, [16](#)
- ◀ Joint transform correlator, [326–329](#)
- ◀ Jones, R. C., [509](#)
- ◀ Jones calculus, [293](#)
- ◀ Jones matrix, [293, 294, 295](#)
 - defined, [509–510](#)

- ◀ Kador, L., [260](#)
- ◀ Keller, J. B., [68](#)
- ◀ Kelley, D. H., [277](#)
- ◀ Kelly model, [277](#)
- ◀ Kinoform, [418](#)
- ◀ Kirchhoff, Gustav, [46, 47, 51](#)
 - integral theorem of, [51–54](#)
- ◀ Kirchoff boundary conditions, [56](#)

- ¶ Kirchoff formulation of diffraction by a planar screen, [2](#), [54–58](#)
 - comparison with Rayleigh-Sommerfeld theory, [63–64](#)
- ¶ Knox, C., [432](#)
- ¶ Kodak, [408](#)
- ¶ Kogelnik, H., [398](#)
- ¶ Koliopoulos, C. L., [345](#), [347](#)
- ¶ Kottler, F., [47](#)
- ¶ Kozma, A., [275](#), [424](#), [425](#)

- ¶ Laguerre-Gaussian beams, [112–113](#), [237](#)
- ¶ Latent image, [271](#)
- ¶ Least-mean square-error filter, [336–337](#)
- ¶ Lee, W. H., [420](#)
- ¶ Leith, E. N., [3](#), [321n](#), [358](#), [375](#)
- ¶ Leith-Upatnieks holograms, [366–375](#)
- ¶ Lens, [2](#), [155](#)
 - Fourier transforming properties of, [2](#), [161–167](#)
 - impulse response of positive, [168–169](#)
 - input placed against the, [162–164](#)
 - input placed behind, [166–167](#)
 - input placed in front of, [164–166](#)
 - negative or diverging, [160](#), [161](#)
 - positive or converging, [160](#), [161](#)
 - thin, as a phase transformation, [155–161](#)
 - types of, [159](#)
- ¶ Lens law, [169–172](#)
- ¶ Lensless Fourier transform holograms, [422–423](#)
- ¶ l'Hôpital's rule, [285](#)
- ¶ Light field photography, [263–266](#)
- ¶ Lighthill, M. J., [8](#)
- ¶ Light propagation, corpuscular theory of, [45](#)
- ¶ Light scattering, [276](#)
- ¶ Lin, L. H., [357](#)
- ¶ Linear exponential, [23](#)
 - multiplication by a, [22](#)
- ¶ Linear filters, effects of, [277](#)
- ¶ Linear imaging system, [26](#)
- ¶ Linearity, [1](#), [5](#), [65](#)
 - of the holographic process, [361](#)
 - superposition integral and, [25–26](#)
- ¶ Linearity theorem, [9](#), [489](#)
- ¶ Linear medium, [48](#)
- ¶ Linear phenomena, [5](#)
- ¶ Linear spatial filter, propagation phenomenon as a, [72–73](#)

- ¶Linear systems
 - invariant, [26–28](#)
 - theory of, [5, 25–28](#)
- ¶Line-spread function of a two-dimensional imaging system, [225–226](#)
- ¶Lippmann, G., [358](#)
- ¶Liquid crystals
 - devices, [269](#)
 - electrical properties of, [291–293](#)
 - mechanical properties of, [288–290](#)
 - microdisplays, [297–298](#)
 - optical properties of, [293–296](#)
 - on silicon, [301](#)
 - spatial light modulators based on, [297–301](#)
- ¶Liquid gate, [274](#)
- ¶Lithographically defined masks, [269](#)
- ¶Lithography, single-step, [281–283](#)
- ¶Local spatial frequencies, [18–21](#)
 - rays and, [496–497](#)
- ¶Lohmann, A. W., [358, 413, 428](#)
- ¶Long-period grating, [463](#)
- ¶Lukosz, W., [253](#)
- ¶Lyot coronography, [241–242, 243, 266](#)

- ¶Mach-Zehnder interferometer, [322](#)
- ¶Maggi, G. A., [67](#)
- ¶Magnification property, [22–23](#)
- ¶Marchand, E. W., [64](#)
- ¶Maréchal, A., [275, 314–316](#)
- ¶Mathematica, 100, [123, 151n, 234](#)
- ¶MatLab, [100](#)
- ¶Matrix theory of paraxial geometrical optics, [155](#)
- ¶Maxwell, J. C., [46, 47, 48, 49](#)
- ¶Media, imaging through distorting, [438–442](#)
- ¶Meridional rays, [498](#)
- ¶Mersereau, R. M., [32](#)
- ¶Mertz, L., [428](#)
- ¶Metamaterials, [47](#)
- ¶Micromaching, [286–287](#)
- ¶Micro-mechanical devices, [269](#)
- ¶Microscopy
 - confocal, [221–225](#)
 - high-resolution volume imagery and, [431–432](#)
 - photoactivated localization, [262](#)
 - stimulated emission depletion, [261–262](#)

- stochastic optical reconstruction, [262](#)
- super-resolved fluorescence, [260–262](#)
- Middleton, D., [32](#)
- Minimum reference angle, [369–371](#)
- Minsky, Marvin, [221](#)
- Miyamoto, A., [68](#)
- Modulated drive voltage, [306–308](#)
- Modulation transfer function, [198, 277](#)
- Moerner, W. E., [260](#)
- Monochromatic illumination, [168–174](#)
- Montgomery, R. M., [345, 347](#)
- Moore’s law, [309–310](#)
- Multiple-exposure holographic interferometry, [432–433](#)
- Multiplex holograms, [387–389](#)
- Multiplication by a linear exponential, [22](#)
- Multiplication property, [22](#)
- Multi-step lithography, [283–287](#)
- Mutual intensity, [192](#)

- Narrowband filters for add/drop multiplexers, [461–462](#)
- Negative lens, [160, 161](#)
- Negative phase contrast, [314](#)
- Negative sidelobes, [24](#)
- Nematic liquid crystals, [289](#)
 - optical properties of, [293–296](#)
- Neural networks, holographic weights for artificial, [443–447](#)
- Newton, Isaac, [45](#)
- Ng, Ren, [263](#)
- Nondispersive medium, [48](#)
- Nonmagnetic medium, [48](#)
- Nonmonochromatic light, [161](#)
- Nonmonochromatic waves, generalization to, [66–67](#)
- Nontanning bleach, [279, 280](#)
- Nuttall apodization, [266–267](#)
- Nyquist density, [32, 33](#)
- Nyquist rate, [31, 32–33](#)

- Object, relation between image and, [172–174](#)
- Obliquity factor, [63, 65](#)
- Offset-reference hologram, [366](#)
- Offset reference-wave digital holography, [427](#)
- One-dimensional Fourier transforms, [12](#)
- Optical communications, Fourier optics in, [453–484](#)

- arrayed waveguide gratings, [474–484](#)
 - fiber Bragg gratings, [454–463](#)
 - spectral holography, [469–474](#)
 - ultrashort pulse shaping, [463–469](#)
- Optical damage in glass, [308](#)
- Optical elements, holographic, [447](#)
- Optical fibers, [454–457](#)
 - recording gratings in, [457–458](#)
- Optical Fourier transform, [167](#)
- Optical imaging systems
 - analysis of complex coherent, [174–177](#)
 - analysis of two, using ray matrices, [175–177](#)
 - comparison of coherent and incoherent imaging, confocal microscopy, [221–225](#)
 - frequency analysis of
 - aberrations and their effects on frequency response, [204–214](#)
 - comparison of coherent and incoherent imaging, [214–221](#)
 - frequency response for diffraction-limited coherent imaging, [194–197](#)
 - frequency response for diffraction-limited incoherent imaging, [197–204](#)
 - generalized treatment of imaging systems, [186–194](#)
 - photographic film or plate in coherent light, [274–276](#)
 - wave-optics analysis of coherent systems, [155–183](#)
 - analysis of complex coherent optical systems, [174–177](#)
 - Fourier transforming properties of lenses, [161–167](#)
 - image formation: monochromatic illumination, [168–174](#)
 - thin lens as phase transformation, [155–161](#)
- Optically driven liquid crystal light modulators, [298–300](#)
- Optical neural networks based on volume holographic weights, [445–447](#)
- Optical properties of liquid crystals, [293–296](#)
- Optical synthesis of character-recognition machine, [332–334](#)
- Optical transfer function (OTF), [197–199](#)
 - of an aberration-free system, [200–201](#)
 - effects of aberrations on the, [206–207](#)
 - examples of diffraction-limited, [201–204](#)
 - general properties of, [199–200](#)
- Optics, 1
 - beam, [108–114](#)
 - binary, [283](#)
 - coherent, [316](#)
 - diffractive (See [Diffractive optics](#))
 - Fourier (See [Fourier optics](#))
 - geometrical, [495–496](#)
 - history of, 2
- Orbital angular momentum, [113](#)
- Orthoscopic image, [372](#)
- Outgoing waves, [56](#)

•Oversampling, [33](#)

•Palermo, C., [321n](#)

•Papoulis, A., [6](#)

•Paraxial approximation, [158](#), [497](#)

•Paraxial Helmholtz equation, [74](#), [108](#)

•Paraxial rays, [158](#)

•Paris, D. P., [413](#)

•Parseval's theorem, [10](#)

•Partial coherence, [193](#)

- theory of, [189](#)

•Penumbra effect, [44](#)

•Perfect image, [105](#)

•Peterson, D. P., [32](#)

•Phase contour interferograms, [420–421](#)

•Phase-contrast microscopes, [314n](#)

•Phase contrast technique, [313](#)

•Phase distributions, [218](#)

- manipulation of, [269](#)

•Phase modulation, [279](#)

•Phase-only filter, [321n](#)

•Phase retrieval from Fourier magnitude, [38–39](#)

•Phase-shifting digital holography, [427–428](#)

•Phase transformation

- physical meaning of, [158–161](#)
- thin lens as a, [155–161](#)

•Phasor, [50](#)

•Photoactivated localization microscopy, [262](#)

•Photographic density, [272–273](#)

•Photographic emulsions, bleaching of, [279–280](#)

•Photographic film

- exposure, development, and fixing of, [270–272](#)
- properties of, [2](#)
- wavefront modulation with, [270–280](#)

•Photographic materials, importance of, [269](#)

•Photographs, improvement of, [314–316](#)

•Photopolymer films, [308](#), [408–409](#)

•Photorefractive crystals, [446](#)

•Photorefractive materials, [308](#), [409–412](#)

•Pico-projector applications, [301](#)

•Planar interface, refraction at a, [500](#)

•Planar screen, [64](#)

- Kirchoff formulation of diffraction by a, [54–58](#)

•Planes

- conjugate, [501–502](#)
- focal, [502–503](#)
- principal, [503–505](#)

Plane waves, angular spectrum of, [68–73](#)

Point-spread functions, [2](#), [26](#)

- transfer function engineering and, [231–267](#)
 - cubic phase mask for increased depth of field, [231–237](#)
 - light field photography, [263–266](#)
 - point-spread function engineering for exoplanet discovery, [241–248](#)
 - resolution beyond the classical diffraction limit, [248–263](#)
 - rotating point-spread functions for depth resolution, [237–240](#)

Poisson, S., [45–46](#)

Polarization transformations, examples of simple, [511–512](#)

Polarization vector, [293](#)

Polychromatic illumination, [189–194](#)

Porter, A. B., [310](#)

Positive lens, [160](#), [161](#)

Positive phase contrast, [314](#)

Power spectral density, [336n](#)

Power spectrum, [164](#), [343](#)

Principal planes, [503–505](#)

Prism, [177–178](#)

Projection-slice theorem, [37–38](#)

Propagation

- phenomenon of, as a linear spatial filter, [72–73](#)
- theory of, [475](#)
- through free space of index n , [499–500](#)

Pseudocopic image, [372](#)

Ptychography, Fourier, [2](#), [252–253](#)

Pulse shaping system, [466–467](#)

Pupil function, [163](#)

Pupils, entrance and exit, [506–507](#)

Q factor, [306](#)

Q-parameter, [108](#)

Quadratic-phase dispersion, [84](#)

Quadratic-phase exponential, [23](#)

- finite integral of the, [81–83](#)

Quadratic-phase factors, eliminating, [169–172](#)

Quadratic-phase functions, [19n](#)

Quadratic-phase transformation, [2](#)

Quantum efficiency, [76](#)

- Radio-wave propagation, [43](#)
- Ragnarsson, S. E., [339](#), [341](#)
- Ragnarsson filter, [340](#)
- Rainbow holograms, [385–387](#)
- Raman-Nath diffraction, [341](#)
- Raman-Nath regime, [304–305](#), [306](#)
- Ratcliffe, J. A., [68](#)
- Ray(s)
 - defined, [496](#)
 - local spatial frequency and, [496–497](#)
 - meridional, [498](#)
- Rayleigh, Lord, [185](#), [188](#)
- Rayleigh criterion of resolution, [216–219](#), [232](#)
- Rayleigh interferometer, [323](#)
- Rayleigh range, [108](#)
- Rayleigh-Sommerfeld formulation of diffraction, [2](#), [46](#), [58–63](#)
 - comparison with Kirchhoff theory, [63–64](#)
- Rayleigh's (Parseval's) theorem, [10](#), [490](#)
- Ray matrix, [174–177](#)
 - analysis of two optical systems using, [175–177](#)
- Ray transfer matrix, [498–501](#)
 - Fresnel diffraction in terms of, [85–88](#)
- RCA Laboratories, [410](#)
- Real-time holographic interferometry, [433](#), [435](#)
- Reciprocity theorem of Helmholtz, [58](#)
- Reconstruction wave, [360–361](#)
- Recording gratings in optical fibers, [457–458](#)
- Rectangle function, [14](#), [23–24](#)
- Rectangular aperture, [90–91](#), [148–149](#)
- Rectangular coordinators, Huygens-Fresnel principle in, [77–78](#)
- Reduced coordinates, [189](#)
- Referenceless on-axis complex hologram (ROACH), [419](#)
- Reference pulse, effects of delay between signal waveform and, [474](#)
- Reflection hologram, [394](#)
- Reflection modulator, [296](#)
- Reflective polarization devices, [513–514](#)
- Refraction, [497](#)
 - defined, [43](#)
 - at a planar interface, [500](#)
 - at a spherical interface, [500–501](#)
- Refractive index, [48](#), [280](#), [283](#)
 - extraordinary, [293](#)
 - ordinary, [293](#)
- Refractive optical elements, [280](#)
- Relief image, [279](#)

❖ Resolution

- beyond the classical diffraction limit, [248–263](#)
- Rayleigh criterion of, [216–219](#)

❖ Resolution cell, [220](#)

❖ Responsivity, [76](#)

❖ Restrick, R. C., III, [428](#)

❖ Reynolds, G. O., [357](#)

❖ Rogers, G. L., [358, 428](#)

❖ Rotating point-spread functions for depth resolution, [237–240](#)

❖ Rotation theorem, [10–11, 491–492](#)

❖ Rubinowicz, A., [67](#)

❖ Sampling a space-limited quadratic-phase exponential, [122–125](#)

❖ Sampling theory, two-dimensional, [28–35](#)

❖ Saxby, G., [357](#)

❖ Scalar diffraction theory, [2, 43–74](#)

- angular spectrum of plane waves and, [68–73](#)
- comparison of Kirchoff and Rayleigh-Sommerfeld theories and, [63–64](#)
- diffraction at boundaries and, [67–68](#)
- generalization to nonmonochromatic waves and, [66–67](#)
- Green's theorem and, [51](#)
- Helmholtz equation and, [50–51](#)
- historical introduction to, [43–47](#)
- Huygens-Fresnel principle and, [64–66](#)
- integral theorem of Helmholtz and Kirchhoff and, [51–54](#)
- Kirchhoff formulation of diffraction by a planar screen and, [54–58](#)
- Rayleigh-Sommerfeld formulation of diffraction and, [58–63](#)
- from a vector to a scalar theory and, [47–50](#)

❖ Scalar field, [50](#)

❖ Scalar phenomenon, [47](#)

❖ Scalar theory, [50](#)

❖ Schlieren method, [313](#)

- for observing phase objects, [352](#)

❖ Schumann, W., [432](#)

❖ Second Rayleigh-Sommerfeld solution, [61](#)

❖ Security applications, holograms for, [447](#)

❖ Self-images, [105](#)

❖ Separable functions, [11–12](#)

❖ Shannon, C. E., [28, 29](#)

❖ Shear theorem, [11, 492](#)

❖ Sheppard, C., [221](#)

❖ Sherman, G. C., [73](#)

❖ Shift property, [22, 25–26](#)

❖ Shift theorem, [10, 490](#)

- Signal waveform, effects of delay between reference pulse and, [474](#)
- Signum function, [14](#)
- Silver halide emulsions, [408](#)
- Similarity theorem, [10](#), [489](#)
- Sinc function, [14](#)
- Single DFT approach, [122](#)
- Single step lithography, [281–283](#)
- Sinusoidal amplitude grating, [93–96](#)
 - Fresnel diffraction by a, [104–108](#)
 - thin, [94](#), [96–98](#)
- Slanted grating, [463](#)
- Smectic liquid crystals, [289](#)
- Smith, H. M., [357](#)
- Snell's Law, [44](#), [455](#), [497](#)
- Sommerfeld, A., [44](#), [46](#), [47](#), [59](#), [67](#)
- Sommerfeld radiation condition, [56](#)
- Space-bandwidth product, [33–35](#)
- Space domain, [326](#)
- Space-integrating correlator, [343–345](#)
- Space-invariance, [27](#), [66](#), [72](#)
- Space-limited quadratic-phase exponential, sampling a, [122–125](#)
- Space-variant system, [173](#)
- Sparrow resolution criterion, [229–230](#)
- Spatial amplitude, manipulation of, [269](#)
- Spatial coherence, [5](#), [6](#), [192](#)
- Spatial filtering, [370n](#)
 - power of, [311](#)
- Spatial frequencies
 - mapping of temporal frequencies to, [464–466](#)
 - space frequency localization and, [17–24](#)
- Spatial light modulation, [2](#), [162](#), [287–288](#), [289](#)
 - acoustic-optic, [304–308](#)
 - based on liquid crystals, [297–301](#)
 - deformable mirror, [301–303](#)
 - liquid crystal, [287–301](#)
- Spatially incoherent light, holography with, [428–431](#)
- Spatial period, [9](#)
- Speckle effect, [220](#)
- Speckle noise, [426](#)
- Spectral holography, [3](#), [469–474](#)
- Spectral pulse shaping, applications of, [467–469](#)
- Spectrum analyzer, [343](#)
- Spherical interface, refraction at a, [500–501](#)
- Sprague, R. A., [345](#), [347](#)
- Square aperture, Fresnel diffraction by a, [99–102](#)

- Square-law mapping, [275](#)
- Square wave phase gratings, [281](#)
- Starlight suppression, apodization for, [242](#), [244–248](#)
- Stationary phrase, [83](#)
- Stepped thickness function, approximation by a, [283–285](#)
- Stereograms, holographic, [384–385](#)
- Stimulated emission depletion microscopy, [261–262](#)
- Stochastic optical reconstruction microscopy, [262](#)
- Strehl definition, [228](#)
- Stroke, G. W., [428](#)
- Structured illumination, [2](#)
- Superposition integral, [26](#), [65](#)
 - linearity and, [25–26](#)
- Super-resolved fluorescence microscopy, [260–262](#)
- Symmetry, circular, [12–14](#)
- Synthesis, [1](#)
- Synthetic aperture Fourier holography, [2](#), [251–252](#)
- System
 - defined, [25](#)
 - linear, [25–28](#)
- Takahashi, H., [474](#)
- Talbot images, [104–108](#)
- Talbot subimage, [106](#)
- Tanning bleach, [279](#)
- Temporal changes, [65](#)
- Temporal frequencies, mapping of to spatial frequencies, [464–466](#)
- Temporal function, spectrum of, [1](#)
- Temporal imaging, [474](#)
- Terminal properties, [186](#)
- Texas Instruments, Inc., [301](#)
- Thick amplitude reflection grating, [405–407](#)
- Thick amplitude transmission grating, [402–404](#)
- Thick holograms, [390–407](#)
- Thickness function, [156–158](#)
- Thick phase reflection grating, [404–405](#)
- Thick phase transmission grating, [401–402](#)
- Thick reflection hologram, [358](#)
- Thin lens, [155](#)
 - passage through, [501](#)
 - as a phase transformation, [155–161](#)
- Thin sinusoidal amplitude grating, [94](#), [96–98](#)
- Thompson, B. J., [432](#)
- ThorLabs, [408](#)

- Three-dimensional scenes, holography of, [371–374](#)
- Tichenor, D. A., [172](#), [341](#)
- Time derivative, [67](#)
- Time-integrating correlator in acousto-optic signal processing systems, [345–347](#)
- Time-invariance, [26](#)
- Time-to-space mapping, [465](#)
- Time-varying phasor, [190](#)
- Toal, V., [357](#)
- Transfer function engineering and point-spread function, [231–267](#)
 - cubic phase mask for increased depth of field, [231–237](#)
 - light field photography, [263–266](#)
 - point-spread function engineering for exoplanet discovery, [241–248](#)
 - resolution beyond the classical diffraction limit, [248–263](#)
 - rotating point-spread functions for depth resolution, [237–240](#)
- Transfer functions, [2](#), [26–28](#)
- Transverse magnifications, [378–379](#)
- Triangle function, [15](#)
- Twisted nematic liquid crystal, [289](#)
- Two-dimensional Fourier transform, [8](#)
- Two-dimensional imaging system, line-spread function of, [225–226](#)
- Two-dimensional sampling theory, [28–35](#)
- Two-dimensional signals and systems, [5–42](#)
 - discrete Fourier transform in, [35–37](#)
 - Fourier analysis in, [6–17](#)
 - linear systems in, [25–28](#)
 - phase retrieval from Fourier magnitude in, [38–39](#)
 - projection-slice theorem in, [37–38](#)
 - sampling theory in, [28–35](#)
 - spatial frequency and space-frequency localization in, [17–24](#)
- Two-point resolution, [216–219](#)

- Ultimate holography, [408](#)
- Ultrashort pulse shaping processing, [3](#), [463–469](#)
- Undersampling, [33](#)
- Unit-amplitude, [180](#)
- Upatnieks, J., [3](#), [358](#), [375](#)
- Upsampling, [253n](#)
- Vacuum permittivity, [48](#)
- VanderLugt, A. B., [321](#)
- VanderLugt filter, [321–326](#), [337–338](#), [340](#), [354–355](#)
 - advantages of, [326](#)
 - processing input data, [324–325](#)
 - synthesis of the frequency-plane mask, [321–324](#)

- VanderLugt geometry, [329](#)
- Vector theory, [49](#)
- Vest, C. M., [432](#)
- Vibration analysis, [437–438](#)
- Vignetting effect, [166](#)
- Viritual image, [362](#)
- VLSI fabrication, [283](#)
- Volume gratings
 - of finite size, [396–398](#)
 - reconstructing wavefronts from a, [392–394](#)
- Volume holographic grating, recording, [391–392](#)
- Volume holography, [443–447](#)
- Vortex phase filter, [242](#)

- Ward, J. H., [432](#)
- Wave field, intensity of, [75–77](#)
- Wavefront modulation, [2, 269–308](#)
 - deformable mirror spatial light modulators, [301–303](#)
 - with diffractive optical elements, [280–287](#)
 - liquid crystal spatial light modulators, [287–301](#)
 - other methods of, [308](#)
 - with photographic film, [270–280](#)
- Wavefront reconstruction, [357–358](#)
 - image formation by holography in, [361–363](#)
 - linearity of the holographic process in, [361](#)
 - reconstruction of the original wavefront in, [360–361](#)
 - recording amplitude and phase in, [358–359](#)
 - recording medium in, [359–360](#)
- Waveguide grating, [3, 478–479](#)
- Wavelength multiplexers, [481](#)
- Wavelength routers, [481–484](#)
- Wave number, [51](#)
- Wave-optics analysis of coherent optical systems, [2, 155–183](#)
 - analysis of complex coherent optical systems, [174–177](#)
 - Fourier transforming properties of lenses, [161–167](#)
 - image formation: monochromatic illumination, [168–174](#)
 - thin lens as phase transformation, [155–161](#)
- Wave propagation, [43](#)
- Wave vector diagram, [306](#)
- Weaver, C. J., [326](#)
- White noise, [329](#)
- Whittaker-Shannon sampling theorem, [29–32](#)
- Wiener filter, [336–337, 339–341](#)
- Wigner distribution function, [21–24](#)

•Wilson, T., [221](#)
•Windowing, [212](#)
•Wolf, E., [38](#), [64](#), [68](#)

•X-ray crystallography, [357](#)

•Yamaguchi, I., [427](#)
•Young, N. O., [428](#)
•Young, Thomas, [45](#), [58](#), [67](#)

•Zelenka, J. S., [424](#)
•Zernike, Frits, [313](#), [314](#), [316](#)
•Zernike phase-contrast microscope, [312–314](#)
•Zero-frequency values, [197](#)
•Zero order, [95](#)
•Zero-spread nonlinearity, [277](#)
•Zinky, W. R., [432](#)



Contents

1. [Cover Page](#)
2. [Title Page](#)
3. [Dedication Page](#)
4. [Copyright Page](#)
5. [The Author](#)
6. [Preface](#)
7. [Contents](#)
8. [1 Introduction](#)
 1. [1.1 Optics, Information, and Communication](#)
 2. [1.2 The Book](#)
9. [2 Analysis of Two-Dimensional Signals and Systems](#)
 1. [2.1 Fourier Analysis in Two Dimensions](#)
 1. [2.1.1 Definition and Existence Conditions](#)
 2. [2.1.2 The Fourier Transform as a Decomposition](#)
 3. [2.1.3 Fourier Transform Theorems](#)
 4. [2.1.4 Separable Functions](#)
 5. [2.1.5 Functions with Circular Symmetry: Fourier-Bessel Transforms](#)
 6. [2.1.6 Some Frequently Used Functions and Some Useful Fourier Transform Pairs](#)
 2. [2.2 Spatial Frequency and Space-Frequency Localization](#)
 1. [2.2.1 Local Spatial Frequencies](#)
 2. [2.2.2 The Wigner Distribution Function](#)
 1. [Real-valued Property](#)
 2. [Shift Property](#)
 3. [Multiplication by a Linear Exponential](#)
 4. [Convolution Property](#)
 5. [Multiplication Property](#)
 6. [Magnification Property](#)
 7. [Fourier Transform Property](#)
 3. [2.3 Linear Systems](#)
 1. [2.3.1 Linearity and the Superposition Integral](#)
 2. [2.3.2 Invariant Linear Systems: Transfer Functions](#)
 4. [2.4 Two-Dimensional Sampling Theory](#)
 1. [2.4.1 The Whittaker-Shannon Sampling Theorem](#)
 2. [2.4.2 Oversampling, Undersampling and Aliasing](#)
 3. [2.4.3 Space-Bandwidth Product](#)
 5. [2.5 The Discrete Fourier Transform](#)

- 6. [2.6 The Projection-Slice Theorem](#)
- 7. [2.7 Phase Retrieval from Fourier Magnitude](#)
- 8. [Problems - Chapter 2](#)
- 10. [**3 Foundations of Scalar Diffraction Theory**](#)
 - 1. [3.1 Historical Introduction](#)
 - 2. [3.2 From a Vector to a Scalar Theory](#)
 - 3. [**3.3 Some Mathematical Preliminaries**](#)
 - 1. [3.3.1 The Helmholtz Equation](#)
 - 2. [3.3.2 Green's Theorem](#)
 - 3. [3.3.3 The Integral Theorem of Helmholtz and Kirchhoff](#)
 - 4. [**3.4 The Kirchhoff Formulation of Diffraction by a Planar Screen**](#)
 - 1. [3.4.1 Application of the Integral Theorem](#)
 - 2. [3.4.2 The Kirchhoff Boundary Conditions](#)
 - 3. [3.4.3 The Fresnel-Kirchhoff Diffraction Formula](#)
 - 5. [**3.5 The Rayleigh-Sommerfeld Formulation of Diffraction**](#)
 - 1. [3.5.1 Choice of Alternative Green's Functions](#)
 - 2. [3.5.2 The Rayleigh-Sommerfeld Diffraction Formula](#)
 - 3. [3.5.3 Reproduction of Boundary Conditions](#)
 - 6. [3.6 Kirchhoff and Rayleigh-Sommerfeld Theories Compared](#)
 - 7. [3.7 Further Discussion of the Huygens-Fresnel Principle](#)
 - 8. [3.8 Generalization to Nonmonochromatic Waves](#)
 - 9. [3.9 Diffraction at Boundaries](#)
 - 10. [**3.10 The Angular Spectrum of Plane Waves**](#)
 - 1. [3.10.1 The Angular Spectrum and Its Physical Interpretation](#)
 - 2. [3.10.2 Propagation of the Angular Spectrum](#)
 - 3. [3.10.3 Effects of a Diffracting Aperture on the Angular Spectrum](#)
 - 4. [3.10.4 The Propagation Phenomenon as a Linear Spatial Filter](#)
- 11. [Problems - Chapter 3](#)
- 11. [**4 Fresnel and Fraunhofer Diffraction**](#)
 - 1. [**4.1 Background**](#)
 - 1. [4.1.1 The Intensity of a Wave Field](#)
 - 2. [4.1.2 The Huygens-Fresnel Principle in Rectangular Coordinates](#)
 - 2. [**4.2 The Fresnel Approximation**](#)
 - 1. [4.2.1 Positive vs. Negative Phases](#)
 - 2. [4.2.2 Accuracy of the Fresnel Approximation](#)
 - 3. [4.2.3 Finite Integral of the Quadratic-Phase Exponential Function](#)
 - 4. [4.2.4 The Fresnel Approximation and the Angular Spectrum](#)
 - 5. [4.2.5 Fresnel Diffraction Between Confocal Spherical Surfaces](#)
 - 6. [4.2.6 Fresnel Diffraction in Terms of Ray Transfer Matrices](#)

- 3. [4.3 The Fraunhofer Approximation](#)
 - 4. [4.4 Examples of Fraunhofer Diffraction Patterns](#)
 - 1. [4.4.1 Rectangular Aperture](#)
 - 2. [4.4.2 Circular Aperture](#)
 - 3. [4.4.3 Thin Sinusoidal Amplitude Grating](#)
 - 4. [4.4.4 Thin Sinusoidal Phase Grating](#)
 - 5. [4.4.5 General Method for Calculating Diffraction Efficiency of Gratings](#)
 - 5. [4.5 Examples of Fresnel Diffraction Calculations](#)
 - 1. [4.5.1 Fresnel Diffraction by a Square Aperture](#)
 - 2. [4.5.2 Fresnel Diffraction by a Circular Aperture](#)
 - 3. [4.5.3 Fresnel Diffraction by a Sinusoidal Amplitude Grating-Talbot Images](#)
 - 6. [4.6 Beam Optics](#)
 - 1. [4.6.1 Gaussian Beams](#)
 - 2. [4.6.2 Hermite-Gaussian Beams](#)
 - 3. [4.6.3 Laguerre-Gaussian Beams](#)
 - 4. [4.6.4 Bessel Beams](#)
 - 7. [Problems - Chapter 4](#)
12. [5 Computational Diffraction and Propagation](#)
 - 1. [5.1 Approaches to Computational Diffraction](#)
 - 2. [5.2 Sampling a Space-Limited Quadratic-Phase Exponential](#)
 - 3. [5.3 The Convolution Approach](#)
 - 1. [5.3.1 Bandwidth and Sampling Considerations](#)
 - 2. [5.3.2 Discrete Convolution Equations](#)
 - 3. [5.3.3 Simulation Results](#)
 - 4. [5.3.4 Convolution by Fourier Transforms](#)
 - 4. [5.4 The Fresnel Transform Approach](#)
 - 1. [5.4.1 Sampling Increments](#)
 - 2. [5.4.2 Sampling Ratio Q](#)
 - 3. [5.4.3 Finding the Required M, Q, and N](#)
 - 4. [5.4.4 The Discrete Diffraction Formulas](#)
 - 5. [5.4.5 Examples of the Dependence of M and N on NF](#)
 - 6. [5.4.6 Summary of Steps Using the Fresnel Transform Approach](#)
 - 7. [5.4.7 Computational Complexity of the Fresnel Transform Approach](#)
 - 5. [5.5 The Fresnel Transfer Function Approach](#)
 - 1. [5.5.1 Sampling Considerations](#)
 - 2. [5.5.2 Finding N, M and Q for each NF](#)
 - 3. [5.5.3 The Discrete Diffraction Formulas](#)
 - 4. [5.5.4 Examples of the Dependence of M, N and Q on NF](#)
 - 5. [5.5.5 Summary of Steps Using the Fresnel Transfer Function Approach](#)

- 6. [5.5.6 Computational Complexity of the Fresnel Transfer Function Approach](#)
 - 6. [5.6 The Exact Transfer Function Approach](#)
 - 1. [5.6.1 Sampling in the Frequency Domain](#)
 - 2. [5.6.2 Sampling in the Space Domain](#)
 - 3. [5.6.3 Simulation Results](#)
 - 4. [5.6.4 Computational Complexity of the Exact Transfer Function Approach](#)
 - 7. [5.7 Comparison of Computational Complexities](#)
 - 8. [5.8 Extension to More Complex Apertures](#)
 - 1. [5.8.1 One-Dimensional Case](#)
 - 2. [5.8.2 Two-Dimensional Apertures Separable in \(x,y\) Coordinates](#)
 - 3. [5.8.3 Circularly-Symmetric Apertures](#)
 - 4. [5.8.4 More General Cases](#)
 - 9. [5.9 Concluding Comments](#)
 - 10. [Problems - Chapter 5](#)
13. [6 Wave-Optics Analysis of Coherent Optical Systems](#)
- 1. [6.1 A Thin Lens as a Phase Transformation](#)
 - 1. [6.1.1 The Thickness Function](#)
 - 2. [6.1.2 The Paraxial Approximation](#)
 - 3. [6.1.3 The Phase Transformation and Its Physical Meaning](#)
 - 2. [6.2 Fourier Transforming Properties of Lenses](#)
 - 1. [6.2.1 Input Placed against the Lens](#)
 - 2. [6.2.2 Input Placed in Front of the Lens](#)
 - 3. [6.2.3 Input Placed behind the Lens](#)
 - 4. [6.2.4 Example of an Optical Fourier Transform](#)
 - 3. [6.3 Image Formation: Monochromatic Illumination](#)
 - 1. [6.3.1 The Impulse Response of a Positive Lens](#)
 - 2. [6.3.2 Eliminating Quadratic-Phase Factors: The Lens Law](#)
 - 3. [6.3.3 The Relation between Object and Image](#)
 - 4. [6.4 Analysis of Complex Coherent Optical Systems](#)
 - 1. [6.4.1 The Ray Matrix Approach](#)
 - 2. [6.4.2 Analysis of Two Optical Systems Using Ray Matrices](#)
 - 5. [Problems - Chapter 6](#)
14. [7 Frequency Analysis of Optical Imaging Systems](#)
- 1. [7.1 Generalized Treatment of Imaging Systems](#)
 - 1. [7.1.1 A Generalized Model](#)
 - 2. [7.1.2 Effects of Diffraction on the Image](#)
 - 3. [7.1.3 Polychromatic Illumination: The Coherent and Incoherent Cases](#)
 - 2. [7.2 Frequency Response for Diffraction-Limited Coherent Imaging](#)
 - 1. [7.2.1 The Amplitude Transfer Function](#)

- 2. [7.2.2 Examples of Amplitude Transfer Functions](#)
 - 3. [7.3 Frequency Response for Diffraction-Limited Incoherent Imaging](#)
 - 1. [7.3.1 The Optical Transfer Function](#)
 - 2. [7.3.2 General Properties of the OTF](#)
 - 3. [7.3.3 The OTF of an Aberration-Free System](#)
 - 4. [7.3.4 Examples of Diffraction-Limited OTFs](#)
 - 4. [7.4 Aberrations and Their Effects on Frequency Response](#)
 - 1. [7.4.1 The Generalized Pupil Function](#)
 - 2. [7.4.2 Effects of Aberrations on the Amplitude Transfer Function](#)
 - 3. [7.4.3 Effects of Aberrations on the OTF](#)
 - 4. [7.4.4 Example of a Simple Aberration: A Focusing Error](#)
 - 5. [7.4.5 Apodization and Its Effects on Frequency Response](#)
 - 5. [7.5 Comparison of Coherent and Incoherent Imaging](#)
 - 1. [7.5.1 Frequency Spectrum of the Image Intensity](#)
 - 2. [7.5.2 Two-Point Resolution](#)
 - 3. [7.5.3 Other Effects](#)
 - 6. [7.6 Confocal Microscopy](#)
 - 1. [7.6.1 Coherent Case](#)
 - 2. [7.6.2 Incoherent Case](#)
 - 3. [7.6.3 Optical Sectioning](#)
 - 7. [Problems - Chapter 7](#)
15. [8 Point-Spread Function and Transfer Function Engineering](#)
- 1. [8.1 Cubic Phase Mask for Increased Depth of Field](#)
 - 1. [8.1.1 Depth of Focus](#)
 - 2. [8.1.2 Depth of Field](#)
 - 3. [8.1.3 The Cubic Phase Mask](#)
 - 2. [8.2 Rotating Point-Spread Functions for Depth Resolution](#)
 - 3. [8.3 Point-Spread Function Engineering for Exoplanet Discovery](#)
 - 1. [8.3.1 The Lyot Coronagraph](#)
 - 2. [8.3.2 Apodization for Starlight Suppression](#)
 - 4. [8.4 Resolution beyond the Classical Diffraction Limit](#)
 - 1. [8.4.1 Analytic Continuation](#)
 - 1. [Underlying Mathematical Fundamentals](#)
 - 2. [Intuitive Explanation of Bandwidth Extrapolation](#)
 - 2. [8.4.2 Synthetic Aperture Fourier Holography](#)
 - 3. [8.4.3 Fourier Ptychography](#)
 - 4. [8.4.4 Coherent Spectral Multiplexing](#)
 - 5. [8.4.5 Incoherent Structured Illumination Imaging](#)
 - 6. [8.4.6 Super-Resolved Fluorescence Microscopy](#)

1. [Fluorescent Labelling](#)
 2. [Localization Precision](#)
 3. [Stimulated Emission Depletion Microscopy \(STED\)](#)
 4. [Photoactivated Localization Microscopy \(PALM\) and Stochastic Optical Reconstruction Microscopy \(STORM\)](#)
 5. [8.5 Light Field Photography](#)
 6. [Problems - Chapter 8](#)
16. [9 Wavefront Modulation](#)
1. [9.1 Wavefront Modulation with Photographic Film](#)
 1. [9.1.1 The Physical Processes of Exposure, Development, and Fixing](#)
 2. [9.1.2 Definition of Terms](#)
 3. [9.1.3 Photographic Film or Plate in Coherent Optical Systems](#)
 4. [9.1.4 The Modulation Transfer Function](#)
 5. [9.1.5 Bleaching of Photographic Emulsions](#)
 2. [9.2 Wavefront Modulation with Diffractive Optical Elements](#)
 1. [9.2.1 Single Step Lithography](#)
 2. [9.2.2 Multistep Lithography](#)
 1. [Approximation by a Stepped Thickness Function](#)
 2. [The Fabrication Process](#)
 3. [9.2.3 Other Types of Diffractive Optics](#)
 4. [9.2.4 A Word of Caution](#)
 3. [9.3 Liquid Crystal Spatial Light Modulators](#)
 1. [9.3.1 Properties of Liquid Crystals](#)
 1. [Mechanical Properties of Liquid Crystals](#)
 2. [Electrical Properties of Liquid Crystals](#)
 3. [Optical Properties of Nematic and Ferroelectric Liquid Crystals](#)
 2. [9.3.2 Spatial Light Modulators Based on Liquid Crystals](#)
 1. [Electrically Driven Liquid Crystal Spatial Light Modulators](#)
 2. [Optically Driven Liquid Crystal Spatial Light Modulators](#)
 3. [Ferroelectric Liquid Crystal Spatial Light Modulators](#)
 4. [Liquid Crystal on Silicon \(LCOS\)](#)
 4. [9.4 Deformable Mirror Spatial Light Modulators](#)
 5. [9.5 Acousto-Optic Spatial Light Modulators](#)
 1. [A CW Drive Voltage](#)
 2. [A Modulated Drive Voltage](#)
 6. [9.6 Other Methods of Wavefront Modulation](#)
 7. [Problems - Chapter 9](#)
17. [10 Analog Optical Information Processing](#)
1. [10.1 Historical Background](#)
 1. [10.1.1 The Abbe-Porter Experiments](#)

- 2. [10.1.2 The Zernike Phase-Contrast Microscope](#)
- 3. [10.1.3 Improvement of Photographs: Maréchal](#)
- 4. [10.1.4 Application of Coherent Optics to More General Data Processing](#)
- 2. [10.2 Coherent Optical Information Processing Systems](#)
 - 1. [10.2.1 Coherent System Architectures](#)
 - 2. [10.2.2 Constraints on Filter Realization](#)
- 3. [10.3 The VanderLugt Filter](#)
 - 1. [10.3.1 Synthesis of the Frequency-Plane Mask](#)
 - 2. [10.3.2 Processing the Input Data](#)
 - 3. [10.3.3 Advantages of the VanderLugt Filter](#)
- 4. [10.4 The Joint Transform Correlator](#)
- 5. [10.5 Application to Character Recognition](#)
 - 1. [10.5.1 The Matched Filter](#)
 - 2. [10.5.2 A Character-Recognition Problem](#)
 - 3. [10.5.3 Optical Synthesis of a Character-Recognition Machine](#)
 - 4. [10.5.4 Sensitivity to Scale Size and Rotation](#)
- 6. [10.6 Image Restoration](#)
 - 1. [10.6.1 The Inverse Filter](#)
 - 2. [10.6.2 The Wiener Filter, or the Least-Mean-Square-Error Filter](#)
 - 3. [10.6.3 Filter Realization](#)
 - 1. [Inverse Filter](#)
 - 2. [Wiener Filter](#)
- 7. [10.7 Acousto-Optic Signal Processing Systems](#)
 - 1. [10.7.1 Bragg Cell Spectrum Analyzer](#)
 - 2. [10.7.2 Space-Integrating Correlator](#)
 - 3. [10.7.3 Time-Integrating Correlator](#)
 - 4. [10.7.4 Other Acousto-Optic Signal Processing Architectures](#)
- 8. [10.8 Discrete Analog Optical Processors](#)
 - 1. [10.8.1 Discrete Representation of Signals and Systems](#)
 - 2. [10.8.2 A Parallel Incoherent Matrix-Vector Multiplier](#)
 - 3. [10.8.3 Methods for Handling Bipolar and Complex Data](#)
- 9. [Problems - Chapter 10](#)
- 18. [11 Holography](#)
 - 1. [11.1 Historical Introduction](#)
 - 2. [11.2 The Wavefront Reconstruction Problem](#)
 - 1. [11.2.1 Recording Amplitude and Phase](#)
 - 2. [11.2.2 The Recording Medium](#)
 - 3. [11.2.3 Reconstruction of the Original Wavefront](#)
 - 4. [11.2.4 Linearity of the Holographic Process](#)

5. [11.2.5 Image Formation by Holography](#)
3. [11.3 The Gabor Hologram](#)
 1. [11.3.1 Origin of the Reference Wave](#)
 2. [11.3.2 The Twin Images](#)
 3. [11.3.3 Limitations of the Gabor Hologram](#)
4. [11.4 The Leith-Upatnieks Hologram](#)
 1. [11.4.1 Recording the Hologram](#)
 2. [11.4.2 Obtaining the Reconstructed Images](#)
 3. [11.4.3 The Minimum Reference Angle](#)
 4. [11.4.4 Holography of Three-Dimensional Scenes](#)
 5. [11.4.5 Practical Problems in Holography](#)
5. [11.5 Image Locations and Magnification](#)
 1. [11.5.1 Image Locations](#)
 2. [11.5.2 Axial and Transverse Magnifications](#)
 3. [11.5.3 An Example](#)
6. [11.6 Some Different Types of Holograms](#)
 1. [11.6.1 Fresnel, Fraunhofer, Image, and Fourier Holograms](#)
 2. [11.6.2 Transmission and Reflection Holograms](#)
 3. [11.6.3 Holographic Stereograms](#)
 4. [11.6.4 Rainbow Holograms](#)
 5. [11.6.5 Multiplex Holograms](#)
 6. [11.6.6 Embossed Holograms](#)
7. [11.7 Thick Holograms](#)
 1. [11.7.1 Recording a Volume Holographic Grating](#)
 2. [11.7.2 Reconstructing Wavefronts from a Volume Grating](#)
 3. [11.7.3 Fringe Orientations for More Complex Recording Geometries](#)
 4. [11.7.4 Gratings of Finite Size](#)
 5. [11.7.5 Diffraction Efficiency—Coupled Mode Theory](#)
 1. [The Analysis](#)
 2. [Solution for a Thick Phase Transmission Grating](#)
 3. [Solution for a Thick Amplitude Transmission Grating](#)
 4. [Solution for a Thick Phase Reflection Grating](#)
 5. [Solution for a Thick Amplitude Reflection Grating](#)
 6. [Summary of Maximum Possible Diffraction Efficiencies](#)
8. [11.8 Recording Materials](#)
 1. [11.8.1 Silver Halide Emulsions](#)
 2. [11.8.2 Photopolymer Films](#)
 3. [11.8.3 Dichromated Gelatin](#)
 4. [11.8.4 Photorefractive Materials](#)

9. [11.9 Computer-Generated Holograms](#)

1. [11.9.1 The Sampling and Computation Problems](#)
2. [11.9.2 The Representational Problem](#)
 1. [Detour-Phase Holograms](#)
 2. [The Kinoform and the ROACH](#)
 3. [Phase Contour Interferograms](#)

10. [11.10 Degradations of Holographic Images](#)

1. [11.10.1 Effects of Film MTF](#)
 1. [Fourier Transform and Lensless Fourier Transform Holograms](#)
 2. [Generalization of the Geometry](#)
2. [11.10.2 Effects of Film Nonlinearities](#)
3. [11.10.3 Effects of Film-Grain Noise](#)
4. [11.10.4 Speckle Noise](#)

11. [11.11 Digital Holography](#)

1. [11.11.1 Offset Reference-Wave Digital Holography](#)
2. [11.11.2 Phase-Shifting Digital Holography](#)

12. [11.12 Holography with Spatially Incoherent Light](#)

13. [11.13 Applications of Holography](#)

1. [11.13.1 Microscopy and High-Resolution Volume Imagery](#)
2. [11.13.2 Interferometry](#)
 1. [Multiple-Exposure Holographic Interferometry](#)
 2. [Real-Time Holographic Interferometry](#)
 3. [Contour Generation](#)
 4. [Vibration Analysis](#)
3. [11.13.3 Imaging through Distorting Media](#)
4. [11.13.4 Holographic Data Storage](#)
5. [11.13.5 Holographic Weights for Artificial Neural Networks](#)
 1. [Model of a Neuron](#)
 2. [Networks of Neurons](#)
 3. [Optical Neural Networks Based on Volume Holographic Weights](#)
6. [11.13.6 Other Applications](#)
 1. [Holographic Optical Elements](#)
 2. [Holographic Display and Holographic Art](#)
 3. [Holograms for Security Applications](#)

14. [Problems - Chapter 11](#)

19. [12 Fourier Optics in Optical Communications](#)

1. [12.1 Introduction](#)
2. [12.2 Fiber Bragg Gratings](#)
 1. [12.2.1 Introduction to Optical Fibers](#)

- 2. [12.2.2 Recording Gratings in Optical Fibers](#)
 - 3. [12.2.3 Effects of an FBG on Light Propagating in the Fiber](#)
 - 1. [Phase Reflection Gratings](#)
 - 4. [12.2.4 Applications of FBGs](#)
 - 1. [Narrowband Filters for Add/Drop Multiplexers](#)
 - 2. [FBG Dispersion Compensators](#)
 - 5. [12.2.5 Gratings Operated in Transmission](#)
 - 3. [12.3 Ultrashort Pulse Shaping and Processing](#)
 - 1. [12.3.1 Mapping of Temporal Frequencies to Spatial Frequencies](#)
 - 2. [12.3.2 Pulse Shaping System](#)
 - 3. [12.3.3 Applications of Spectral Pulse Shaping](#)
 - 1. [Application to Code Division Multiple Access](#)
 - 2. [Application to Fiber Dispersion Compensation](#)
 - 4. [12.4 Spectral Holography](#)
 - 1. [12.4.1 Recording the Hologram](#)
 - 2. [12.4.2 Reconstructing the Signals](#)
 - 3. [12.4.3 Effects of Delay between the Reference Pulse and the Signal Waveform](#)
 - 5. [12.5 Arrayed Waveguide Gratings](#)
 - 1. [12.5.1 Component Parts of an Arrayed Waveguide Grating](#)
 - 1. [Integrated Optics Waveguides](#)
 - 2. [Integrated Star Couplers](#)
 - 3. [Waveguide Grating](#)
 - 4. [The Overall System](#)
 - 2. [12.5.2 Applications of AWGs](#)
 - 1. [Wavelength Multiplexers and Demultiplexers](#)
 - 2. [Wavelength Routers](#)
 - 6. [Problems - Chapter 12](#)
20. [A Delta Functions and Fourier Transform Theorems](#)
 - 1. [A.1 Delta Functions](#)
 - 2. [A.2 Derivation of Fourier Transform Theorems](#)
21. [B Introduction to Paraxial Geometrical Optics](#)
 - 1. [B.1 The Domain of Geometrical Optics](#)
 - 1. [The Concept of a Ray](#)
 - 2. [Rays and Local Spatial Frequency](#)
 - 2. [B.2 Refraction, Snell's Law, and the Paraxial Approximation](#)
 - 3. [B.3 The Ray-Transfer Matrix](#)
 - 1. [Elementary Ray-Transfer Matrices](#)
 - 4. [B.4 Conjugate Planes, Focal Planes, and Principal Planes](#)
 - 1. [Conjugate Planes](#)

- 2. [Focal Planes](#)
- 3. [Principal Planes](#)
- 5. [B.5 Entrance and Exit Pupils](#)
- 22. [C Polarization and Jones Matrices](#)
 - 1. [C.1 Definition of the Jones Matrix](#)
 - 2. [C.2 Examples of Simple Polarization Transformations](#)
 - 3. [C.3 Reflective Polarization Devices](#)
- 23. [D The Grating Equation](#)
- 24. [Bibliography](#)
- 25. [Index](#)

Landmarks

1. [Cover](#)
2. [Table of Contents](#)
3. [Begin Reading](#)
4. [A Delta Functions and Fourier Transform Theorems](#)
5. [B Introduction to Paraxial Geometrical Optics](#)
6. [C Polarization and Jones Matrices](#)
7. [D The Grating Equation](#)
8. [Bibliography](#)
9. [Index](#)

List of Illustrations

1. [Figure 2.1 Lines of zero phase for the function \$\exp\[j2\pi\(f_X x + f_Y y\)\]\$.](#)
2. [Figure 2.2 Special functions.](#)
3. [Figure 2.3 The circle function and its transform.](#)
4. [Figure 2.4 The spectrum of the finite chirp function, \$L_X = 20\$, \$\beta = 1\$.](#)
5. [Figure 2.5 The sampled function.](#)
6. [Figure 2.6 Spectra of \(a\) the original function and \(b\) the sampled data \(only three periods are shown in each direction for this infinitely periodic function\).](#)
7. [Figure 2.7 Illustration of aliasing. \(a\) A central slice through the two-dimensional spectrum shown in Fig. 2.6 \(a\). \(b\) A central slice through the spectral islands shown in Fig. 2.6 \(b\), representing the spectrum that results from sampling at the Nyquist rate. \(c\) The spectrum that results from sampling at 3/4 the Nyquist rate, showing the result of overlap of the spectral islands. In both \(b\) and \(c\), the dotted rectangles represent the transfer function of the interpolation filter. Clearly the original spectrum is not recovered in \(c\).](#)
8. [Figure 3.1 Snell's law at a sharp boundary \(\$n_2 > n_1\$ \).](#)
9. [Figure 3.2 The penumbra effect.](#)
10. [Figure 3.3 Arrangement used for observing diffraction of light.](#)
11. [Figure 3.4 Huygens envelope construction.](#)
12. [Figure 3.5 Surface of integration.](#)
13. [Figure 3.6 Kirchhoff formulation of diffraction by a plane screen.](#)
14. [Figure 3.7 Point-source illumination of a plane screen.](#)
15. [Figure 3.8 Rayleigh-Sommerfeld formulation of diffraction by a plane screen.](#)
16. [Figure 3.9 The wave vector \$\vec{k}\$.](#)
17. [Figure 4.1 Diffraction geometry.](#)
18. [Figure 4.2 Determining the sign of the phases of exponential representations of \(a\) spherical waves and \(b\) plane waves.](#)
19. [Figure 4.3 Magnitude of the integral of the quadratic-phase exponential function.](#)
20. [Figure 4.4 Light, dark, and transition regions behind a rectangular slit aperture.](#)
21. [Figure 4.5 Confocal spherical surfaces.](#)
22. [Figure 4.6 Input and output of an optical system with an arbitrary \$ABCD\$ matrix.](#)
23. [Figure 4.7 Cross section of the Fraunhofer diffraction pattern of a rectangular aperture.](#)
24. [Figure 4.8 The Fraunhofer diffraction pattern of a rectangular aperture \(\$\ell_X / \ell_Y = 2\$ \).](#)
25. [Figure 4.9 Cross section of the Fraunhofer diffraction pattern of a circular aperture.](#)
26. [Figure 4.10 Fraunhofer diffraction pattern of a circular aperture.](#)
27. [Figure 4.11 Amplitude transmittance function of the sinusoidal amplitude grating.](#)
28. [Figure 4.12 Fraunhofer diffraction pattern for a thin sinusoidal amplitude grating.](#)
29. [Figure 4.13 Fraunhofer diffraction pattern for a thin sinusoidal phase grating. The \$\pm 1\$ orders have nearly vanished in this example. Note that in this example there is some overlap of the diffraction orders.](#)

30. [Figure 4.14 Diffraction efficiency \$J_q^2\(m/2\)\$ vs. \$m/2\$](#) for three values of q .
31. [Figure 4.15 Normal Fresnel diffraction patterns at different distances from a square aperture.](#) The width of the diffraction pattern increases as the Fresnel number N_F shrinks. The width of the original rectangular aperture is indicated by the width of the shaded area.
32. [Figure 4.16 Fresnel diffraction patterns at different distances from a circular aperture.](#) The width of the diffraction pattern increases as the Fresnel number N_F shrinks. The diameter of the original circular aperture is indicated by the width of the shaded area. The patterns for $N_F = 10, 100$ and 1000 have a zero of intensity at their center.
33. [Figure 4.17 Geometry for diffraction calculation.](#)
34. [Figure 4.18 Locations of Talbot image planes behind the grating.](#)
35. [Figure 4.19 Overlap of the \$-1, 0, +1\$ diffraction orders deep in the Fresnel zone.](#)
36. [Figure 4.20 \(a\) Normalized beam width, \(b\) radius of curvature, and \(c\) Gouy phase as a function of \$z/z_0\$](#)
37. [Figure 4.21 Mode intensities for Hermite-Gaussian modes numbered \(left to right, top row\) \$\(0,0\), \(1,0\)\(0,1\)\$ and \(bottom row\) \$\(1,1\), \(2,2\), \(3,3\)\$.](#)
38. [Figure 4.22 Modal intensity and phase distributions for Laguerre-Gaussian beams with various mode numbers.](#) The top row shows the modal intensities, and the bottom row shows the corresponding modal phase distributions, with black representing 0 radians and white 2π radians. The abrupt transitions of the phase from white to black are a result of phase wrapping into the primary interval $(0, 2\pi)$.
39. [Figure P4.9](#)
40. [Figure P4.11](#)
41. [Figure P4.12](#)
42. [Figure P4.15](#)
43. [Figure P4.18](#)
44. [Figure 5.1 Equivalent width of the squared magnitude of the Fourier transform of the finite-width quadratic-phase exponential.](#) The straight solid lines are the asymptotes that are approached in the regions $N_F < 0.25$ ($N_F > 0.25$) (on the right) and $N_F > 0.25$ ($N_F < 0.25$) (on the left). On the right the bandwidth is determined by the quadratic-phase exponential function. On the left the bandwidth is determined by the finite size of the aperture over which that function is defined.
45. [Figure 5.2 Dependence of \$M, K\$ and \$N\$ on \$N_F\$ for the direct convolution method, as determined by simulations.](#) The two cases are (a) normalized intensity sidelobe level $\leq 10^{-2}$ at edge of the diffraction pattern and (b) normalized intensity sidelobe level $\leq 10^{-4}$ at edge of the pattern. The normalization is by the maximum value of the intensity in the diffraction pattern.
46. [Figure 5.3 Diffraction pattern intensity distributions obtained by the direct convolution method with \$N_F = 1\$, \$M = 150\$ and a normalized intensity aliasing criterion of \$10^{-4}\$.](#) (a) Linear plot of the central part of the diffraction pattern. (b) Logarithmic plot of the entire diffraction pattern.

47. Figure 5.4 Plots of N^N and M^M and Q^Q as a function of NF^N_F for the Fresnel transform approach when the normalized intensity level at the edge of the diffraction pattern is no greater than (a) 10^{-2} and (b) 10^{-4} .
48. Figure 5.5 (a) Values of M^M , N^N and Q^Q required for the Fresnel transfer function approach as a function of Fresnel number NF^N_F for aliasing intensity constraints (a) 10^{-2} and (b) 10^{-4}
49. Figure 5.6 Plots of central slices through the two-dimensional diffraction pattern intensity on a linear scale (left) and log scale (right) using the exact transfer function method for two values of $\Delta x/\lambda$: (a) $\Delta x/\lambda = 111$ $\Delta x/\lambda = 111$ and (b) $\Delta x/\lambda = 0.509$ $\Delta x/\lambda = 0.509$. For both cases, $NF = 10$ $N_F = 10$ and $M = 180$ $M = 180$. For (a), $Q = 4.5$ and $N = 810$ $N = 810$ were used, while for (b) $Q = 24$ $Q = 24$ and $N = 4320$ $N = 4320$ were used.
50. Figure 5.7 Computational complexities C_{2D}^D of the direct convolution approach and C_{2D}^{FFT} of the FFT approach to convolution for a normalized intensity aliasing criterion of (a) 10^{-2} and (b) 10^{-4} .
51. Figure 5.8 Computational complexities C_{2D}^{FFT} of the FFT approach to convolution and C_{2D}^{FRT} of the Fresnel transform approach to diffraction pattern calculation for normalized intensity aliasing criterions of (a) 10^{-2} and (b) 10^{-4} .
52. Figure 5.9 Computational complexities of the Fresnel transform approach C_{2D}^{FRT} and the Fresnel transfer function approach C_{2D}^{FTF} for a normalized intensity aliasing criterion of (a) 10^{-2} and (b) 10^{-4} .
53. Figure 5.10 Aperture consisting of a pair of rectangular openings.
54. Figure 5.11 Diffraction patterns of the double rectangle aperture for $NF_1 = 100$, 1 and 0.01 $N_{F1} = 100$, 1 and 0.01. The horizontal scales are different for the three patterns. $M_1^M_1$ has been chosen to be 10 when $NF_1 = 0.01$ $N_{F1} = 0.01$. The dashed curve represents the sinc^2 normalized intensity distribution due to one of the sub-apertures by itself.
55. Figure 5.12 Cross-sections of diffraction patterns calculated for a circular aperture when $NF = 10$. Part (a) shows the pattern obtained by numerical integration of (4-67), while part (b) shows the result obtained by the discrete projection and FFT approach described here.
56. Figure 6.1 The thickness function. (a) Front view, (b) side view
57. Figure 6.2 Calculation of the thickness function. (a) Geometry for $\Delta_1^{\Delta_1}$, (b) geometry for $\Delta_2^{\Delta_2}$, and (c) geometry for $\Delta_3^{\Delta_3}$.
58. Figure 6.3 Various types of lenses.
59. Figure 6.4 Effects of a converging lens and a diverging lens on a normally incident plane wave.
60. Figure 6.5 Geometries for performing the Fourier transform operation with a positive lens.

61. [Figure 6.6 Vignetting of the input. The shaded area in the input plane represents the portion of the input transparency that contributes to the Fourier transform at \$\(u_1, v_1\)\$.](#)
62. [Figure 6.7 Optically obtained Fourier transform of the character 3.](#)
63. [Figure 6.8 Geometry for image formation.](#)
64. [Figure 6.9 Converging illumination of the object.](#)
65. [Figure 6.10 Region of object space contributing to the field at a particular image point.](#)
66. [Figure 6.11 First problem analyzed.](#)
67. [Figure 6.12 Second problem analyzed.](#)
68. [Figure P6.2](#)
69. [Figure P6.3](#)
70. [Figure P6.4](#)
71. [Figure P6.7](#)
72. [Figure P6.8](#)
73. [Figure P6.9](#)
74. [Figure P6.10](#)
75. [Figure P6.11](#)
76. [Figure P6.15](#)
77. [Figure P6.18](#)
78. [Figure 7.1 Generalized model of an imaging system.](#)
79. [Figure 7.2 The Abbe theory of image formation.](#)
80. [Figure 7.3 Amplitude transfer functions for diffraction-limited systems with \(a\) square and \(b\) circular exit pupils.](#)
81. [Figure 7.4 Geometrical interpretations of the OTF of a diffraction-limited system. \(a\) The pupil function-total area is the denominator of the OTF; \(b\) two displaced pupil functions—the shaded area is the numerator of the OTF.](#)
82. [Figure 7.5 Light from patches separated by \$\(\lambda z_i |f_X|, \lambda z_i |f_Y|\)\$ interferes to produce a sinusoidal fringe at frequency \$\(f_X, f_Y\)\$. The shaded areas on the pupil are the areas within which the light patches can reside while retaining this special separation.](#)
83. [Figure 7.6 Calculation of the OTF for a square aperture.](#)
84. [Figure 7.7 The optical transfer function of a diffraction-limited system with a square pupil.](#)
85. [Figure 7.8 Calculation of the area of overlap of two displaced circles. \(a\) Overlapping circles, \(b\) geometry of the calculation.](#)
86. [Figure 7.9 The optical transfer function of a diffraction-limited system with a circular pupil. \(a\) Three-dimensional perspective, \(b\) cross section.](#)
87. [Figure 7.10 Geometry for defining the aberration function.](#)
88. [Figure 7.11 OTF for a focusing error in a system with a square pupil. \(a\) Three-dimensional plot with \$f_X/2f_o\$ along one axis and \$W_m/\lambda\$ along the other axis. \(b\) Cross section along the \$f_X\$ axis with \$W_m/\lambda\$ as a parameter. Note that only when \$W_m/\lambda > 0.5\$ does the OTF go negative over a certain frequency range.](#)
89. [Figure 7.12 \(a\) Focused and \(b\) misfocused images of a spoke target.](#)
90. [Figure 7.13 Geometrical optics prediction of the point-spread function of a system having a square pupil function and a severe focusing error.](#)
91. [Figure 7.14 Apodization of a rectangular aperture by a Gaussian function. \(a\) Intensity transmissions with and without apodization. \(b\) Point-spread functions with and without apodization.](#)

92. [Figure 7.15 Optical transfer functions with and without a Gaussian apodization.](#)
93. [Figure 7.16 Pupil amplitude transmittance and the corresponding OTF with and without a particular “inverse” apodization.](#)
94. [Figure 7.17 Calculation of the spectrum of the image intensity for object A.](#)
95. [Figure 7.18 Image intensity for two equally bright incoherent point sources separated by the Rayleigh resolution distance. Circular aperture assumed. The vertical lines show the locations of the two sources.](#)
96. [Figure 7.19 Image intensities for two equally bright coherent point sources separated by the Rayleigh resolution distance, with the phase difference between the two sources as a parameter. Circular aperture assumed. The vertical lines show the locations of the two point sources.](#)
97. [Figure 7.20 Images of a step in coherent and incoherent light. Circular pupil assumed.](#)
98. [Figure 7.21 Photographs of the image of an edge in \(a\) coherent and \(b\) incoherent illumination. \[From \[77\]. Copyright 1966 by the Optical Society of America, Inc., reprinted with permission.\]](#)
99. [Figure 7.22 Images illustrating the speckle effect. The object is a transparency illuminated through a diffuser. \(a\) Image in incoherent light. \(b\) Image in coherent light. \(c\) Close-up image of a particular letter in coherent light. \[Photo courtesy of P. Chavel and T. Avignon, Institut d’Optique.\]](#)
100. [Figure 7.23 Confocal reflection microscope geometry. The solid lines represent rays incident on and reflected from an in-focus object point. The dashed lines represent the rays reflected from an out-of-focus object point.](#)
101. [Figure 7.24 Optical transfer functions for a conventional microscope and a confocal microscope, assuming incoherent emission.](#)
102. [Figure P7.1](#)
103. [Figure P7.3](#)
104. [Figure P7.4](#)
105. [Figure P7.5](#)
106. [Figure P7.6](#)
107. [Figure P7.7](#)
108. [Figure P7.10](#)
109. [Figure P7.12](#)
110. [Figure 8.1 Geometries for calculating \(a\) depth of focus, \(b\) depth of field.](#)
111. [Figure 8.2 Exact MTFs with various levels of defocus and a cubic phase aberration. \(a\) MTFs with no cubic phase mask, defocus parameter \$W_{m2}/\lambda = 0, 0.5, 1.0, 1.5, 2.0,\$ and \$2.5.\$ \(b\) MTFs with defocus parameter fixed at \$W_{m2}/\lambda=2.0\$ \$W_{m2}/\lambda = 2.0\$, with the cubic phase parameter \$W_{m3}/\lambda = 4.0, 8.0, 10.0, 12.0,\$ and \$14.0.\$ While the curves for \$W_{m3}/\lambda > 4.0\$ \$W_{m3}/\lambda > 4.0\$ look as if they are approaching an asymptote in the mid-frequency range, in fact they are approaching the approximate values given by \(8-15\) and are inversely proportional to \$W_{m3}\$ \$\sqrt{W_{m3}}\$ \(c\) MTFs for a fixed cubic phase parameter \$W_{m3}/\lambda = W_{m3}/\lambda = 8,\$ with variable amounts of defocus. The defocus parameter \$W_{m2}/\lambda\$ \$W_{m2}/\lambda\$ for these curves takes the values \$0.5, 1.0, 1.5, 2.0,\$ and \$2.5.\$ The curves are almost indistinguishable, confirming that in the mid-frequency range, the MTF is, to a good approximation, independent of the amount of defocusing.](#)

112. Figure 8.3 Phase structure of the OTF when $W_{m2}/\lambda = 2.5$ and $W_{m3}/\lambda = 8$
113. Figure 8.4 Rotation of the intensity point-spread function as z/z_0 grows from 0 to 100.
 $z/z_0 = 0$ is the plane of best focus, and $z/z_0 = \pm 1$ are the boundaries of the Rayleigh range. (a) In this example, $2p/|l|=1$, and the four modes $(l,p)=(2,1),(4,2)(6,3)$ ($l, p) = (2, 1), (4, 2) (6, 3)$ and (8, 4) are added. The total rotation represented by these density plots is 180 degrees, so over $z=+\infty$ the total rotation is 360 degrees. However, the symmetry of the point-spread function restricts the unambiguous range to $z/z_0 = \pm \tan(\pi/4)$. (b) In this example, $2p/|l|=2$, and the four modes $(l,p)=(1,1),(2,2),(3,3)$, ($l, p) = (1, 1), (2, 2), (3, 3)$, and (4, 4) are added. The total rotation represented by these density plots is approximately 270 degrees, so over $z/z_0 = \pm \infty$, the total rotation is $1\frac{1}{2}$ times 360 degrees. Rotation can be restricted to 360 degrees if z is restricted to the range $z/z_0 = \pm \tan(\pi/3)$.
114. Figure 8.5 Geometry of the Lyot Coronagraph. The heavy lines on the left represent rays from the star, while the lighter solid lines represent rays from the planet. The light dotted lines to the right of the occulting stop represent paths the light from the star would have taken if the occulting stop were not present.
115. Figure 8.6 Simulation of the effects of the Lyot coronagraph. The planet/star intensity ratio is 10^{-6} in this example. (a) The telescope pupil (plane P_1) when starlight alone is incident. The simulation uses 400×400 pixels and the pupil has a diameter of 200 pixels. (b) Intensity distribution of the starlight incident on the initial image plane P_2 . (c) Starlight passed by the stop of diameter of 26 pixels in plane P_2 . (d) The resulting image of the pupil plane P_3 . The presence of the central stop in the image plane has diffracted light to the edges of the pupil. (e) The Lyot stop, having a clear opening with a diameter of 80 pixels. (f) An overexposed image of the star alone in plane P_4 . (g) Image of the planet alone in plane P_4 . The planet is assumed offset from the star by 50 image pixels. (h) Image of the star and the planet in plane P_2 . The planet is not detectable. (i) Image of the star and the planet in plane P_4 . The planet can be seen next to an overexposed image of the star. Note that in practice, the planet would be much closer to the star than the 50 image pixels assumed here, and the image of the planet would be in the midst of sidelobes of the image of the star, but for illustration purposes here we have chosen a larger separation.
116. Figure 8.7 (a) Plots of $g(x) = \psi_0(c, x)/\psi_0(c, 0)$ for $c=2, 10, 10, 50$. (b) Corresponding plots of the derivative of $g(x)$ with respect to x

- for the same choices of c . A large derivative corresponds to a hard edge that is nearly vertical.
117. Figure 8.8 The “cat’s eye” apodizing mask when $g(x) = \psi_0(10, x)/\psi_0(10, 0)$. Note there are no hard edges parallel to the vertical axis. R is the radius of the circular aperture.
 118. Figure 8.9 Intensity point-spread functions for no apodization and for cat’s-eye apodizations with $c=2, 10, 50$. The width of the main lobe when $c=10$ is about 2.7 times the width of the main lobe when there is no apodization.
 119. Figure 8.10 (a) Linear density plot of the full two-dimensional point-spread function of a system with a cat’s eye apodizing mask. The center of the PSF is strongly overexposed in this depiction. (b) Logarithmic plot of the same function, revealing some of the structure in the region where the point-spread function of the star would be suppressed. The intensity dynamic range represented in this figure is about 10^8 .
 120. Figure 8.11 (a) Object intensity distribution, and (b) object spectrum and the OTF.
 121. Figure 8.12 Geometry for synthetic aperture holography.
 122. Figure 8.13 Holographic camera for recording complex-valued images. M_1 and M_2 are mirrors, while BS_1 and BS_2 are beam splitters. The Gaussian illumination beam is split into two paths, the upper path being the reference path and the lower path the object path. The light passes through a grating on a piezo stage that is used to accurately translate the grating. The spatial filter selects certain orders of the grating and equalizes their strengths. The object is illuminated by these orders, ($0, \pm 1$ and ± 2 orders in the example to be given). The variable iris represents the finite entrance pupil of the system, which in this example corresponds to an $NA=0.063$. The multiplexed image falls on a CMOS detector, where it interferes with the tilted reference wave, yielding an interference pattern that encodes the phase. How to obtain the complex field from the hologram will be covered in Chapter 11.
 123. Figure 8.14 Reconstructed image spectrum log-magnitude when the $0, \pm 1$ and ± 2 orders of the grating illuminate the object. Coherent superresolution imaging via grating-based illumination, Jeffrey P. Wilde, Joseph W. Goodman, Yonina C. Eldar, and Yuzuru Takashima, *Appl. Opt.* **56**(1), A79–A88 (2017).
 124. Figure 8.15 Images obtained (a) without multiplexing and (b) with multiplexing. Coherent superresolution imaging via grating-based illumination, Jeffrey P. Wilde, Joseph W. Goodman, Yonina C. Eldar, and Yuzuru Takashima, *Appl. Opt.* **56**(1), A79–A88 (2017).
 125. Figure 8.16 Frequency domain depiction of the three object spectral islands and the OTF.
 126. Figure 8.17 Energy band diagram illustrating fluorescent emission.
 127. Figure 8.18 This figure shows a diffraction-limited image (upper left) on a super-resolution image (lower left) of immunolabeled microtubules in a BSC-1 cell over a 14x14 micron field of view. Three-dimensional information was determined by the double-helix point spread function technique. Reproduced from Hsiao-lu D. Lee, Steffen J. Sahl, Matthew D. Lew, and W. E. Moerner, “The double-helix microscope super-resolves extended biological structures by localizing single blinking molecules in three dimensions with nanoscale precision,” *Appl. Phys. Lett.* **100**, 153701 (2012), with permission of AIP Publishing.
 128. Figure 8.19 Geometry of the light field camera.

129. [Figure 8.20 Light field transformations: \(a\) the original light field, \(b\) light field after a Fourier transform, \(c\) light field after forward propagation by distance \$z\$, \(d\) light field after backward propagation by distance \$z\$.](#)
130. [Figure 8.21 Effects at the \$i\$ th super-pixel of misfocus on the light field: \(a\) the imaging geometries, \(b\) the light fields at the detector plane.](#)
131. [Figure 9.1 Structure of a photographic film or plate.](#)
132. [Figure 9.2 Pictorial representation of the photographic process. \(a\) Exposure, \(b\) latent image, \(c\) after development, and \(d\) after fixing. Only the emulsion is shown.](#)
133. [Figure 9.3 The Hurter-Driffield curve for a typical emulsion.](#)
134. [Figure 9.4 A liquid gate for removing film thickness variations. The thickness variations are greatly exaggerated.](#)
135. [Figure 9.5 Typical amplitude transmittance versus exposure curve.](#)
136. [Figure 9.6 The Kelley model of the photographic process. \(a\) Full model; \(b\) simplified model.](#)
137. [Figure 9.7 Measurement of the MTF by projecting back through the H&D curve.](#)
138. [Figure 9.8 Typical measured MTF curve.](#)
139. [Figure 9.9 A relief image produced by a tanning bleach. \(a\) Original density image, \(b\) Relief image after bleaching.](#)
140. [Figure 9.10 Fraction of the photoresist remaining versus the exposure dose \$D\$ \(positive resist assumed\).](#)
141. [Figure 9.11 Example of two rows of a four-level mask pattern.](#)
142. [Figure 9.12 Ideal sawtooth thickness profile for a blazed grating, and binary optic approximation to that profile \(\$N=2\$ \)](#)
143. [Figure 9.13 Diffraction efficiencies of various orders of a stepped approximation to a sawtooth grating. The parameter \$p^P\$ determines the particular diffraction order, with the order number given by \$p2N+1\$, and the number of discrete levels is \$2N^{2^N}\$.](#)
144. [Figure 9.14 Steps in the fabrication of a four-level binary optic element.](#)
145. [Figure 9.15 Molecular arrangements for different types of liquid crystals. \(a\) Nematic liquid crystal, \(b\) smectic liquid crystal, and \(c\) cholesteric liquid crystal. The layers in \(b\) and \(c\) have been separated for clarity. Only a small column of molecules is shown.](#)
146. [Figure 9.16 Molecular arrangements in a twisted nematic liquid crystal. The lines between the alignment layers indicate the direction of molecular alignment at various depths within the cell.](#)
147. [Figure 9.17 Ferroelectric liquid crystal \(a\) smectic-C* layered structure, and \(b\) allowed molecular orientations.](#)
148. [Figure 9.18 Structure of an electrically controlled liquid crystal cell.](#)
149. [Figure 9.19 Twisted nematic liquid crystal with a voltage applied. Only a small column of molecules is shown.](#)
150. [Figure 9.20 Ferroelectric liquid crystal molecules align in one of two allowed directions, depending on the direction of the field. The angles of orientation in the two states are separated by \$2\theta_t\$.](#)
151. [Figure 9.21 Intensity modulation with a reflective NLC cell.](#)
152. [Figure 9.22 Hughes liquid crystal SLM.](#)
153. [Figure 9.23 Electrical model for the optically written SLM. \$R_{PS}\$ and \$C_{PS}\$ are the resistance and capacitance of the photosensor, \$CDM\$ \$C_{DM}\$ is the capacitance of the dielectric](#)

- mirror, and RLC R_{LC} and CLC C_{LC} are the resistance and capacitance of the liquid crystal layer.
154. [Figure 9.24 Readout of the Hughes liquid crystal SLM with \(a\) no write light present, and \(b\) write light present.](#)
 155. [Figure 9.25 Deformable mirror pixel structures for \(a\) a membrane SLM and \(b\) a cantilever beam SLM.](#)
 156. [Figure 9.26 Torsion beam DMD">\(a\) top view and \(b\) side view.](#)
 157. [Figure 9.27 Acousto-optic cells operating in the \(a\) Raman-Nath regime and the \(b\) Bragg regime.](#)
 158. [Figure 9.28 Wave vector diagram for Bragg interaction. \$\vec{k}_i\$ is the incident optical wave vector, \$\vec{k}_1\$ is the optical wave vector of the component diffracted into the first diffraction order, and \$\vec{K}\$ is the acoustical wave vector.](#)
 159. [Figure P9.4 Profiles of ideal and quantized gratings.](#)
 160. [Figure 10.1 The Abbe-Porter experiment.](#)
 161. [Figure 10.2 Photograph of \(a\) the spectrum of mesh and \(b\) the original mesh.](#)
 162. [Figure 10.3 Mesh filtered with a horizontal slit in the focal plane. \(a\) Spectrum, \(b\) image.](#)
 163. [Figure 10.4 Mesh filtered with a vertical slit in the focal plane. \(a\) Spectrum, \(b\) image.](#)
 164. [Figure 10.5 Compensation for image blur. \(a\) Focal-plane filter; \(b\) transfer functions.](#)
 165. [Figure 10.6 Architectures for coherent optical information processing.](#)
 166. [Figure 10.7 Example of an anamorphic processor.](#)
 167. [Figure 10.8 Reachable regions of the frequency plane for \(a\) a purely absorbing filter, \(b\) an absorbing filter and binary phase control, \(c\) a pure phase filter, and \(d\) a filter that achieves arbitrary distributions of absorption and phase control.](#)
 168. [Figure 10.9 Recording the frequency-plane mask for a VanderLugt filter.](#)
 169. [Figure 10.10 Two alternative systems for producing the frequency-plane transparency \(a\) Modified Mach-Zehnder interferometer; \(b\) modified Rayleigh interferometer.](#)
 170. [Figure 10.11 Locations of the various terms of the processor output.](#)
 171. [Figure 10.12 The joint transform correlator. \(a\) Recording the filter, \(b\) obtaining the filtered output.](#)
 172. [Figure 10.13 Optical interpretation of the matched-filtering operation.](#)
 173. [Figure 10.14 Block diagram of a character-recognition system.](#)
 174. [Figure 10.15 Photographs of \(a\) the impulse response of a VanderLugt filter, and \(b\) the response of the filter to the letters Q, W, and P.](#)
 175. [Figure 10.16 Synthesis of a bank of matched filters with a single frequency-plane filter. \(a\) Recording the frequency-plane filter; \(b\) format of the matched filter portion of the output.](#)
 176. [Figure 10.17 Magnitudes of the transfer function of a Wiener filter. The image is assumed to have been blurred by a point-spread function consisting of a circular disk of radius w. The signal-to-noise ratio is varied from 1000 to 1. The phase of the filter changes between 0 and \$\pi\$ radians between alternate zeros of this transfer function.](#)
 177. [Figure 10.18 Deblurring of the blur point-spread function. \(a\) The original blur, \(b\) the magnitude of the deblur point-spread function, and \(c\) the point-spread function of the blur-deblur sequence. \[From \[343\]. Copyright 1975 by the Optical Society of America, Inc., reprinted with permission.\]](#)
 178. [Figure 10.19 Bragg cell spectrum analyzer.](#)
 179. [Figure 10.20 Acousto-optic space-integrating correlator.](#)

180. [Figure 10.21 Time-integrating correlator.](#)
181. [Figure 10.22 A fully parallel incoherent matrix-vector multiplier.](#)
182. [Figure 10.23 Optical elements comprising the parallel matrix-vector multiplier">\(a\) perspective view, \(b\) top view, and \(c\) side view.](#)
183. [Figure P10.1](#)
184. [Figure P10.10](#)
185. [Figure P10.11](#)
186. [Figure 11.1 Interferometric recording.](#)
187. [Figure 11.2 Wavefront reconstruction with \(a\) the original reference wave \$A\$ as illumination, and \(b\) the conjugate reference wave \$A^*\$ as illumination.](#)
188. [Figure 11.3 Imaging by wavefront reconstruction. \(a\) Recording the hologram of a point-source object; \(b\) generation of the virtual image; \(c\) generation of the real image.](#)
189. [Figure 11.4 Recording a Gabor hologram.](#)
190. [Figure 11.5 Formation of twin images from a Gabor hologram.](#)
191. [Figure 11.6 Recording a Leith-Upatnieks hologram.](#)
192. [Figure 11.7 Reconstruction of images from a Leith-Upatnieks hologram.](#)
193. [Figure 11.8 Spectra of \(a\) the object and \(b\) the hologram.](#)
194. [Figure 11.9 Holographic imaging of a three-dimensional scene. \(a\) Recording the hologram; \(b\) reconstructing the virtual image; \(c\) reconstructing the real image.](#)
195. [Figure 11.10 Photograph of a portion of a hologram of a diffuse three-dimensional scene.](#)
196. [Figure 11.11 Photographs showing the three-dimensional character of the virtual image reconstructed from a hologram.](#)
197. [Figure 11.12 Generalized \(a\) recording and \(b\) reconstruction geometries.](#)
198. [Figure 11.13 Recording a microwave hologram. The sources that provide the object and reference illuminations are derived from the same microwave signal generator to assure coherence.](#)
199. [Figure 11.14 Recording a lensless Fourier transform hologram.](#)
200. [Figure 11.15 \(a\) Recording a reflection hologram, and \(b\) reconstructing an image in reflected light.](#)
201. [Figure 11.16 Photograph of a virtual image reconstructed from a reflection hologram.](#)
202. [Figure 11.17 Recording a holographic stereogram \(top view\). \(a\) Recording the holograms, and \(b\) viewing the image.](#)
203. [Figure 11.18 The rainbow hologram. \(a\) The first recording step, and \(b\) the second recording step.](#)
204. [Figure 11.19 Reconstruction of the image from a rainbow hologram; \(a\) Reconstruction geometry, \(b\) slit sizes at different wavelengths.](#)
205. [Figure 11.20 Constructing a multiplex hologram. \(a\) Recording the still-frame sequence, and \(b\) recording the multiplex hologram. M indicates a mirror, BS a beam splitter. The reference wave arrives at the recording plane from above.](#)
206. [Figure 11.21 Viewing the image with a multiplex hologram.](#)
207. [Figure 11.22 Recording an elementary hologram with a thick emulsion.](#)
208. [Figure 11.23 Wave vector diagram illustrating the length and direction of the grating vector.](#)
209. [Figure 11.24 Reconstruction geometry.](#)
210. [Figure 11.25 Cone of incident wave vectors that satisfies the Bragg condition.](#)
211. [Figure 11.26 Orientation of interference fringes within a recording medium. \(a\) Two plane waves forming slant fringes, \(b\) a plane wave and a spherical wave, \(c\) two plane waves impinging from opposite sides of the emulsion, and \(d\) a plane wave and a spherical wave impinging from opposite sides of the recording medium.](#)

212. [Figure 11.27 Slice through the hyperboloids of fringe maxima for the case of two point sources. The dark lines represent interference fringes, while the lighter lines are the wavefronts.](#)
213. [Figure 11.28 Grating-vector clouds and their effect on closing the \$\vec{k}\$ -vector triangle. The dotted vectors correspond to \$\vec{k}\$ vectors when the grating is recorded, and the solid vectors correspond to the \$\vec{k}\$ vectors when reconstruction takes place. Changes of the lengths of the \$\vec{k}\$ vectors correspond to reconstruction at a different wavelength than was used for recording. In part \(a\), a change of the length of \$\vec{k} \rightarrow \vec{p}\$ does not prevent closure of the \$\vec{k}\$ -vector diagram. In part \(b\), a change of the angle of \$\vec{k} \rightarrow \vec{p}\$ does not prevent closure.](#)
214. [Figure 11.29 Geometry for analysis of a thick hologram.](#)
215. [Figure 11.30 Normalized intensities of the diffracted and undiffracted waves as a function of \$\Phi\$ for the Bragg matched case.](#)
216. [Figure 11.31 Diffraction efficiency of a thick phase transmission grating when Bragg mismatch is present.](#)
217. [Figure 11.32 Maximum possible Bragg matched diffraction efficiency versus \$\Phi_a'\$ for a thick amplitude transmission grating.](#)
218. [Figure 11.33 Diffraction efficiency of a thick amplitude transmission grating with Bragg mismatch.](#)
219. [Figure 11.34 Diffraction efficiency of a thick Bragg matched phase reflection grating.](#)
220. [Figure 11.35 Diffraction efficiency of a thick phase reflection grating when Bragg mismatch is present.](#)
221. [Figure 11.36 Bragg matched diffraction efficiency of a thick amplitude reflection grating.](#)
222. [Figure 11.37 Diffraction efficiency of a thick amplitude reflection hologram when Bragg mismatch is present.](#)
223. [Figure 11.38 Relations between \(a\) an incident sinusoidal intensity pattern and the resulting distributions of \(b\) charge, \(c\) electric field, and \(d\) refractive index change in a photorefractive material.](#)
224. [Figure 11.39 The detour-phase concept. The subcells are moved within a cell to control the phase of the transmitted light. Zero-phase lines of the reconstruction wavefront are shown.](#)
225. [Figure 11.40 A single cell in a detour-phase hologram.](#)
226. [Figure 11.41 \(a\) Binary detour-phase hologram; \(b\) image reconstructed from such a hologram. Courtesy of International Business Machines Corporation, © \(1969\) International Business Machines Corporation.](#)
227. [Figure 11.42 \(a\) The gray level display that leads to a kinoform, and \(b\) the image obtained from that kinoform. Courtesy of International Business Machines Corporation, © \(1969\) International Business Machines Corporation.](#)
228. [Figure 11.43 Plot of a phase contour interferogram for a quadratic-phase approximation to a spherical lens.](#)
229. [Figure 11.44 Nonlinear effects in holography for a diffuse object. \(a\) Images obtained under nearly linear recording conditions, and \(b\) images obtained under highly nonlinear recording conditions. \[From \[139\]. Copyright 1967 by the Optical Society of America. Reprinted with permission.\]](#)
230. [Figure 11.45 Triangular interferometer for incoherent holography.](#)

231. [Figure 11.46 Double-exposure holographic interferometry with a Q-switched ruby laser.](#) [Reproduced from L.O. Heflinger, R.F. Wuerker and R.E. Brooks, "Holographic Interferometry", *J. Appl. Phys.* 37, 642–649 (1966) with the permission of AIP Publishing]
232. [Figure 11.47 Contour generation by the two-source method.](#)
233. [Figure 11.48 Contour generation by the two-wavelength method.](#) [From [169]. Copyright 1967 by the Optical Society of America, Inc., reprinted with permission.]
234. [Figure 11.49 Recording a hologram of a vibrating object.](#)
235. [Figure 11.50 Holographic images of a diaphragm vibrating in two different modes.](#) [From [287]. Copyright 1965 by the Optical Society of America, Inc., reprinted with permission.]
236. [Figure 11.51 Use of the original distorting medium for compensating aberrations.](#) (a) Recording the hologram and (b) reconstructing the image.
237. [Figure 11.52 Use of a hologram compensating plate.](#) (a) Recording the compensating plate; (b) cancellation of the aberrations.
238. [Figure 11.53 Aberration-free imaging when the object and reference waves are identically distorted.](#) (a) Recording the hologram; (b) obtaining the image.
239. [Figure 11.54 Page-oriented holographic storage.](#)
240. [Figure 11.55 A volume holographic storage system.](#) The case of angle multiplexing is illustrated.
241. [Figure 11.56 \(a\) Model of a single neuron; \(b\) sigmoidal nonlinearity.](#)
242. [Figure 11.57 A four-layer neural network.](#)
243. [Figure 11.58 Illustration of a single weighted interconnection using a hologram.](#) In practice, many such interconnections would be realized simultaneously.
244. [Figure P11.1](#)
245. [Figure P11.5](#)
246. [Figure 12.1 A short section of fiber.](#)
247. [Figure 12.2 Two methods for recording an FBG">\(a\) interferometric method, and \(b\) phase grating method.](#)
248. [Figure 12.3 Plot of diffraction efficiency \$\eta\$ versus grating length \$\ell\$ and fractional wavelength de-tuning \$x\$ for \$\lambda_B=1550\$ nm, \$n_1=1.45\$ and \$\delta n=10^{-4}\$.](#)
249. [Figure 12.4 Response of a reflection grating with \$\delta n=10^{-4}\$ and \(free-space\) Bragg wavelength 1550 nm for \(a\) length 1 mm, \(b\) length 5 mm, \(c\) length 1 cm, \(d\) length 1 m.](#)
250. [Figure 12.5 Typical structure of an FBG add/drop multiplexer.](#)
251. [Figure 12.6 Dispersion compensation using a chirped FBG.](#)
252. [Figure 12.7 \(a\) Simple amplitude transmission grating. \(b\) Reflection grating.](#)
253. [Figure 12.8 Geometry for mapping optical frequency onto space.](#)
254. [Figure 12.9 Pulse shaping by spectral filtering.](#)
255. [Figure 12.10 Typical CDMA system.](#)
256. [Figure 12.11 Recording a Spectral hologram.](#)
257. [Figure 12.12 Fringes in the focal plane.](#)
258. [Figure 12.13 Reconstructing the temporal signal.](#)
259. [Figure 12.14 Architecture of an arrayed waveguide grating.](#)
260. [Figure 12.15 Cross section of a rectangular waveguide.](#)
261. [Figure 12.16 Star coupler showing \(a\) fan out from a particular input port to all output ports, and \(b\) fan in from all input ports to a particular output port. Similar operations happen for all input ports and all output ports simultaneously.](#)

262. [Figure 12.17 Star coupler geometry.](#) f is the radius of both circular arcs.
263. [Figure 12.18 Gratings in \(a\) free space and \(b\) waveguides.](#)
264. [Figure 12.19 AWG used as \(a\) demultiplexer and \(b\) multiplexer](#)
265. [Figure 12.20 Imaging analog to an AWG.](#)
266. [Figure 12.21 Illustration of AWG wavelength routing properties.](#) (a) Imaging λ_0 from central port to central port. (b) Imaging λ_0 from an off-center input port to an output port at the inverted image position. (c) Imaging wavelength $\lambda_1 = \lambda_0 + \delta\lambda$ from the central input port to an offset image port. (d) Imaging λ_1 from a top input port to a “wrapped-around” output port.
267. [Figure 12.22 Wavelength routing properties of an AWG.](#) The first wavelength subscript corresponds to the input port number, and the second subscript corresponds to the wavelength offset from λ_0 in increments of $\delta\lambda$.
268. [Figure P12.5](#)
269. [Figure B.1 Input and output of an optical system.](#)
270. [Figure B.2 Elementary structures for ray-transfer matrix calculations.](#) (a) Free space, (b) a planar interface, (c) a spherical interface, and (d) a thin lens.
271. [Figure B.3 Definition of focal points.](#) (a) Rear focal point of a positive lens, (b) front focal point of a positive lens, (c) front focal point of a negative lens, and (d) rear focal point of a negative lens.
272. [Figure B.4 Definitions of principal planes.](#) (a) First principal plane P_1 , (b) second principal plane P_2 .
273. [Figure B.5 Relations between principal planes, focal lengths, and object/image distances.](#)
274. [Figure B.6 Entrance and exit pupils.](#) (a) Entrance and exit pupils coincide with the physical pupil, (b) the exit pupil coincides with the physical pupil, and (c) the entrance pupil coincides with the physical pupil.
275. [Figure C.1 Coordinate rotation.](#) The direction of wave propagation is out of the page.
276. [Figure C.2 Reflective polarization device.](#)
277. [Figure D.1 Transmission grating geometry.](#)

List of Tables

1. [Table 2.1: Transform pairs for some functions separable in rectangular coordinates.](#)
2. [Table 4.1: Locations of maxima and minima of the Airy pattern.](#)
3. [Table 11.1: Maximum possible diffraction efficiencies of volume sinusoidal gratings.](#)

1. [i](#)
2. [ii](#)
3. [iii](#)
4. [iv](#)
5. [v](#)
6. [vi](#)
7. [vii](#)
8. [viii](#)
9. [ix](#)
10. [x](#)
11. [xi](#)
12. [xii](#)
13. [xiii](#)
14. [xiv](#)
15. [1](#)
16. [2](#)
17. [3](#)
18. [5](#)
19. [6](#)
20. [7](#)
21. [8](#)
22. [9](#)
23. [10](#)
24. [11](#)
25. [12](#)
26. [13](#)
27. [14](#)
28. [15](#)
29. [16](#)
30. [17](#)
31. [18](#)
32. [19](#)
33. [20](#)
34. [21](#)
35. [22](#)
36. [23](#)
37. [24](#)
38. [25](#)

- 39. [26](#)
- 40. [27](#)
- 41. [28](#)
- 42. [29](#)
- 43. [30](#)
- 44. [31](#)
- 45. [32](#)
- 46. [33](#)
- 47. [34](#)
- 48. [35](#)
- 49. [36](#)
- 50. [37](#)
- 51. [38](#)
- 52. [39](#)
- 53. [40](#)
- 54. [41](#)
- 55. [42](#)
- 56. [43](#)
- 57. [44](#)
- 58. [45](#)
- 59. [46](#)
- 60. [47](#)
- 61. [48](#)
- 62. [49](#)
- 63. [50](#)
- 64. [51](#)
- 65. [52](#)
- 66. [53](#)
- 67. [54](#)
- 68. [55](#)
- 69. [56](#)
- 70. [57](#)
- 71. [58](#)
- 72. [59](#)
- 73. [60](#)
- 74. [61](#)
- 75. [62](#)
- 76. [63](#)
- 77. [64](#)
- 78. [65](#)
- 79. [66](#)
- 80. [67](#)
- 81. [68](#)
- 82. [69](#)
- 83. [70](#)
- 84. [71](#)
- 85. [72](#)
- 86. [73](#)

- 87. [74](#)
- 88. [75](#)
- 89. [76](#)
- 90. [77](#)
- 91. [78](#)
- 92. [79](#)
- 93. [80](#)
- 94. [81](#)
- 95. [82](#)
- 96. [83](#)
- 97. [84](#)
- 98. [85](#)
- 99. [86](#)
- 100. [87](#)
- 101. [88](#)
- 102. [89](#)
- 103. [90](#)
- 104. [91](#)
- 105. [92](#)
- 106. [93](#)
- 107. [94](#)
- 108. [95](#)
- 109. [96](#)
- 110. [97](#)
- 111. [98](#)
- 112. [99](#)
- 113. [100](#)
- 114. [101](#)
- 115. [102](#)
- 116. [103](#)
- 117. [104](#)
- 118. [105](#)
- 119. [106](#)
- 120. [107](#)
- 121. [108](#)
- 122. [109](#)
- 123. [110](#)
- 124. [111](#)
- 125. [112](#)
- 126. [113](#)
- 127. [114](#)
- 128. [115](#)
- 129. [116](#)
- 130. [117](#)
- 131. [118](#)
- 132. [119](#)
- 133. [120](#)
- 134. [121](#)

- 135. [122](#)
- 136. [123](#)
- 137. [124](#)
- 138. [125](#)
- 139. [126](#)
- 140. [127](#)
- 141. [128](#)
- 142. [129](#)
- 143. [130](#)
- 144. [131](#)
- 145. [132](#)
- 146. [133](#)
- 147. [134](#)
- 148. [135](#)
- 149. [136](#)
- 150. [137](#)
- 151. [138](#)
- 152. [139](#)
- 153. [140](#)
- 154. [141](#)
- 155. [142](#)
- 156. [143](#)
- 157. [144](#)
- 158. [145](#)
- 159. [146](#)
- 160. [147](#)
- 161. [148](#)
- 162. [149](#)
- 163. [150](#)
- 164. [151](#)
- 165. [152](#)
- 166. [153](#)
- 167. [154](#)
- 168. [155](#)
- 169. [156](#)
- 170. [157](#)
- 171. [158](#)
- 172. [159](#)
- 173. [160](#)
- 174. [161](#)
- 175. [162](#)
- 176. [163](#)
- 177. [164](#)
- 178. [165](#)
- 179. [166](#)
- 180. [167](#)
- 181. [168](#)
- 182. [169](#)

- 183. [170](#)
- 184. [171](#)
- 185. [172](#)
- 186. [173](#)
- 187. [174](#)
- 188. [175](#)
- 189. [176](#)
- 190. [177](#)
- 191. [178](#)
- 192. [179](#)
- 193. [180](#)
- 194. [181](#)
- 195. [182](#)
- 196. [183](#)
- 197. [185](#)
- 198. [186](#)
- 199. [187](#)
- 200. [188](#)
- 201. [189](#)
- 202. [190](#)
- 203. [191](#)
- 204. [192](#)
- 205. [193](#)
- 206. [194](#)
- 207. [195](#)
- 208. [196](#)
- 209. [197](#)
- 210. [198](#)
- 211. [199](#)
- 212. [200](#)
- 213. [201](#)
- 214. [202](#)
- 215. [203](#)
- 216. [204](#)
- 217. [205](#)
- 218. [206](#)
- 219. [207](#)
- 220. [208](#)
- 221. [209](#)
- 222. [210](#)
- 223. [211](#)
- 224. [212](#)
- 225. [213](#)
- 226. [214](#)
- 227. [215](#)
- 228. [216](#)
- 229. [217](#)
- 230. [218](#)

- 231. [219](#)
- 232. [220](#)
- 233. [221](#)
- 234. [222](#)
- 235. [223](#)
- 236. [224](#)
- 237. [225](#)
- 238. [226](#)
- 239. [227](#)
- 240. [228](#)
- 241. [229](#)
- 242. [230](#)
- 243. [231](#)
- 244. [232](#)
- 245. [233](#)
- 246. [234](#)
- 247. [235](#)
- 248. [236](#)
- 249. [237](#)
- 250. [238](#)
- 251. [239](#)
- 252. [240](#)
- 253. [241](#)
- 254. [242](#)
- 255. [243](#)
- 256. [244](#)
- 257. [245](#)
- 258. [246](#)
- 259. [247](#)
- 260. [248](#)
- 261. [249](#)
- 262. [250](#)
- 263. [251](#)
- 264. [252](#)
- 265. [253](#)
- 266. [254](#)
- 267. [255](#)
- 268. [256](#)
- 269. [257](#)
- 270. [258](#)
- 271. [259](#)
- 272. [260](#)
- 273. [261](#)
- 274. [262](#)
- 275. [263](#)
- 276. [264](#)
- 277. [265](#)
- 278. [266](#)

- 279. [267](#)
- 280. [269](#)
- 281. [270](#)
- 282. [271](#)
- 283. [272](#)
- 284. [273](#)
- 285. [274](#)
- 286. [275](#)
- 287. [276](#)
- 288. [277](#)
- 289. [278](#)
- 290. [279](#)
- 291. [280](#)
- 292. [281](#)
- 293. [282](#)
- 294. [283](#)
- 295. [284](#)
- 296. [285](#)
- 297. [286](#)
- 298. [287](#)
- 299. [288](#)
- 300. [289](#)
- 301. [290](#)
- 302. [291](#)
- 303. [292](#)
- 304. [293](#)
- 305. [294](#)
- 306. [295](#)
- 307. [296](#)
- 308. [297](#)
- 309. [298](#)
- 310. [299](#)
- 311. [300](#)
- 312. [301](#)
- 313. [302](#)
- 314. [303](#)
- 315. [304](#)
- 316. [305](#)
- 317. [306](#)
- 318. [307](#)
- 319. [308](#)
- 320. [309](#)
- 321. [310](#)
- 322. [311](#)
- 323. [312](#)
- 324. [313](#)
- 325. [314](#)
- 326. [315](#)

- 327. [316](#)
- 328. [317](#)
- 329. [318](#)
- 330. [319](#)
- 331. [320](#)
- 332. [321](#)
- 333. [322](#)
- 334. [323](#)
- 335. [324](#)
- 336. [325](#)
- 337. [326](#)
- 338. [327](#)
- 339. [328](#)
- 340. [329](#)
- 341. [330](#)
- 342. [331](#)
- 343. [332](#)
- 344. [333](#)
- 345. [334](#)
- 346. [335](#)
- 347. [336](#)
- 348. [337](#)
- 349. [338](#)
- 350. [339](#)
- 351. [340](#)
- 352. [341](#)
- 353. [342](#)
- 354. [343](#)
- 355. [344](#)
- 356. [345](#)
- 357. [346](#)
- 358. [347](#)
- 359. [348](#)
- 360. [349](#)
- 361. [350](#)
- 362. [351](#)
- 363. [352](#)
- 364. [353](#)
- 365. [354](#)
- 366. [355](#)
- 367. [356](#)
- 368. [357](#)
- 369. [358](#)
- 370. [359](#)
- 371. [360](#)
- 372. [361](#)
- 373. [362](#)
- 374. [363](#)

375. [364](#)
376. [365](#)
377. [366](#)
378. [367](#)
379. [368](#)
380. [369](#)
381. [370](#)
382. [371](#)
383. [372](#)
384. [373](#)
385. [374](#)
386. [375](#)
387. [376](#)
388. [377](#)
389. [378](#)
390. [379](#)
391. [380](#)
392. [381](#)
393. [382](#)
394. [383](#)
395. [384](#)
396. [385](#)
397. [386](#)
398. [387](#)
399. [388](#)
400. [389](#)
401. [390](#)
402. [391](#)
403. [392](#)
404. [393](#)
405. [394](#)
406. [395](#)
407. [396](#)
408. [397](#)
409. [398](#)
410. [399](#)
411. [400](#)
412. [401](#)
413. [402](#)
414. [403](#)
415. [404](#)
416. [405](#)
417. [406](#)
418. [407](#)
419. [408](#)
420. [409](#)
421. [410](#)
422. [411](#)

- 423. [412](#)
- 424. [413](#)
- 425. [414](#)
- 426. [415](#)
- 427. [416](#)
- 428. [417](#)
- 429. [418](#)
- 430. [419](#)
- 431. [420](#)
- 432. [421](#)
- 433. [422](#)
- 434. [423](#)
- 435. [424](#)
- 436. [425](#)
- 437. [426](#)
- 438. [427](#)
- 439. [428](#)
- 440. [429](#)
- 441. [430](#)
- 442. [431](#)
- 443. [432](#)
- 444. [433](#)
- 445. [434](#)
- 446. [435](#)
- 447. [436](#)
- 448. [437](#)
- 449. [438](#)
- 450. [439](#)
- 451. [440](#)
- 452. [441](#)
- 453. [442](#)
- 454. [443](#)
- 455. [444](#)
- 456. [445](#)
- 457. [446](#)
- 458. [447](#)
- 459. [448](#)
- 460. [449](#)
- 461. [450](#)
- 462. [451](#)
- 463. [453](#)
- 464. [454](#)
- 465. [455](#)
- 466. [456](#)
- 467. [457](#)
- 468. [458](#)
- 469. [459](#)
- 470. [460](#)

471. [461](#)

472. [462](#)

473. [463](#)

474. [464](#)

475. [465](#)

476. [466](#)

477. [467](#)

478. [468](#)

479. [469](#)

480. [470](#)

481. [471](#)

482. [472](#)

483. [473](#)

484. [474](#)

485. [475](#)

486. [476](#)

487. [477](#)

488. [478](#)

489. [479](#)

490. [480](#)

491. [481](#)

492. [482](#)

493. [483](#)

494. [484](#)

495. [485](#)

496. [487](#)

497. [488](#)

498. [489](#)

499. [490](#)

500. [491](#)

501. [492](#)

502. [493](#)

503. [494](#)

504. [495](#)

505. [496](#)

506. [497](#)

507. [498](#)

508. [499](#)

509. [500](#)

510. [501](#)

511. [502](#)

512. [503](#)

513. [504](#)

514. [505](#)

515. [506](#)

516. [507](#)

517. [509](#)

518. [510](#)

- 519. [511](#)
- 520. [512](#)
- 521. [513](#)
- 522. [514](#)
- 523. [515](#)
- 524. [516](#)
- 525. [517](#)
- 526. [518](#)
- 527. [519](#)
- 528. [520](#)
- 529. [521](#)
- 530. [522](#)
- 531. [523](#)
- 532. [524](#)
- 533. [525](#)
- 534. [526](#)
- 535. [527](#)
- 536. [528](#)
- 537. [529](#)
- 538. [530](#)
- 539. [531](#)
- 540. [532](#)
- 541. [533](#)
- 542. [534](#)
- 543. [535](#)
- 544. [536](#)
- 545. [537](#)
- 546. [538](#)
- 547. [539](#)
- 548. [540](#)
- 549. [541](#)
- 550. [542](#)
- 551. [543](#)
- 552. [544](#)
- 553. [545](#)
- 554. [546](#)

Table of Contents

[Title Page](#)

[Dedication Page](#)

[Copyright Page](#)

[The Author](#)

[Preface](#)

[Contents](#)

[1 Introduction](#)

[1.1 Optics, Information, and Communication](#)

[1.2 The Book](#)

[2 Analysis of Two-Dimensional Signals and Systems](#)

[2.1 Fourier Analysis in Two Dimensions](#)

[2.1.1 Definition and Existence Conditions](#)

[2.1.2 The Fourier Transform as a Decomposition](#)

[2.1.3 Fourier Transform Theorems](#)

[2.1.4 Separable Functions](#)

[2.1.5 Functions with Circular Symmetry: Fourier-Bessel Transforms](#)

[2.1.6 Some Frequently Used Functions and Some Useful Fourier Transform Pairs](#)

[2.2 Spatial Frequency and Space-Frequency Localization](#)

[2.2.1 Local Spatial Frequencies](#)

[2.2.2 The Wigner Distribution Function](#)

[Real-valued Property](#)

[Shift Property](#)

[Multiplication by a Linear Exponential](#)

[Convolution Property](#)

[Multiplication Property](#)

[Magnification Property](#)

[Fourier Transform Property](#)

[2.3 Linear Systems](#)

[2.3.1 Linearity and the Superposition Integral](#)

[2.3.2 Invariant Linear Systems: Transfer Functions](#)

[2.4 Two-Dimensional Sampling Theory](#)

[2.4.1 The Whittaker-Shannon Sampling Theorem](#)

[2.4.2 Oversampling, Undersampling and Aliasing](#)

[2.4.3 Space-Bandwidth Product](#)

[2.5 The Discrete Fourier Transform](#)

[2.6 The Projection-Slice Theorem](#)

[2.7 Phase Retrieval from Fourier Magnitude](#)

[Problems - Chapter 2](#)

[3 Foundations of Scalar Diffraction Theory](#)

[3.1 Historical Introduction](#)

[3.2 From a Vector to a Scalar Theory](#)

[3.3 Some Mathematical Preliminaries](#)

[3.3.1 The Helmholtz Equation](#)

[3.3.2 Green's Theorem](#)

[3.3.3 The Integral Theorem of Helmholtz and Kirchhoff](#)

[3.4 The Kirchhoff Formulation of Diffraction by a Planar Screen](#)

[3.4.1 Application of the Integral Theorem](#)

[3.4.2 The Kirchhoff Boundary Conditions](#)

[3.4.3 The Fresnel-Kirchhoff Diffraction Formula](#)

[3.5 The Rayleigh-Sommerfeld Formulation of Diffraction](#)

[3.5.1 Choice of Alternative Green's Functions](#)

[3.5.2 The Rayleigh-Sommerfeld Diffraction Formula](#)

[3.5.3 Reproduction of Boundary Conditions](#)

[3.6 Kirchhoff and Rayleigh-Sommerfeld Theories Compared](#)

[3.7 Further Discussion of the Huygens-Fresnel Principle](#)

[3.8 Generalization to Nonmonochromatic Waves](#)

[3.9 Diffraction at Boundaries](#)

[3.10 The Angular Spectrum of Plane Waves](#)

[3.10.1 The Angular Spectrum and Its Physical Interpretation](#)

[3.10.2 Propagation of the Angular Spectrum](#)

[3.10.3 Effects of a Diffracting Aperture on the Angular Spectrum](#)

[3.10.4 The Propagation Phenomenon as a Linear Spatial Filter](#)

[Problems - Chapter 3](#)

[4 Fresnel and Fraunhofer Diffraction](#)

[4.1 Background](#)

[4.1.1 The Intensity of a Wave Field](#)

[4.1.2 The Huygens-Fresnel Principle in Rectangular Coordinates](#)

[4.2 The Fresnel Approximation](#)

[4.2.1 Positive vs. Negative Phases](#)

[4.2.2 Accuracy of the Fresnel Approximation](#)

[4.2.3 Finite Integral of the Quadratic-Phase Exponential Function](#)

[4.2.4 The Fresnel Approximation and the Angular Spectrum](#)

[4.2.5 Fresnel Diffraction Between Confocal Spherical Surfaces](#)

[4.2.6 Fresnel Diffraction in Terms of Ray Transfer Matrices](#)

[4.3 The Fraunhofer Approximation](#)

[4.4 Examples of Fraunhofer Diffraction Patterns](#)

[4.4.1 Rectangular Aperture](#)

[4.4.2 Circular Aperture](#)

[4.4.3 Thin Sinusoidal Amplitude Grating](#)

[4.4.4 Thin Sinusoidal Phase Grating](#)

[4.4.5 General Method for Calculating Diffraction Efficiency of Gratings](#)

[4.5 Examples of Fresnel Diffraction Calculations](#)

[4.5.1 Fresnel Diffraction by a Square Aperture](#)

[4.5.2 Fresnel Diffraction by a Circular Aperture](#)

[4.5.3 Fresnel Diffraction by a Sinusoidal Amplitude Grating-Talbot Images](#)

[4.6 Beam Optics](#)

[4.6.1 Gaussian Beams](#)

[4.6.2 Hermite-Gaussian Beams](#)

[4.6.3 Laguerre-Gaussian Beams](#)

[4.6.4 Bessel Beams](#)

Problems - Chapter 4

5 Computational Diffraction and Propagation

5.1 Approaches to Computational Diffraction

5.2 Sampling a Space-Limited Quadratic-Phase Exponential

5.3 The Convolution Approach

5.3.1 Bandwidth and Sampling Considerations

5.3.2 Discrete Convolution Equations

5.3.3 Simulation Results

5.3.4 Convolution by Fourier Transforms

5.4 The Fresnel Transform Approach

5.4.1 Sampling Increments

5.4.2 Sampling Ratio Q

5.4.3 Finding the Required M, Q, and N

5.4.4 The Discrete Diffraction Formulas

5.4.5 Examples of the Dependence of M and N on NF

5.4.6 Summary of Steps Using the Fresnel Transform Approach

5.4.7 Computational Complexity of the Fresnel Transform Approach

5.5 The Fresnel Transfer Function Approach

5.5.1 Sampling Considerations

5.5.2 Finding N, M and Q for each NF

5.5.3 The Discrete Diffraction Formulas

5.5.4 Examples of the Dependence of M, N and Q on NF

5.5.5 Summary of Steps Using the Fresnel Transfer Function Approach

5.5.6 Computational Complexity of the Fresnel Transfer Function Approach

5.6 The Exact Transfer Function Approach

5.6.1 Sampling in the Frequency Domain

5.6.2 Sampling in the Space Domain

5.6.3 Simulation Results

5.6.4 Computational Complexity of the Exact Transfer Function Approach

5.7 Comparison of Computational Complexities

5.8 Extension to More Complex Apertures

5.8.1 One-Dimensional Case

5.8.2 Two-Dimensional Apertures Separable in (x,y) Coordinates

5.8.3 Circularly-Symmetric Apertures

5.8.4 More General Cases

5.9 Concluding Comments

Problems - Chapter 5

6 Wave-Optics Analysis of Coherent Optical Systems

6.1 A Thin Lens as a Phase Transformation

6.1.1 The Thickness Function

6.1.2 The Paraxial Approximation

6.1.3 The Phase Transformation and Its Physical Meaning

6.2 Fourier Transforming Properties of Lenses

6.2.1 Input Placed against the Lens

6.2.2 Input Placed in Front of the Lens

[6.2.3 Input Placed behind the Lens](#)
[6.2.4 Example of an Optical Fourier Transform](#)
[6.3 Image Formation: Monochromatic Illumination](#)
[6.3.1 The Impulse Response of a Positive Lens](#)
[6.3.2 Eliminating Quadratic-Phase Factors: The Lens Law](#)
[6.3.3 The Relation between Object and Image](#)
[6.4 Analysis of Complex Coherent Optical Systems](#)
[6.4.1 The Ray Matrix Approach](#)
[6.4.2 Analysis of Two Optical Systems Using Ray Matrices](#)

[Problems - Chapter 6](#)

[7 Frequency Analysis of Optical Imaging Systems](#)
[7.1 Generalized Treatment of Imaging Systems](#)
[7.1.1 A Generalized Model](#)
[7.1.2 Effects of Diffraction on the Image](#)
[7.1.3 Polychromatic Illumination: The Coherent and Incoherent Cases](#)
[7.2 Frequency Response for Diffraction-Limited Coherent Imaging](#)
[7.2.1 The Amplitude Transfer Function](#)
[7.2.2 Examples of Amplitude Transfer Functions](#)
[7.3 Frequency Response for Diffraction-Limited Incoherent Imaging](#)
[7.3.1 The Optical Transfer Function](#)
[7.3.2 General Properties of the OTF](#)
[7.3.3 The OTF of an Aberration-Free System](#)
[7.3.4 Examples of Diffraction-Limited OTFs](#)
[7.4 Aberrations and Their Effects on Frequency Response](#)
[7.4.1 The Generalized Pupil Function](#)
[7.4.2 Effects of Aberrations on the Amplitude Transfer Function](#)
[7.4.3 Effects of Aberrations on the OTF](#)
[7.4.4 Example of a Simple Aberration: A Focusing Error](#)
[7.4.5 Apodization and Its Effects on Frequency Response](#)
[7.5 Comparison of Coherent and Incoherent Imaging](#)
[7.5.1 Frequency Spectrum of the Image Intensity](#)
[7.5.2 Two-Point Resolution](#)
[7.5.3 Other Effects](#)
[7.6 Confocal Microscopy](#)
[7.6.1 Coherent Case](#)
[7.6.2 Incoherent Case](#)
[7.6.3 Optical Sectioning](#)
[Problems - Chapter 7](#)
[8 Point-Spread Function and Transfer Function Engineering](#)
[8.1 Cubic Phase Mask for Increased Depth of Field](#)
[8.1.1 Depth of Focus](#)
[8.1.2 Depth of Field](#)
[8.1.3 The Cubic Phase Mask](#)
[8.2 Rotating Point-Spread Functions for Depth Resolution](#)
[8.3 Point-Spread Function Engineering for Exoplanet Discovery](#)
[8.3.1 The Lyot Coronagraph](#)
[8.3.2 Apodization for Starlight Suppression](#)

[8.4 Resolution beyond the Classical Diffraction Limit](#)

[8.4.1 Analytic Continuation](#)

[Underlying Mathematical Fundamentals](#)
[Intuitive Explanation of Bandwidth Extrapolation](#)

[8.4.2 Synthetic Aperture Fourier Holography](#)

[8.4.3 Fourier Ptychography](#)

[8.4.4 Coherent Spectral Multiplexing](#)

[8.4.5 Incoherent Structured Illumination Imaging](#)

[8.4.6 Super-Resolved Fluorescence Microscopy](#)

[Fluorescent Labelling](#)

[Localization Precision](#)

[Stimulated Emission Depletion Microscopy \(STED\)](#)

[Photoactivated Localization Microscopy \(PALM\) and Stochastic Optical Reconstruction Microscopy \(STORM\)](#)

[8.5 Light Field Photography](#)

[Problems - Chapter 8](#)

[9 Wavefront Modulation](#)

[9.1 Wavefront Modulation with Photographic Film](#)

[9.1.1 The Physical Processes of Exposure, Development, and Fixing](#)

[9.1.2 Definition of Terms](#)

[9.1.3 Photographic Film or Plate in Coherent Optical Systems](#)

[9.1.4 The Modulation Transfer Function](#)

[9.1.5 Bleaching of Photographic Emulsions](#)

[9.2 Wavefront Modulation with Diffractive Optical Elements](#)

[9.2.1 Single Step Lithography](#)

[9.2.2 Multistep Lithography](#)

[Approximation by a Stepped Thickness Function](#)

[The Fabrication Process](#)

[9.2.3 Other Types of Diffractive Optics](#)

[9.2.4 A Word of Caution](#)

[9.3 Liquid Crystal Spatial Light Modulators](#)

[9.3.1 Properties of Liquid Crystals](#)

[Mechanical Properties of Liquid Crystals](#)

[Electrical Properties of Liquid Crystals](#)

[Optical Properties of Nematic and Ferroelectric Liquid Crystals](#)

[9.3.2 Spatial Light Modulators Based on Liquid Crystals](#)

[Electrically Driven Liquid Crystal Spatial Light Modulators](#)

[Optically Driven Liquid Crystal Spatial Light Modulators](#)

[Ferroelectric Liquid Crystal Spatial Light Modulators](#)

[Liquid Crystal on Silicon \(LCOS\)](#)

[9.4 Deformable Mirror Spatial Light Modulators](#)

[9.5 Acousto-Optic Spatial Light Modulators](#)

[A CW Drive Voltage](#)

[A Modulated Drive Voltage](#)

[9.6 Other Methods of Wavefront Modulation](#)

[Problems - Chapter 9](#)

[10 Analog Optical Information Processing](#)

[10.1 Historical Background](#)

[10.1.1 The Abbe-Porter Experiments](#)

[10.1.2 The Zernike Phase-Contrast Microscope](#)

[10.1.3 Improvement of Photographs: Maréchal](#)

[10.1.4 Application of Coherent Optics to More General Data Processing](#)

[10.2 Coherent Optical Information Processing Systems](#)

[10.2.1 Coherent System Architectures](#)

[10.2.2 Constraints on Filter Realization](#)

[10.3 The VanderLugt Filter](#)

[10.3.1 Synthesis of the Frequency-Plane Mask](#)

[10.3.2 Processing the Input Data](#)

[10.3.3 Advantages of the VanderLugt Filter](#)

[10.4 The Joint Transform Correlator](#)

[10.5 Application to Character Recognition](#)

[10.5.1 The Matched Filter](#)

[10.5.2 A Character-Recognition Problem](#)

[10.5.3 Optical Synthesis of a Character-Recognition Machine](#)

[10.5.4 Sensitivity to Scale Size and Rotation](#)

[10.6 Image Restoration](#)

[10.6.1 The Inverse Filter](#)

[10.6.2 The Wiener Filter, or the Least-Mean-Square-Error Filter](#)

[10.6.3 Filter Realization](#)

[Inverse Filter](#)

[Wiener Filter](#)

[10.7 Acousto-Optic Signal Processing Systems](#)

[10.7.1 Bragg Cell Spectrum Analyzer](#)

[10.7.2 Space-Integrating Correlator](#)

[10.7.3 Time-Integrating Correlator](#)

[10.7.4 Other Acousto-Optic Signal Processing Architectures](#)

[10.8 Discrete Analog Optical Processors](#)

[10.8.1 Discrete Representation of Signals and Systems](#)

[10.8.2 A Parallel Incoherent Matrix-Vector Multiplier](#)

[10.8.3 Methods for Handling Bipolar and Complex Data](#)

[Problems - Chapter 10](#)

[11 Holography](#)

[11.1 Historical Introduction](#)

[11.2 The Wavefront Reconstruction Problem](#)

[11.2.1 Recording Amplitude and Phase](#)

[11.2.2 The Recording Medium](#)

[11.2.3 Reconstruction of the Original Wavefront](#)

[11.2.4 Linearity of the Holographic Process](#)

[11.2.5 Image Formation by Holography](#)

[11.3 The Gabor Hologram](#)

[11.3.1 Origin of the Reference Wave](#)

[11.3.2 The Twin Images](#)

[11.3.3 Limitations of the Gabor Hologram](#)

[11.4 The Leith-Upatnieks Hologram](#)

[11.4.1 Recording the Hologram](#)

[11.4.2 Obtaining the Reconstructed Images](#)

[11.4.3 The Minimum Reference Angle](#)

[11.4.4 Holography of Three-Dimensional Scenes](#)

[11.4.5 Practical Problems in Holography](#)

[11.5 Image Locations and Magnification](#)

[11.5.1 Image Locations](#)

[11.5.2 Axial and Transverse Magnifications](#)

[11.5.3 An Example](#)

[11.6 Some Different Types of Holograms](#)

[11.6.1 Fresnel, Fraunhofer, Image, and Fourier Holograms](#)

[11.6.2 Transmission and Reflection Holograms](#)

[11.6.3 Holographic Stereograms](#)

[11.6.4 Rainbow Holograms](#)

[11.6.5 Multiplex Holograms](#)

[11.6.6 Embossed Holograms](#)

[11.7 Thick Holograms](#)

[11.7.1 Recording a Volume Holographic Grating](#)

[11.7.2 Reconstructing Wavefronts from a Volume Grating](#)

[11.7.3 Fringe Orientations for More Complex Recording Geometries](#)

[11.7.4 Gratings of Finite Size](#)

[11.7.5 Diffraction Efficiency—Coupled Mode Theory](#)

[The Analysis](#)

[Solution for a Thick Phase Transmission Grating](#)

[Solution for a Thick Amplitude Transmission Grating](#)

[Solution for a Thick Phase Reflection Grating](#)

[Solution for a Thick Amplitude Reflection Grating](#)

[Summary of Maximum Possible Diffraction Efficiencies](#)

[11.8 Recording Materials](#)

[11.8.1 Silver Halide Emulsions](#)

[11.8.2 Photopolymer Films](#)

[11.8.3 Dichromated Gelatin](#)

[11.8.4 Photorefractive Materials](#)

[11.9 Computer-Generated Holograms](#)

[11.9.1 The Sampling and Computation Problems](#)

[11.9.2 The Representational Problem](#)

[Detour-Phase Holograms](#)

[The Kinoform and the ROACH](#)

Phase Contour Interferograms

11.10 Degradations of Holographic Images

11.10.1 Effects of Film MTF

Fourier Transform and Lensless Fourier Transform Holograms
Generalization of the Geometry

11.10.2 Effects of Film Nonlinearities

11.10.3 Effects of Film-Grain Noise

11.10.4 Speckle Noise

11.11 Digital Holography

11.11.1 Offset Reference-Wave Digital Holography

11.11.2 Phase-Shifting Digital Holography

11.12 Holography with Spatially Incoherent Light

11.13 Applications of Holography

11.13.1 Microscopy and High-Resolution Volume Imagery

11.13.2 Interferometry

Multiple-Exposure Holographic Interferometry
Real-Time Holographic Interferometry
Contour Generation
Vibration Analysis

11.13.3 Imaging through Distorting Media

11.13.4 Holographic Data Storage

11.13.5 Holographic Weights for Artificial Neural Networks

Model of a Neuron
Networks of Neurons
Optical Neural Networks Based on Volume Holographic Weights

11.13.6 Other Applications

Holographic Optical Elements
Holographic Display and Holographic Art
Holograms for Security Applications

Problems - Chapter 11

12 Fourier Optics in Optical Communications

12.1 Introduction

12.2 Fiber Bragg Gratings

12.2.1 Introduction to Optical Fibers

12.2.2 Recording Gratings in Optical Fibers

12.2.3 Effects of an FBG on Light Propagating in the Fiber Phase Reflection Gratings

12.2.4 Applications of FBGs

Narrowband Filters for Add/Drop Multiplexers
FBG Dispersion Compensators

12.2.5 Gratings Operated in Transmission

12.3 Ultrashort Pulse Shaping and Processing

12.3.1 Mapping of Temporal Frequencies to Spatial Frequencies

12.3.2 Pulse Shaping System

12.3.3 Applications of Spectral Pulse Shaping

Application to Code Division Multiple Access

[Application to Fiber Dispersion Compensation](#)

[12.4 Spectral Holography](#)

[12.4.1 Recording the Hologram](#)

[12.4.2 Reconstructing the Signals](#)

[12.4.3 Effects of Delay between the Reference Pulse and the Signal Waveform](#)

[12.5 Arrayed Waveguide Gratings](#)

[12.5.1 Component Parts of an Arrayed Waveguide Grating](#)

[Integrated Optics Waveguides](#)

[Integrated Star Couplers](#)

[Waveguide Grating](#)

[The Overall System](#)

[12.5.2 Applications of AWGs](#)

[Wavelength Multiplexers and Demultiplexers](#)

[Wavelength Routers](#)

[Problems - Chapter 12](#)

[A Delta Functions and Fourier Transform Theorems](#)

[A.1 Delta Functions](#)

[A.2 Derivation of Fourier Transform Theorems](#)

[B Introduction to Paraxial Geometrical Optics](#)

[B.1 The Domain of Geometrical Optics](#)

[The Concept of a Ray](#)

[Rays and Local Spatial Frequency](#)

[B.2 Refraction, Snell's Law, and the Paraxial Approximation](#)

[B.3 The Ray-Transfer Matrix](#)

[Elementary Ray-Transfer Matrices](#)

[B.4 Conjugate Planes, Focal Planes, and Principal Planes](#)

[Conjugate Planes](#)

[Focal Planes](#)

[Principal Planes](#)

[B.5 Entrance and Exit Pupils](#)

[C Polarization and Jones Matrices](#)

[C.1 Definition of the Jones Matrix](#)

[C.2 Examples of Simple Polarization Transformations](#)

[C.3 Reflective Polarization Devices](#)

[D The Grating Equation](#)

[Bibliography](#)

[Index](#)