

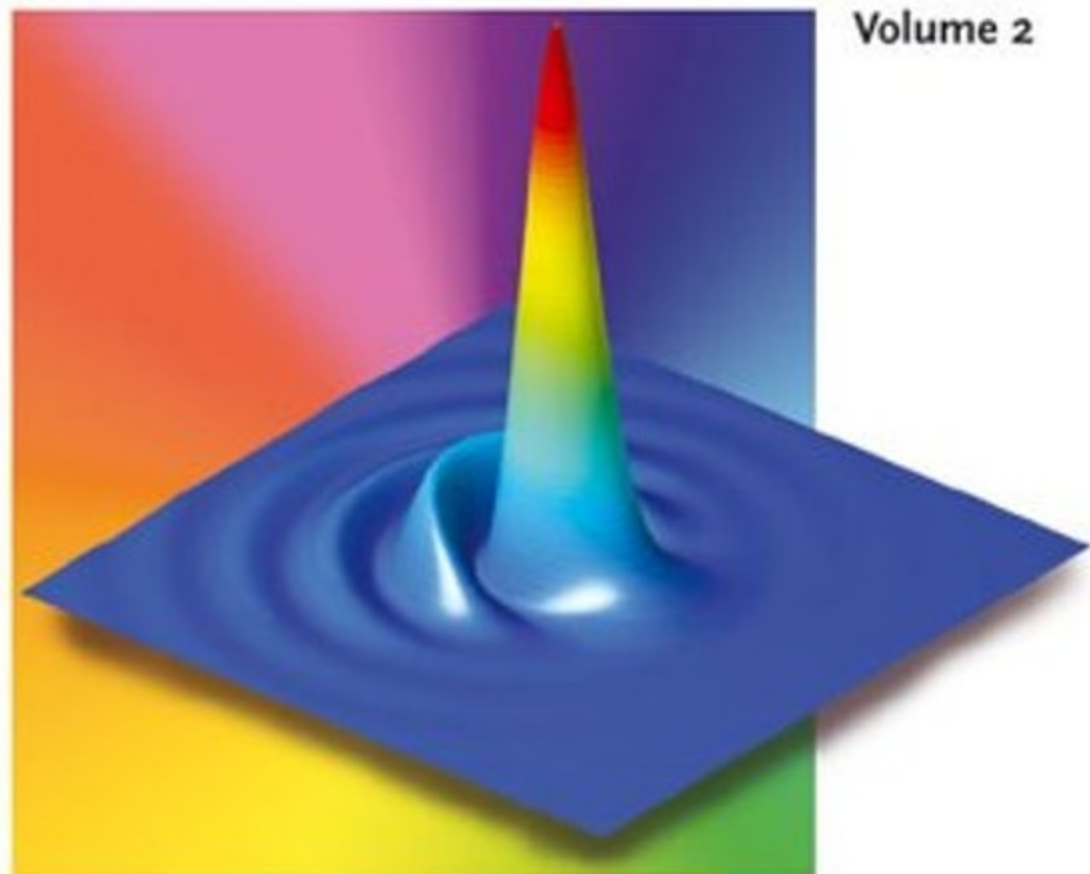
Edited by
Herbert Gross

 WILEY-VCH

Handbook of Optical Systems

W. Singer, M. Totzek, H. Gross
Physical Image Formation

Volume 2



Handbook of Optical Systems

Edited by
Herbert Gross

Volume 2: Physical Image Formation
Wolfgang Singer, Michael Totzeck, Herbert Gross



WILEY-VCH Verlag GmbH & Co. KGaA

Handbook of Optical Systems

Edited by Herbert Gross

Volume 2:

Physical Image Formation

Handbook of Optical Systems

Edited by
Herbert Gross

Volume 1: Fundamentals of Technical Optics

Volume 2: Physical Image Formation

Volume 3: Aberration Theory and Correction of Optical Systems

Volume 4: Survey of Optical Instruments

Volume 5: Metrology of Optical Components and Systems

Volume 6: Advanced Physical Optics



WILEY-
VCH

WILEY-VCH Verlag GmbH & Co. KGaA

Handbook of Optical Systems

Edited by
Herbert Gross

Volume 2: Physical Image Formation
Wolfgang Singer, Michael Totzeck, Herbert Gross



WILEY-VCH Verlag GmbH & Co. KGaA

Herbert Gross

Head of Optical Design Department
Carl Zeiss AG, Oberkochen, Germany
e-mail: Gross@zeiss.de

Wolfgang Singer

Department LIT-TD
Carl Zeiss SMT AG, Oberkochen, Germany
e-mail: singer@zeiss.de

Michael Totzeck

Department LIT-TD
Carl Zeiss SMT AG, Oberkochen, Germany
e-mail: m.totzeck@zeiss.de

■ All books published by Wiley-VCH are carefully produced. Nevertheless, authors, editors, and publisher do not warrant the information contained in these books, including this book, to be free of errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate.

Library of Congress Card No.:

applied for

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library.

Bibliographic information published by

Die Deutsche Bibliothek

Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at
<<http://dnb.ddb.de>>.

© 2005 WILEY-VCH Verlag GmbH & Co. KGaA,
Weinheim

All rights reserved (including those of translation into other languages). No part of this book may be reproduced in any form – nor transmitted or translated into machine language without written permission from the publishers. Registered names, trademarks, etc. used in this book, even when not specifically marked as such, are not to be considered unprotected by law.

Printed in the Federal Republic of Germany.

Printed on acid-free paper.

Cover Design 4t Matthes + Traut Werbeagentur
GmbH, Darmstadt

Typesetting Kühn & Weyh, Satz und Medien,
Freiburg

Printing Druckhaus Darmstadt GmbH, Darmstadt

Bookbinding Litges & Dopf Buchbinderei GmbH,
Darmstadt

ISBN 3-527-40378-7 (Vol. 2)

ISBN 3-527-40382-5 (Set)

Wolfgang Singer

Wolfgang Singer was born in 1964 and studied Physics at the University of Erlangen. He received his Ph.D. at the Institute of Applied Optics in 1995 with a thesis on microoptics, propagation theory and tomography. He spent his post doctorate at the Institute de Microtechnique in Neuchatel, where he developed diffractive diffusors for DUV illumination systems. From 1996 to 1998, he was assistant at the Institute of Applied Optics at the University of Stuttgart. Since 1998, he has been with Carl Zeiss SMT AG, working in the department of optical design and simulation for lithographic optics. His work includes tolerancing of objectives and the design of illumination systems of EUV systems. He became principal scientist and was engaged at the scientific training programme at Carl Zeiss. His special interests are imaging theory and partial coherence, and he has written his own simulation software. He holds 50 patents and has published about 30 papers and contributions to textbooks.

Michael Totzeck

Michael Totzeck was born in 1961. He received his diploma degree in Physics in 1987 and his Ph.D. in 1989, both from the Technical University of Berlin, where he also did his habilitation in 1995. In 1991 he was awarded the Carl-Ramsauer-Award of the AEG AG for his Ph.D. thesis on near field diffraction. From 1995 to 2002, he headed a group on high resolution microscopy at the Institute of Applied Optics in Stuttgart, working by experimental, theoretical and numerical means on optical metrology at the resolution limit. He has been with the Carl Zeiss SMT AG since 2002, working in the department for optical design. His current research topic is electromagnetic imaging with high-NA optical systems. He has published 40 papers on diffraction theory, near-field optics, high-resolution microscopy, interferometry, metrology, optical singularities, polarization-optics and physics education.

Herbert Gross

Herbert Gross was born in 1955. He studied Physics at the University of Stuttgart and joined Carl Zeiss in 1982. Since then he has been working in the department of optical design. His special areas of interest are the development of simulation methods, optical design software and algorithms, the modelling of laser systems and simulation of problems in physical optics, and the tolerancing and the measurement of optical systems. Since 1995, he has been heading the central optical design department at Zeiss. He served as a lecturer at the University of Applied Sciences at Aalen and at the University of Lausanne, and gave seminars for the Photonics Net of Baden Württemberg as well as several company internal courses. In 1995, he received his PhD at the University of Stuttgart on a work on the modelling of laser beam propagation in the partial coherent region. He has published several papers and has given many talks at conferences.

Contents of Volume I

1	Introduction	1
2	Paraxial Imaging	5
3	Interfaces	61
4	Materials	111
5	Raytracing	173
6	Radiometry	229
7	Light Sources	269
8	Sensor Technology and Signal Processing	323
9	Theory of Color Vision	379
10	Optical Systems	425
11	Aberrations	485
12	Wave Optics	523
13	Plano-optical Components	569
14	Gratings	647
15	Special Components	693
16	Optical Measurement and Testing Techniques	759

Contents

Introduction XIX

17	The Wave Equation	1
17.1	Introduction	2
17.2	From Maxwell to Helmholtz	2
17.2.1	Maxwell's Equations and the Inhomogeneous Wave Equation	2
17.2.2	Wave Equation in Homogeneous Media and the Scalar Wave Equation	4
17.2.3	The Dispersion Relation of the Harmonic Wave Solution	6
17.3	Elementary Waves in Free Space	9
17.3.1	The Electromagnetic Plane Wave	9
17.3.2	Spherical Wave	11
17.3.3	Dipole Wave	11
17.3.4	Radiated Field of a Harmonic Current Distribution	13
17.3.5	A Note on Plane and Spherical Waves	13
17.4	Energy, Irradiance and Intensity	14
17.5	The Angular Spectrum	17
17.5.1	Spatial Frequency Representation	17
17.5.2	Transformation of the Three-dimensional Spectrum into Two Dimensions	19
17.5.3	Free-space Propagation of Transverse Fields	20
17.5.4	Periodic Fields with Discrete Spectra	22
17.5.5	Boundary Conditions and the Spatial Frequency Spectrum	23
17.5.6	Vector Field Representation by Spatial Frequencies	24
17.6	Evanescence Waves	26
17.7	Approximative Solutions to the Wave Equation	28
17.7.1	Geometrical Optics and the Eikonal Equation	28
17.7.2	Paraxial Wave Equation	29
17.7.3	Transport of Intensity	30
17.7.4	Gaussian Beams	31
17.7.5	Ray Equivalent of Gaussian Beams	36
17.7.6	Gaussian Beams in Two Dimensions	37
17.8	Literature	39

18	Scalar Diffraction	41
18.1	Introduction	42
18.2	Kirchhoff Diffraction Integral	44
18.2.1	Inconsistency of the Kirchhoff Diffraction Integral	48
18.3	1 st and 2 nd Rayleigh–Sommerfeld Diffraction Integral	48
18.4	Two-dimensional Diffraction	50
18.5	Huygens Principle	52
18.6	Fourier Space Formulation	54
18.7	Examples of Scalar Diffraction Patterns	57
18.7.1	Diffraction Fields Behind Slits	57
18.7.2	Diffraction by a Rectangular Aperture	59
18.8	Fresnel Diffraction	60
18.8.1	Computation	61
18.8.2	Validity	62
18.9	Collin’s Fresnel Diffraction Integral	64
18.9.1	Definition	64
18.9.2	Example	67
18.10	Fraunhofer Diffraction	69
18.11	Grating Diffraction	71
18.11.1	Ronchi Grating	71
18.11.2	The Sinusoidal Phase Grating and Surface Fabrication Errors	76
18.12	Scalar Diffraction at Dielectric Objects	79
18.13	Babinet’s Principle	82
18.14	Scalar Scattering	85
18.15	Boundary Diffraction Waves	89
18.15.1	Geometrical Theory of Diffraction	90
18.15.2	An Empirical Boundary Diffraction Wave	94
18.16	Literature	96
19	Interference and Coherence	99
19.1	Basic Principles	100
19.1.1	Introduction	100
19.1.2	Two-beam Interference and Double Slit Diffraction	102
19.1.3	Contributions of Different Points of the Light Source	105
19.1.4	The High-frequency Term	107
19.1.5	The Low-frequency Term	108
19.1.6	Different Light Source Points with Statistical Phase	109
19.2	Mathematical Description of Coherence	113
19.2.1	Coherence Function	113
19.2.2	Wigner Distribution Function	116
19.2.3	Moments of the Wigner Distribution Function	120
19.2.4	Smoothing of the Wigner Distribution Function and Diffraction Focus	121
19.2.5	Wigner Distribution Function of Coherent Fields	122
19.2.6	Ambiguity Function	123

19.2.7	The Characterizing Functions in their Context	125
19.3	Temporal Coherence	126
19.3.1	Superposition of Signals with Different Frequency	126
19.3.2	Spectral Distribution of a Light Source	127
19.3.3	Bandwidth-limited Signals	128
19.3.4	Axial Coherence Length	130
19.3.5	Thermal Light Sources	133
19.3.6	Temporal Coherence in the Michelson Interferometer	134
19.4	Spatial Coherence	135
19.4.1	Introduction	135
19.4.2	Propagation of the Coherence Function	138
19.4.3	Van Cittert-Zernike Theorem	140
19.4.4	The Coherence Function of a Circular Source	140
19.4.5	Coherence Function behind a Double Slit	143
19.4.6	Propagation of the Wigner Distribution Function	146
19.5	Gaussian Schell Beams	149
19.5.1	Definition of Gaussian Schell Beams	149
19.5.2	Coherence and Wigner Functions of Gaussian Schell Beams	154
19.5.3	Basis Mode Expansion of Partial Coherent Fields	156
19.6	Statistical Optics and Speckle	159
19.6.1	Photon Statistics	159
19.6.2	The Speckle Effect	161
19.6.3	Speckle Parameters and Surface Structure	163
19.6.4	Computation of Speckle Effects	165
19.6.5	Speckle Reduction	169
19.7	Array Homogenizer	172
19.7.1	Setup of the System	172
19.7.2	Pupil Filling	175
19.7.3	Coherence Effects	176
19.7.4	Example Calculation	177
19.8	Miscellaneous	179
19.8.1	General Coherence Length	179
19.8.2	General Degree of Coherence	182
19.8.3	Coherence and Polarization	183
19.9	Literature	184
20	The Geometrical Optical Description and Incoherent Imaging	187
20.1	Introduction	188
20.2	Characteristic Functions	189
20.2.1	Geometrical Optics and the Wave Equation	189
20.2.2	The Characteristic Functions	191
20.2.3	Geometrical-optical imaging	194
20.2.4	The Canonical Pupil	196
20.2.5	A Note on Diffractive Optical Elements	199

20.3	The Ideal Wave-optical Image of a Point and Geometrical-optical Image Formation	200
20.3.1	The Scalar Luneburg Integral	200
20.3.2	Energy Discussions for Optical Imaging	204
20.3.3	The Airy Disc	206
20.3.4	Incoherent Resolution	210
20.4	Aberrations of Optical Systems	211
20.4.1	The Small-aberration Limit: The Strehl Ratio	211
20.4.2	Expansion of the Wave-front Error into Zernike Polynomials	212
20.4.3	Point Images for Different Aberrations	217
20.4.4	Distortion, Defocus and Astigmatism	219
20.4.5	Spherical Aberrations Z_9 , Coma Z_7 and Z_8	220
20.4.6	Line of Sight	221
20.4.7	Wave Aberrations for Annular Pupils	224
20.4.8	Extended Zernike Expansion	227
20.5	Helmholtz–Lagrange Invariant and Phase-space Description	231
20.5.1	The Phase Space	231
20.5.2	The Resolution Limit in the Space Domain and in the Spatial Frequency Domain	234
20.5.3	The Space–Bandwidth Product	236
20.6	Literature	237
21	The Abbe Theory of Imaging	239
21.1	Introduction	240
21.2	Phenomenological Description of Imaging	244
21.2.1	The Explanation of Image Formation According to Abbe and the Abbe Resolution	244
21.2.2	The Information About an Object Contained in an Image	249
21.2.3	Koehler Illumination and the Visibility	252
21.2.4	The Siedentopf Illumination Principle	255
21.2.5	Imaging with Different Colours	259
21.2.6	Aplanatic Correction and Geometrical Optics	260
21.3	The Mathematical Description of Fourier Optical Imaging	262
21.3.1	Imaging with Uncorrelated Light Sources	262
21.3.2	Consideration of Magnification	267
21.4	Coherence in Imaging	269
21.4.1	The Coherent Image	269
21.4.2	Incoherent Imaging	272
21.4.3	One-Dimensional Incoherent Imaging	273
21.4.4	Systems with Rotational Symmetry	275
21.4.5	Conditions for Incoherent, Partially Coherent and Coherent Imaging	277
21.4.6	Imaging with Correlated Light Sources	280
21.5	Literature	281

22	Coherence Theory of Optical Imaging	283
22.1	Introduction	284
22.2	Theoretical Description of Partially Coherent Image Formation	284
22.2.1	Hopkins Transmission Cross Coefficient	284
22.2.2	Image Fidelity	287
22.2.3	Hopkins Formalism for Periodic Objects	288
22.2.4	Aberrations in the Linear Grating Image	293
22.3	The Coherence Function and the Coherence Transfer Function	296
22.4	The Phase Space Description	300
22.4.1	Transformation of Coherence and Wigner Distribution Function	300
22.4.2	Propagation of the Wigner Distribution Function in Free Space	303
22.4.3	Compilation of the Transformations	307
22.5	Optical Imaging in the Presence of Aberrations	309
22.5.1	Linear Systems and Classification of Aberrations	309
22.5.2	Random Non-stationary Aberrations: Stray Light and Flare	314
22.6	Literature	317
23	Three-dimensional Imaging	319
23.1	Introduction	320
23.2	The Ewald Sphere and the Generalized Pupil	321
23.2.1	The Ewald Sphere	321
23.2.2	The Generalized Aperture and the Three-dimensional Point-spread Function	322
23.3	The Three-dimensional Transfer Function	327
23.3.1	Born Approximation and the Laue Equation	327
23.3.2	Dändliker's Representation and the Shape of the Three-dimensional Transfer Function	330
23.3.3	Resolution, Depth Resolution and Depth of Focus	335
23.3.4	3D-Transfer Functions in Microscopy	338
23.3.5	Magnification and a Comment on Absolute Instruments	340
23.4	Selected Examples of the Three-Dimensional Transfer Function	343
23.4.1	Transfer Function for Incoherent Imaging with $\sigma=1$	343
23.4.2	Partial Coherent Image Examples	344
23.4.3	'Tayloring' of the 3D-Transfer Function	346
23.4.5	Influence of Aberrations	351
23.5	Literature	352
24	Image Examples of Selected Objects	355
24.1	Introduction	356
24.2	Two-point Resolution	356
24.2.1	Incoherent Versus Coherent Two-point Resolution	356
24.2.2	Image of a Double Slit for Coherent and Incoherent Illumination	360
24.2.3	Phase Shift and Oblique Illumination	364
24.3	The Image of an Edge	365
24.3.1	The Coherent Image of an Amplitude and Phase Edge	365

24.3.2	The Incoherent Image of an Amplitude Edge	369
24.3.3	Partially Coherent Edge Image	370
24.3.4	The Determination of the Optical Transfer Function from the Edge Image	375
24.4	The Line Image	376
24.4.1	The Line Image of a Rotational-symmetrical Lens	376
24.4.2	Coherent Line or Slit Image	377
24.4.3	Incoherent Line or Slit Image	380
24.5	The Grating Image	381
24.5.1	The Coherent Linear Grating Image	381
24.5.2	The Coherent Grating Image with Aberrations	384
24.5.3	The Influence of the Coherence Parameter σ on the Grating Image	386
24.5.4	Influence of the Shape of the Effective Light Source on the Grating Image	389
24.5.5	Wigner Distribution Function for Gratings, Talbot Effect and Propagation-invariant Fields	394
24.6	Pinhole Imaging and Quasi-point Sources	399
24.6.1	Introduction	399
24.6.2	Incoherent Image of a Circular Object	400
24.6.3	Quasi-point Source	402
24.6.4	Pinhole with Coherent Illumination	404
24.6.5	Pinhole with Partial Coherent Illumination	405
24.6.6	Defocusing Planes and Deconvolution	406
24.7	Literature	407
25	Special System Examples and Applications	409
25.1	Introduction	410
25.2	Point-spread Functions for Annular Pupils	410
25.2.1	Introduction	410
25.2.2	Annular Pupils, Central Obscuration and Pupil Filters	411
25.3	Point-spread Functions of Non-uniform Illuminated Pupils	416
25.3.1	Introduction	416
25.3.2	General Gaussian Apodization	417
25.3.3	Gaussian Profile with Truncation	418
25.4	Engineering of the Point-spread Function by Pupil Masks	423
25.4.1	Introduction	423
25.4.2	Characterization of the Three-dimensional Point-spread Function	423
25.4.3	Characterization of Extended Depth of Focus	426
25.4.4	Relation Between Axial and Transverse Resolution	427
25.4.5	Ambiguity Function as Defocussed Transfer Function	429
25.4.6	Image Multiplexing	430
25.4.7	Fundamental Relationships	432
25.4.8	Calculation of Masks	432
25.5	Special Pupil Masks	433
25.5.1	Introduction	433

- 25.5.2 Phase Masks According to Toraldo 434
- 25.5.3 Logarithmic Phase Mask 435
- 25.5.4 Chirped Ring Pupil 437
- 25.5.5 Complex Filter Described by Zernike Expansions 439
- 25.5.6 Cubic Phase Plates for Extended Depth of Focus 442
- 25.5.7 Structured Illumination 447
- 25.6 Selected Practical Applications for Pupil Filtering Techniques 450
- 25.6.1 Phase Contrast Filtering, Dark-field Illumination 450
- 25.6.2 Frequency Doubling 453
- 25.6.3 Defect Filtering 455
- 25.6.4 Ronchi Test 456
- 25.7 Literature 463

- 26 Polarization 465**
- 26.1 Introduction 467
- 26.2 Polarization States 467
- 26.2.1 Representation of Polarization States 468
- 26.2.2 Jones Vector 468
- 26.2.3 Ellipse of Polarization 470
- 26.2.4 Orthogonal Jones Vectors 471
- 26.2.5 Jones Vectors in Different Bases 472
- 26.2.6 Unpolarized Light 472
- 26.2.7 Partial Polarization 473
- 26.2.8 Polarization Matrix 473
- 26.2.9 Stokes Vector 475
- 26.2.10 Poincaré Sphere 478
- 26.3 Jones Matrix 479
- 26.3.1 Definition 479
- 26.3.2 Jones Matrix Acting on a Jones Vector 480
- 26.3.3 Succession of Jones Matrices 480
- 26.3.4 Jones Matrix Acting on a Polarization Matrix 481
- 26.3.5 Examples of Jones Matrices 481
- 26.3.6 Rotated and Mirrored Jones Matrix 482
- 26.3.7 Jones Matrix for Different Basis Polarization States 483
- 26.3.8 Eigenpolarizations of a Jones Matrix 483
- 26.3.9 Jones Matrix of a Retarder 484
- 26.3.10 Jones Matrix of a Partial Polarizer 487
- 26.3.11 Pauli's Spin Matrices 489
- 26.3.12 Jones Matrix Decomposition 489
- 26.4 Müller Matrix 491
- 26.4.1 Definition 491
- 26.4.2 Examples 492
- 26.5 Müller–Jones Matrix 493
- 26.6 Light in Anisotropic Media 494
- 26.6.1 Anisotropic Media 494

26.6.2	Principal Refractive Indices of an Anisotropic Medium Without Spatial Dispersion and Optical Activity	495
26.6.3	Fresnel Ellipsoid	496
26.6.4	Index Ellipsoid	497
26.6.5	Types of Birefringent Media	497
26.7	Eigenwaves in Anisotropic Media	501
26.7.1	Plane Waves in Anisotropic Media	501
26.7.2	Eigenwaves and their Polarization	502
26.7.3	Properties of the Eigenpolarizations	506
26.7.4	The Intersection Ellipse	506
26.8	Jones Matrix of Propagation	507
26.9	Jones Matrices of Propagation for Common Media	508
26.9.1	Eigenpolarizations and ν -values	508
26.9.2	Coordinate Systems	509
26.9.3	Uniaxial Crystal	509
26.9.4	Biaxial Crystal	510
26.9.5	CaF ₂ with Spatial Dispersion at $\lambda = 193$ nm	511
26.10	Beam-splitting in an Anisotropic Medium	511
26.11	Examples of Polarization-optical Elements	516
26.11.1	Quarter-wave and Half-wave Retarder	516
26.11.2	Babinet–Soleil Compensator	516
26.11.3	Faraday Rotator	518
26.11.4	Brewster Plate	519
26.12	Literature	520
27	Vector Diffraction	523
27.1	Introduction	524
27.2	Focus Computation for Polarized Fields	525
27.2.1	Geometry for Focus Computation	525
27.2.2	Richards–Wolf integral	526
27.2.3	Plane Wave Expansion	531
27.2.4	Focus Fields for Various Input Polarizations	533
27.3	Vector Kirchhoff Diffraction Integral	538
27.4	Analytical Solutions	538
27.4.1	Plane Interface: Fresnel's Equations	540
27.4.2	Diffraction at a Circular Cylinder	542
27.4.3	Mie Scattering	547
27.5	Numerical Methods for Arbitrary Structures	553
27.6	Coupled Dipole Method	553
27.7	Integral Equation Approach and Moment Method	555
27.7.1	The Moment Method	555
27.7.2	Form of Scattering Operator	556
27.7.3	Scattering in Three-layer Medium	557
27.8	Fourier Modal Method	563
27.8.1	Theory	563

27.8.2	Diffraction Efficiency	568
27.9	Finite-difference Method	568
27.9.1	Boundary Conditions	570
27.9.2	Implicit Paraxial Wave Equation in Two Dimensions	572
27.9.3	Paraxial Wave Equation in Cylindrical Coordinates	572
27.9.4	ADI-formulation of the Paraxial Wave Equation in Three Dimensions	575
27.9.5	Split-step-beam Propagation Method	576
27.10	Rigorous Diffraction in Optical Imaging	579
27.10.1	Dielectrics and Metals	579
27.11	Simulation of Polarized Imaging by use of Rigorous Diffraction	583
27.12	Literature	587
28	Polarization and Optical Imaging	589
28.1	Introduction	590
28.2	The Image-forming Field	590
28.3	Interference of Electromagnetic Waves	592
28.3.1	Two-beam Vector Interference	592
28.3.2	Contrast for High-NA, s- and p-polarization	593
28.3.3	Influence of Recording Medium	594
28.3.4	Vector Effect in Optical Microscopy	595
28.3.5	Vector Effect in Optical Lithography	595
28.4	Polarized Ray Trace	596
28.4.1	Definition of Ray, Beam and Path	597
28.4.2	Ray-splitting at Anisotropic Elements	597
28.4.3	Refraction and Reflection at Birefringent Interfaces	598
28.4.4	The Single-path Approximation	599
28.5	Optical Systems with Polarization Effects	604
28.6	Polarized Imaging Model	605
28.6.1	Scalar Image	606
28.6.2	Vector Image for Completely Polarized Illumination	607
28.6.3	Vector Image for Partially Polarized Illumination	609
28.7	Vector Point-spread Function	610
28.7.1	VPSF for Complete Polarization	610
28.7.2	VPSF for Unpolarized Illumination	611
28.8	Polarized Optical Transfer Function	612
28.8.1	Polarized Illumination	612
28.8.2	Unpolarized Illumination	612
28.9	Jones Matrix Pupil	612
28.9.1	Definition for Completely Polarized Illumination	613
28.9.2	Separation of a Scalar Factor	614
28.9.3	Decomposition into Retardance and Diattenuation	615
28.9.4	Example	616
28.10	Jones Matrix Pupils in the Polarization Matrix Calculus	617
28.11	Jones-matrix-based System Optimization	619

28.12	Aberrations of the Transmitted Wavefront	620
28.13	Jones–Zernike Wavefront Aberrations	621
28.13.1	Principle of the Modal Characterization of a Jones Pupil	621
28.13.2	Jones–Zernike Expansion	621
28.13.3	Properties of the Jones–Zernike Polynomials	623
28.14	Literature	625
A1	Mathematical Appendix	627
A.1	Linear Systems	629
A.2	Fourier Series and Fourier Integral	631
A.2.1	Compilation of Basic Properties of the Fourier Transform	632
A.2.2	Special Functions and their Fourier Transforms	634
A.3	Convolution and Correlation	637
A.3.1	Convolution	637
A.3.2	Correlation	637
A.3.3	Power Spectral Density and RMS Value	638
A.4	Discrete Signals	639
A.4.1	The Sampling Theorem	639
A.4.2	Leakage	641
A.4.3	Indexing of the Numerical Discrete Fast Fourier Transform	642
A.5	z-Transform	644
A.5.1	Definition	644
A.5.2	Numerical Evaluation of the z-transform	646
A.5.3	Sinc Interpolation	648
A.6	Hankel Transform	648
A.6.1	Definition	648
A.6.2	Numerical Computation	649
A.7	Practical Calculation of Diffraction Integrals	655
A.7.1	The Oscillation Problem	655
A.7.2	Spatial and Spectral Resolution	660
A.7.3	Periodic Boundary Conditions	662
A.7.4	x-z Sampling of the Ewald Sphere	663
A.7.5	Equivalent Diffraction Setups	663
A.7.6	Optimal Conditioning of the Fresnel Diffraction	666
A.7.7	Numerical Algorithms	669
A.7.8	Fresnel Integrals	672
A.8	Orthogonal Polynomials on Rectangular Domains	675
A.8.1	Chebyshev Polynomials	675
A.8.2	One-dimensional Legendre Polynomials	677
A.8.3	Two-dimensional Chebyshev Polynomials	678
A.8.4	Legendre Polynomials in Two Dimensions	679
A.9	Literature	683
Index		685

Introduction

Physical image formation – no other subject is quite as closely connected with the name of Ernst Abbe. Ernst Abbe derived the first comprehensive theoretical description of optical image formation, he never published his work apart from a few remarks, which he made in 1873. Ernst Abbe died in 1905. It is therefore a privilege to dedicate this second volume of the Handbook of Optical Systems to the 100th anniversary of his death.

Abbe discovered the optical image as an interference phenomenon. The image intensity is interpreted by the interference of the coherent waves, which are formed by emission or diffraction at the object, and transmitted and transformed by the optical imaging system. In order to achieve an image that is similar to the object, the properties of the transformation need to follow certain rules, the first of which is the Abbe sine condition. In the case of non-self-luminous objects, the physical optical description of imaging also includes consideration of the illumination and coherence properties of the light source. In order to consider polarization and vector diffraction effects at the object, a complete vector theory is required. These theoretical descriptions have been developed or adapted from other disciplines of optics, sometimes triggered by the question of image formation. In this volume, we cover the most important topics within optical image formation.

Since the time of Abbe, the complexity of optical systems has increased considerably and the level of knowledge about them has improved a great deal. In modern systems, the use of laser light sources makes it necessary to consider the questions of coherence. Imaging at high numerical apertures, or the use of crystal materials, generates polarization effects. Digital image detection with pixellated solid-state sensors involves the consideration of sampled images. Modern image formation methods are frequently scanning methods, such as confocal imaging, where each object point is individually illuminated and imaged. The resulting image is obtained electronically, and further non-linear processes, such as digital filtering, are frequently applied. Here in Volume 2 we will concentrate on the physical effects involved in image formation, but computerized methods of imaging will be discussed in later volumes.

The general physical imaging problem is electromagnetic in its nature. The image is formed as an interference pattern of waves or a wave front, transmitted by an optical imaging system. For non-self-luminous objects, the waves that finally form

the image originate from diffraction at the object. Diffraction and interference can be conveniently described by wave optics. In a rather abstract way, the task of developing an optical system is to find a distribution of refractive indices and boundaries of the imaging system between the source and the detector that produce in the detector plane a wave field that is, as far as possible, like the object field. A purely geometrical formulation of the system description is not able to cover all the effects that take place. However, a full wave optical treatment of physical image formation is rarely possible. The dimensions and the complexity of optical systems typically do not allow for a rigorous solution of Maxwell's equations. Therefore, the solution is typically separated into different steps. The nature of the physical effects involved in image formation can be split into two different groups, and the steps are selected in accordance with the dominating physical effect, which might be either:

- the propagation of a wave field, i.e., through free space or optical systems, or
- an interaction, either with diffracting objects, detectors or the superposition of wave-fields to form interference patterns.

For each step, there are certain levels of approximate solutions, each of which can be solved by an appropriate solution for propagation or interaction, depending on the required accuracy. For example, there are several powerful methods, which are used in the calculation of field transport and propagation problems as numerical solutions to the wave equation. The remaining difficulties are mostly given by the interfaces between the different sections of the description. In the simplest case, different sampling grids have to be adapted by interpolation or re-sampling. In general, however, the transition is more complicated. While a transfer from ray optics to wave optics is straightforward, for example, transfer from wave to ray optics, generally requires further restrictions and is usually limited to coherent wave fronts. Here in Volume 2 of the Handbook series we will concentrate our description on the most relevant steps for optical imaging, such as diffraction at the object, consideration of partial coherence and vector effects. The standard description of optical imaging systems is shown in figure I-1.

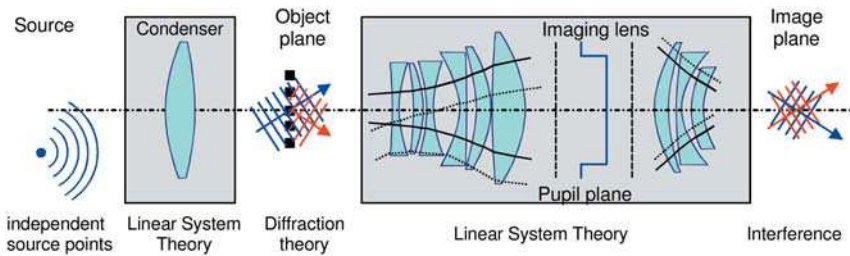


Figure I-1: Physical model for the description of optical imaging.

The physical model of optical imaging aims, whenever possible, to provide a linear system description. In the most convenient description, the effective light source is described by incoherent source points and the object is illuminated by an assembly of incoherent plane waves. Diffraction is often taken into account by the Thin

Element Approximation (TEA), which is equivalent to Fourier-transforming the object transmission function. Imaging is considered, in general, by a complex transfer function and the image intensity is formed by squaring the linearly transmitted amplitudes. The standard procedure is shown in table I-1.

Table I-1: Standard physical model of optical image formation

Step	Physical Principle	Preferred Methods of Solution	Simplification
Source	Linear decomposition	<ul style="list-style-type: none"> • Source points • Laser modes 	<ul style="list-style-type: none"> • Coherence properties
Condenser	Propagation	<ul style="list-style-type: none"> • Linear system theory 	<ul style="list-style-type: none"> • Fourier transformation of effective light source
Object	Interaction – Diffraction theory	<ul style="list-style-type: none"> • Thin element approximation • Scalar diffraction theory • Vector Diffraction theory 	<ul style="list-style-type: none"> • Decomposition of illumination into plane waves
Imaging lens	Propagation and filtering	<ul style="list-style-type: none"> • Linear System theory 	<ul style="list-style-type: none"> • Complex transfer function by ray-tracing
Image formation	Interference	<ul style="list-style-type: none"> • Intensity–square • Incoherent superposition 	<ul style="list-style-type: none"> • Time average

In the following the contents and also the limits of Volume 2 are outlined.

i) Coherence and Illumination

The amplitude spectrum in image space and the image intensity for coherent image formation, together with the diffraction spectrum of the object, are all obtained by a linear system approach. The only remaining question is how to consider the coherence properties of the light source. With coherent illumination of, e.g., a grating, the light propagation cancels out in certain directions, while it is constructively superposed in others. The image is formed as an interference phenomenon of the coherent diffracted waves. Illumination conditions, however, are typically not coherent, and in fact, coherent conditions usually have an adverse effect on image formation. The role of coherence is introduced in section 19.

The coherence properties of optical imaging are generally given by the properties of the light source and the illumination system. As will be shown in sections 21–24, the optical image is predominantly influenced by the physical properties of the illumination conditions. The description of the illumination in optical image formation is conveniently reduced to an effective light source. The illumination system will be considered as an optical system, with the effective light source being provided by a physical light source such as a laser beam. Inside the illumination system, the use of components with micro-structured surfaces eventually makes it necessary to cal-

culate the interaction of the light propagation and the light transfer using non-geometrical models, and then the transition between wave and ray optics is more complicated. The illumination systems and the physical effects involved will be discussed in Volume 4. The discussion of physical image formation in Volume 2 is – as usual – limited to the effective light source model.

ii) Diffraction – Coherent Interaction of Light and Matter

One of the consequences of Abbe's theory is that the information contained in the image is restricted to the information which is contained in the light scattered by the object. For most applications, a scalar description of scattering is sufficient, but further approximations may be applied for convenience. Diffraction, however, can be considered in physical optical imaging in two ways. On the one hand, as far as non self-luminous objects are concerned, diffraction theory is required for the treatment of the interaction of light with the object to be imaged. On the other hand, diffraction at the aperture stop of an imaging lens causes a diffraction-limited spot size. In general the task is to compute the diffraction field of either object or aperture stop, depending on the incident light amplitude distribution. There are several methods used for physically modelling the interaction of light with objects and the method of choice has to be decided upon depending on the appropriate conditions for numerical solution. Possible algorithms are, for example, the Kirchhoff integral, the Fresnel paraxial approximation or the Fraunhofer far-field approximation.

Light is a transverse electromagnetic wave – in the classical limit – and obeys Maxwell's equations. For this reason it can be polarized. Polarization has an impact on optical imaging because the contrast of the interference depends on the mutual polarization of the contributing waves. Furthermore, polarization effects can be used to generate an image contrast as in polarization microscopy and some projection devices. The mathematical description of polarization states and their change due to various optical components is the subject of section 26, while polarization optical imaging is discussed in chapter 28.

If either the level of accuracy required to describe the diffraction at the object becomes extraordinarily high or if the structure sizes become comparable to the wavelength, then scalar approximation theories are no longer sufficient. Some frequently-used rigorous methods that solve Maxwell's equations in object space are treated in chapter 27. However, apart from some examples, it is beyond the scope of this volume to consider near-field effects of the interaction of light with small structures in detail, as is required, for instance, in the interpretation of the images in near-field optical microscopy. Some common near-field optical measurement methods will be discussed in Volume 6.

iii) Propagation of Coherent Light in Optical Systems

If the propagators are chosen in an appropriate way and the sampling conditions are considered properly, the field propagation through nearly every real system can be calculated to an acceptable accuracy. To illustrate the complexity of the propagation problem, a simple example of an optical system is considered. Propagation through optical systems, including interfaces, is frequently simplified by a geomet-

rical optical description, without the need to sacrifice accuracy. Refraction or reflection follows well-known laws, but dielectric coatings might influence the amplitude and the phase, for example. The complex transmission and reflection coefficients generally require a wave optical treatment. These wave optical effects have to be included in the geometrical optical description by, e.g., reference tables or functional descriptions of the coefficients. The consideration of coefficients alone might not be sufficient since, e.g., for multi-layer coatings with a thickness of several wavelengths the position of the reflection or refraction of a ray is not clear. This simple example shows that the level of detail involved in the steps might be quite complex. A full description of all details is beyond the scope of Volume 2. In section 20, however, the formal transition from geometrical optics to wave optics is discussed for the example of imaging systems.

For a wave optical treatment, the effect of the phase elements such as thin lenses or phase filters of an optical system can be considered as for complex filters. The thin phase element is considered by a simple projection of the phase shift of the element. The role of the z -dimension is neglected and the output wave is obtained from the incident wave by multiplication with a complex transfer function, given by the amplitude transmission A of the element and the phase shift Φ , which is added to the phase of the incident wave. In general, however, the complex transmission function depends on the angle of incidence. It should be noted that this TEA, although called a thin-element approximation, is really a small-angle approximation. The TEA is frequently applied in so-called Fourier Optics, which is a simplified wave optical treatment of image formation and optical filter techniques. Within the scope of this approximation, compound optical systems may be treated using the paraxial Collins integral, based on the Fresnel approximation and then by applying the matrix approach of first-order optics. Separated parts of an optical system with nearly-perfect correction can be described as paraxial lens groups. The characterization of the subsystem is possible by a simple paraxial ABCD matrix. An extension of the approximation is possible for small residual aberrations of the optical system. This approach, based on Fresnel's integral, is outlined in section 18. The paraxial formulation of Fourier Optics, however, frequently leads to the incorrect assumption that the application of optical imaging theory – and Fourier Optics in general – is restricted to the paraxial regime. In this volume we will concentrate on general optical imaging theory with no lack of approximations. Fortunately, wave-optical propagators by optical imaging systems are not necessary, since the treatment using geometrical optical approximation produces results with a very high accuracy. The application of wave optical propagators is restricted to special applications and will not be considered further in Volume 2.

In addition, the design and analysis of optical imaging systems is frequently also based on ray tracing and the geometrical optical description. The advantages of ray tracing are many: it is powerful, fast, flexible and applicable over large scales, and there are a variety of commercially available software packages in existence. With certain approximations and, e.g., statistical methods, it is possible to consider even scattering or diffractive optical elements. However, for a comprehensive treatment of optical design, simulated by ray tracing, the reader is referred to Volume 3.

iv) Conversion between Rays and Waves

For a description of physical image formation it is necessary to achieve the transfer between geometrical ray optics and wave optics. As will be shown in section 20, conversion from rays to waves and vice versa is generally only possible for coherent wave fields obeying the law of Malus – a property called orthotomy. Both forms of description are then equivalent and can be converted into each other. The conditions can be summarized as:

1. A wave front must exist. This wave front must be continuous, and there should be no phase singularities.
2. The conversion is generally not possible in the region of a caustic, where wave fronts cannot be defined.

Having defined a coherent wave field $U(x,y,z)$ by a complex amplitude distribution with the phase $\Phi(x,y,z)$ and amplitude distribution $A(x,y,z)$, the local direction of light rays is directly given by the Eikonal equation (see sections 17 and 20). As a second condition, the transfer between the two forms of description has to satisfy the conservation of energy. For this purpose, a ray is either characterized by a weighting coefficient g or – in a statistical approach – the ray density is selected in accordance with the amplitude A . For a more general treatment of the transfer from ray optics to wave optics and vice versa we refer you to a later volume of this Handbook of Optical Systems.

Ultimately the wave optical treatment of propagation through imaging systems is preferred or sometimes even necessary. For micro-optical systems, numerical methods may be applied, such as finite differences for small distances and gradient index media, or mode-expansion methods inside light-guiding structures, such as waveguides. In Volume 6, we address beam-propagation methods in order to model the propagation of light in guiding structures and more complex environments, and also the description of laser light.

The image formed as an interference phenomenon is considered by a wave optical description, either in the space domain by a superposition of, e.g., plane waves; or in the spatial frequency domain by the spatial frequency spectrum representation of the wave field in the image plane. For an introduction, see the wave optical description which is outlined in section 17.

17

The Wave Equation

17.1	Introduction	2
17.2	From Maxwell to Helmholtz	2
17.2.1	Maxwell's Equations and the Inhomogeneous Wave Equation	2
17.2.2	Wave Equation in Homogeneous Media and the Scalar Wave Equation	4
17.2.3	The Dispersion Relation of the Harmonic Wave Solution	6
17.3	Elementary Waves in Free Space	9
17.3.1	The Electromagnetic Plane Wave	9
17.3.2	Spherical Wave	11
17.3.3	Dipole Wave	11
17.3.4	Radiated Field of a Harmonic Current Distribution	13
17.3.5	A Note on Plane and Spherical Waves	13
17.4	Energy, Irradiance and Intensity	14
17.5	The Angular Spectrum	17
17.5.1	Spatial Frequency Representation	17
17.5.2	Transformation of the Three-dimensional Spectrum into Two Dimensions	19
17.5.3	Free-space Propagation of Transverse Fields	20
17.5.4	Periodic Fields with Discrete Spectra	22
17.5.5	Boundary Conditions and the Spatial Frequency Spectrum	23
17.5.6	Vector Field Representation by Spatial Frequencies	24
17.6	Evanescence Waves	26
17.7	Approximative Solutions to the Wave Equation	28
17.7.1	Geometrical Optics and the Eikonal Equation	28
17.7.2	Paraxial Wave Equation	29
17.7.3	Transport of Intensity	30
17.7.4	Gaussian Beams	31
17.7.5	Ray Equivalent of Gaussian Beams	36
17.7.6	Gaussian Beams in Two Dimensions	37
17.8	Literature	39

17.1

Introduction

In the classical limit, light is an electromagnetic wave. Consequently, classical optics can be derived from Maxwell's equations together with the material equations. Excluded from this are quantum effects in the interaction with media (namely quantum mechanics) and in the electromagnetic field itself (namely quantum electrodynamics). The intention of this section is to state clearly the assumptions and approximations at the transition from Maxwell's equations to the monochromatic homogeneous Helmholtz equation.

One central point of the present chapter is the wave equation that describes in differential form the development of waves in space and time. Only in homogeneous and isotropic media can the coupled vector wave equation be separated into scalar wave equations with plane or spherical waves as solutions. The phenomenon of evanescent waves as non-propagating solutions of the wave equation can be observed especially at interfaces. In inhomogeneous media and in anisotropic media a solution of the (in general coupled) vector wave equation is more complicated. In inhomogeneous media we end up with diffraction theories for the scalar domain (discussed in chapter 18) and the electromagnetic domain (discussed in chapter 27). Furthermore, the optical properties of optical media show a nonlinear behavior with the electromagnetic field amplitude or its intensity. However, such nonlinear optical effects will not be considered in this volume.

The wave propagation in homogeneous media is the most simple one, but it is extremely useful for the solution of practical problems. These can be frequently divided into an interaction section and a propagation section, linked by a surface. In a first step the field at the surface is obtained from the solution of a diffraction problem, and in a second step the wave propagation is used to compute the resulting field in an arbitrary region behind the interface. The wave propagation is itself particularly simple for fields that do not change their functional form during propagation. These elementary solutions of the wave equation will be discussed here, too. The most prominent one is the plane wave expansion. Finally, we will establish the link to geometrical optics and the eikonal equation.

17.2

From Maxwell to Helmholtz

17.2.1

Maxwell's Equations and the Inhomogeneous Wave Equation

Maxwell's equations describe the classical properties of electromagnetic waves [17-1]. In their differential form, they are given by Biot-Savart's law for the magnetic field strength \vec{H} and Faraday's induction law for the electric field strength \vec{E} :

$$\nabla \times \vec{H} = \frac{\partial}{\partial t} \vec{D} + \vec{j}, \quad (17-1a)$$

$$\nabla \times \vec{E} = -\frac{\partial}{\partial t} \vec{B}, \quad (17-1b)$$

with the displacement current \vec{D} and the magnetic induction \vec{B} given by

$$\nabla \cdot \vec{D} = \rho, \quad (17-1c)$$

$$\nabla \cdot \vec{B} = 0. \quad (17-1d)$$

The sources for the displacement current \vec{D} are given by electric charges described by the charge density ρ , while there are no ‘charges’ for the magnetic induction \vec{B} . With eq. (17-1a), the sources for a magnetic field are charged currents given by the current density \vec{j} and the displacement current \vec{D} . The current density obeys the equation of continuity

$$\nabla \cdot \vec{j} = -\frac{\partial \rho}{\partial t} \quad (17-2)$$

and, in lowest order, Ohm’s law

$$\vec{j} = \sigma \vec{E} \quad (17-3)$$

with the specific conductivity σ that is, in general, a tensor.

In linear, isotropic and homogeneous media, the properties of the supporting media are described by the following constitutive relations

$$\vec{D} = \epsilon_r \epsilon_0 \vec{E}, \quad (17-4a)$$

$$\vec{B} = \mu_r \mu_0 \vec{H} \quad (17-4b)$$

with the relative dielectric constant ϵ_r and the relative magnetic permeability μ_r . The dielectric constant of vacuum ϵ_0 (measurement), and the magnetic permeability of vacuum μ_0 (an SI definition) are given by

$$\begin{aligned} \epsilon_0 &= 8.85 \dots \cdot 10^{-10} \frac{\text{As}}{\text{Vm}}, \\ \mu_0 &= 4\pi \cdot 10^{-7} \frac{\text{Vs}}{\text{Am}}. \end{aligned} \quad (17-5)$$

With space-variant ϵ_r and μ_r , also \vec{D} , \vec{E} , \vec{B} and \vec{H} are usually functions of space and time.

Since after insertion of the constitutive relations (17-4) into Maxwell’s equations (17-1) the magnetic field strength is connected to the electric field strength by (17-1), further discussion is restricted to the electric field strength. Combining the time de-

derivative of Biot–Savart’s law (17-1a) with the rotation of Faraday’s induction law (17-1b) the following time-dependent wave equations for inhomogeneous, isotropic, charge-free ($\rho = 0$) and non-dispersive media is obtained:

$$\Delta \vec{E} = \varepsilon_r \varepsilon_0 \mu_r \mu_0 \frac{\partial^2}{\partial t^2} \vec{E} + \sigma \mu_r \mu_0 \frac{\partial}{\partial t} \vec{E} - \vec{\nabla} \left(\frac{\vec{\nabla} \varepsilon_r}{|\varepsilon_r|} \cdot \vec{E} \right). \quad (17-6)$$

Solutions of eq. (17-6) are in general damped electromagnetic waves. In non-magnetic isolators with $\sigma = 0$ and $\mu_r = 1$, eq. (17-6) can be written as

$$\Delta \vec{E} - \frac{n^2}{c^2} \frac{\partial^2}{\partial t^2} \vec{E} = -\vec{\nabla} \left(\frac{\vec{\nabla} \varepsilon_r}{|\varepsilon_r|} \cdot \vec{E} \right) = -2\vec{\nabla} \left(\frac{1}{n} \vec{E} \cdot \vec{\nabla} n \right) \quad (17-7)$$

with the velocity c of light in vacuum (SI constant)

$$c = \frac{1}{\sqrt{\varepsilon_0 \mu_0}} = 2.99792458 \cdot 10^8 \frac{\text{m}}{\text{s}} \quad (17-8)$$

and the refractive index $n(r)$ given by

$$n(\vec{r}) = \sqrt{\varepsilon_r \mu_r}. \quad (17-9)$$

In media of inhomogeneous refractive index, i.e., in gradient index media and at boundaries, the wave equation (17-7) is given by three coupled differential equations of second order. Generally, the inhomogeneous coupled wave equation (17-7) cannot be solved analytically. Analytical solutions exist only in homogeneous media (see section 17.3). At interfaces, boundary conditions can be formulated and the solutions in homogeneous regions separated by the interfaces can be determined analytically (see chapter 18).

In weak inhomogeneous media, the right-hand side of eq. (17-7) can be further simplified using [17-2]

$$\frac{\partial}{\partial x} \left(\frac{1}{n} \vec{E} \cdot \vec{\nabla} n \right) = \frac{\vec{\nabla} n}{n} \cdot \frac{\partial}{\partial x} \vec{E} + \vec{E} \cdot \vec{\nabla} \frac{\partial n}{\partial x} \approx \frac{\vec{\nabla} n}{n} \cdot \frac{\partial}{\partial x} \vec{E}. \quad (17-10)$$

After insertion of a harmonic wave for the electric field, one obtains from (17-7)

$$\Delta \vec{E} + n^2 k_0^2 \vec{E} = -2i \frac{\vec{k}}{n} \left(\vec{\nabla} n \cdot \vec{E} \right). \quad (17-11)$$

17.2.2

Wave Equation in Homogeneous Media and the Scalar Wave Equation

For the special case of homogeneous ($\vec{\nabla} \varepsilon_r = 0$) and non-conducting ($\sigma = 0$) media, the conventional form of the wave equation for the electrical field is obtained from eq. (17-6)

$$\Delta \vec{E} = \frac{n^2}{c^2} \frac{\partial^2}{\partial t^2} \vec{E}. \quad (17-12)$$

The wave equation in homogeneous media is still a vector equation, but the vector components of eq. (17-12) are no longer coupled and eq. (17-12) can be solved for each vector component separately. In homogeneous and isotropic media, the wave equations (17-12) for the electric and magnetic field are six independent differential equations and for each field vector component of \vec{E} or \vec{H} , respectively, the same equation holds. If polarization effects can be neglected, frequently only one vector component is considered, the discussion below is therefore restricted to a single vector field component E_x . Temporal and spatial dependency of eq. (17-12) are separated into different terms, consequently time and space dependence of the electrical field can be separated into

$$E_x(\vec{r}, t) = U(\vec{r}) \cdot F(t). \quad (17-13)$$

Fundamental solutions of the space and time-dependent parts are given by the harmonical dependencies of the electrical field. For the moment the discussion will be restricted to electrical fields which are constant in x and y (i.e., $\Delta_x E = \Delta_y E = 0$) in vacuum (i.e., $n = 1$), and the solutions of (17-13) can be written as

$$F(t) = \text{Re}\{e^{i\omega t}\} = \cos(\omega \cdot t), \quad (17-14a)$$

$$U(z) = \text{Re}\{e^{ik_0 z}\} = \cos(k_0 \cdot z) \quad (17-14b)$$

with the temporal frequency $\omega/2\pi = 1/T$ given by the inverse of the time period T and the spatial frequency $k_0/2\pi = 1/\lambda$ given by the inverse of the spatial period λ . Figure 17-1 illustrates both solutions for two times $t = 0$ and $t = t'$. The wave defined according (17-14) is propagating in the negative z -direction. Therefore different signs are usually applied for both arguments of the cosine functions in eq. (17-14).

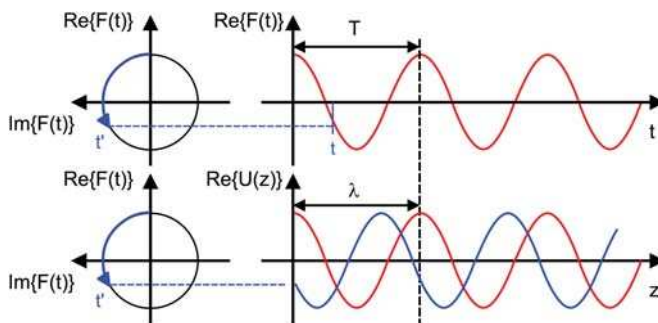


Figure 17-1: Harmonic time dependence of a harmonic wave.

The relation between temporal and spatial dependence can easily be obtained after insertion of eq. (17-14a) into both sides of the wave equation (17-12):

$$\frac{n^2}{c^2} \frac{\partial^2}{\partial t^2} U = -\frac{\omega^2}{c^2} U, \quad (17-15a)$$

$$\Delta U = -k_0^2 U. \quad (17-15b)$$

Since both equations (17-15a) and (17-15b) have to be identical, the spatial frequency $k_0/2\pi = 1/\lambda$ or the propagation constant k_0 can be identified by the ratio of frequency ω and phase velocity c :

$$k_0 = \frac{\omega}{c} = \frac{2\pi}{\lambda}. \quad (17-16)$$

After separation of the time-dependent part and consideration of a refractive index n , the usual form of the Helmholtz equation is obtained after insertion of the propagation constant k_0 :

$$\Delta U + n^2 k_0^2 U = \Delta U + k^2 U = 0. \quad (17-17)$$

Equations (17-17) are still three independent equations for the vector components of the electrical field. Only in the paraxial region may the discussion be reduced to a single vector component of the electrical field and the vector Helmholtz equations then reduce to a single scalar Helmholtz equation for the scalar field amplitude U . The scalar approximation is valid:

- for each of the vector components in homogeneous media or waves that are linearly polarized along the symmetry axis (separable problems);
- approximately, for small propagation angles (paraxial approximation);
- and scalar waves, for instance sound waves in gases.

17.2.3

The Dispersion Relation of the Harmonic Wave Solution

The property of the elementary solution to the Helmholtz equation can easily be discussed after insertion of the elementary solution into Maxwell's equations. According to eq. (17-14), the elementary solution may be given by a monochromatic plane wave:

$$U(\vec{r}, t) = e^{i(\vec{k}\vec{r} - \omega t)} \quad (17-18)$$

with the propagation vector $\vec{k} = (k_x, k_y, k_z)$. After insertion of the harmonic wave for all field components for \vec{E} , \vec{H} , \vec{D} and \vec{B} , from Maxwell's equations it follows that

$$\text{rot}\vec{E} = i(\vec{k} \times \vec{E}) = i\omega\mu_0\vec{H}, \quad (17-19a)$$

$$\text{rot}\vec{H} = i(\vec{k} \times \vec{H}) = -i\omega\vec{D}, \quad (17-19b)$$

or

$$\vec{k} \times \vec{E} = \omega \mu_0 \vec{H}, \quad (17-20a)$$

$$\vec{k} \times \vec{H} = -\omega \vec{D}. \quad (17-20b)$$

In charge-free media it can be derived that

$$\text{div } \vec{D} = i\vec{k} \cdot \vec{D} = 0, \quad (17-21)$$

$$\text{div } \vec{B} = i\mu_0 \vec{k} \cdot \vec{H} = 0. \quad (17-22)$$

As will be shown in section 17.4, the energy propagation is along the Poynting vector \vec{S} , defined by the vector product of \vec{E} and \vec{H} :

$$\vec{S} = \vec{E} \times \vec{H}. \quad (17-23)$$

From eqs (17-20) to (17-23) it follows that:

- \vec{D} , \vec{H} and \vec{k} are perpendicular to each other;
- \vec{S} , \vec{E} and \vec{H} are perpendicular to each other;
- since \vec{D} and \vec{E} are in general not parallel, \vec{k} and \vec{S} are generally not parallel either, i.e., the energy propagation in general does not follow the wave vector;
- in isotropic media \vec{E} and \vec{D} are parallel and \vec{S} points along the direction of \vec{k} .

The vector relations are illustrated in figure 17-2.

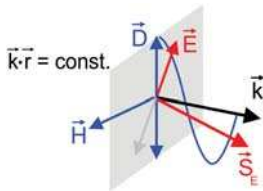


Figure 17-2: Relations between the wave vector \vec{k} , the Poynting vector \vec{S} and the field vectors \vec{H} , \vec{D} and \vec{E} .

Inserting (17-20a) into (17-20b), the displacement current is then given by the dispersion relation [17-3]

$$\vec{D} = -\frac{1}{\omega} \left(\vec{k} \times \left(\frac{1}{\omega \mu_0} \vec{k} \times \vec{E} \right) \right) = -\frac{1}{\omega^2 \mu_0} \left[(\vec{k} \cdot \vec{E}) \cdot \vec{k} - k^2 \vec{E} \right]. \quad (17-24)$$

In the most frequent case of isotropic media, the displacement current is parallel to the electric field, given by the lowest order of eq. (17-24), $\vec{D} = \epsilon_r \epsilon_0 \vec{E}$. This linear dependence of eq. (17-4) is a convenient simplification of the more complex influence of optical media on electromagnetic fields. It can be used in most applications, e.g., in isotropic media, where ϵ_r and μ_r are scalar quantities. Within the scope of the linear approximation (17-4) the displacement current is parallel to the electric field and energy propagation is along the wave propagation direction.

In general, however, the dielectric ‘constant’ may be a function of many parameters. In anisotropic media, dielectric permittivity ϵ_r and magnetic permeability μ_r are tensors and ϵ_r in general depends on the propagation direction \vec{k} (see chapter 26). Furthermore, any medium will be influenced by the incident electromagnetic field. Higher order dependencies of, e.g., the displacement current on the electrical field, are even not considered in eq. (17-24), i.e., ϵ_r and μ_r do not depend on the values of \vec{E} and \vec{H} at the point and time considered. Equations (17-4) or (17-24) may also include a differential time interval. For the time dependence this results in frequency dispersion and, for the spatial dependence, in spatial dispersion. In summary, the dielectric ‘constant’ is not constant at all and thus for the displacement current we may write

$$\vec{D} = \epsilon_0 \epsilon_r \left(\lambda; \vec{k}; \vec{E}; I; T; \vec{F} \dots \right) \cdot \vec{E}. \quad (17-25)$$

The refractive index of a medium may thus depend on:

- the wavelength (dispersion, important to be considered, e.g. in optical design);
- the direction of propagation or the field vector component (relevant in anisotropic media such as in the case of birefringence);
- the electric field amplitude (Pockels effect);
- the (electric) intensity (Kerr effect, thermal effects);
- many other effects, such as in the acousto-optical effect, the Faraday-effect...

The effects are present in almost any material, the question is only the magnitude of the effect. For the optical design of imaging systems it is, however, undesirable to have imaging properties depending on the intensity or similar effects. Therefore optical material with linear properties, such as conventional glasses and isotropic crystals, are preferred. Unfortunately not all effects can be avoided all the time, and therefore thermal effects, or, particularly at short wavelengths, anisotropic crystals, have to be applied. Beside the latter all nonlinear effects will not be considered in this volume.

In isotropic, linear media the displacement current is proportional to the electric field according eq. (17-4). After insertion of (17-4) and (17-21) on the left and right-hand sides of eq. (17-24), the dispersion relation for homogeneous electromagnetic waves follows, in analogy with eq. (17-16):

$$\epsilon_r \epsilon_0 = \frac{1}{\omega^2 \mu_0} k^2. \quad (17-26)$$

The dispersion relation of electromagnetic waves corresponds to the Ewald equation, stating that the magnitude of the wave vector in isotropic homogeneous media is given by $n \cdot 2\pi / \lambda$:

$$|k| = \sqrt{k_x^2 + k_y^2 + k_z^2} = n \cdot \sqrt{\epsilon_0 \mu_0} \cdot \omega = n \cdot \frac{\omega}{c} = nk_0 = n \frac{2\pi}{\lambda}. \quad (17-27)$$

For monochromatic light, any wavevector \vec{k} pointing from the origin along any arbitrary direction therefore has the same length and ends on a spherical surface, the so-called Ewald sphere (see sections 17-5 and 23-1).

17.3

Elementary Waves in Free Space

17.3.1

The Electromagnetic Plane Wave

Elementary solutions of eq. (17-17) in homogeneous, non-conducting, isotropic media of refractive index n are harmonic plane waves, described by a wave vector $\vec{k} = (k_x, k_y, k_z)$. Each vector component of the electric and magnetic field satisfy

$$\vec{E}(\vec{r}, t) = \vec{E} e^{i(\vec{k}\vec{r} - \omega t)}, \quad (17-28a)$$

$$\vec{H}(\vec{r}, t) = \vec{H} e^{i(\vec{k}\vec{r} - \omega t)}. \quad (17-28b)$$

As discussed above in the example given in figure 17-1, the sign of the time dependence, in combination with the sign of the exponent in eq. (17-28), determines the propagation direction. Because of this one has to be careful when implementing diffraction equations obtained from diffraction papers or textbooks: always check the sign of the time dependence and take the complex conjugate of the equations if the sign is different.

The reason why a plane wave has the mathematical form of eq. (17-28) becomes clear when we look at figure 17-3: For a plane wave the surfaces of constant phase are planes oriented perpendicular to the propagation direction determined by the wave vector \vec{k} . Therefore, the projection of the space vector \vec{r} (pointing to the equi-phase surfaces) onto the propagation vector \vec{k} must be constant and the phase planes are given by

$$\vec{k} \cdot \vec{r} = \text{const.} \quad (17-29)$$

which are just the Hessian normal form of a plane with normal vector \vec{k} .

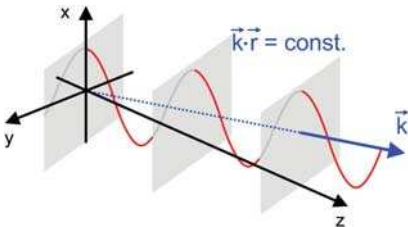


Figure 17-3: The plane wave.

As was shown in chapter 17.2.3, both the electric and the magnetic field vector are situated perpendicular to the propagation direction and perpendicular to each other (fig-

ure 17-4). Following eq. (17-20), the magnetic field vector is further perpendicular to the electrical field vector in isotropic media. The wave is transversal and \vec{E} , \vec{H} and \vec{k} form a right-handed coordinate system. For the considered homogeneous, isotropical, non-conducting medium, the electric and magnetic fields are in phase and their amplitudes are related by

$$|H| = \sqrt{\frac{\epsilon_r \epsilon_0}{\mu_r \mu_0}} |E| = n Z_0 |E|. \tag{17-30}$$

Z_0 is the free-space impedance

$$Z_0 = \sqrt{\frac{\epsilon_0}{\mu_0}} = 377 \Omega. \tag{17-31}$$

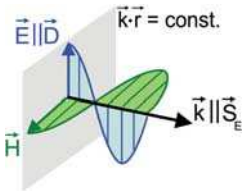
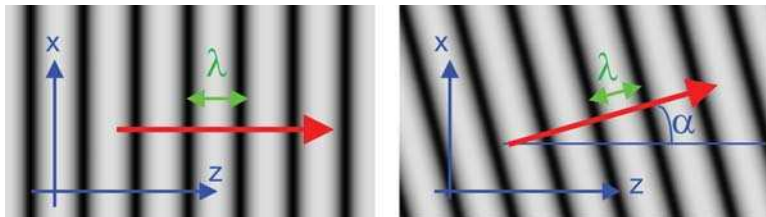


Figure 17-4: \vec{E} , \vec{H} and \vec{k} of a linearly polarized, plane electromagnetic wave in an isotropic, non-conducting medium.

It should be noted that the complex notation is preferable for the linear system theory description, e.g., in Fourier optics. Linear system theory becomes simplified by the application of multiplicative operators. Multiplicative operators are considerably less complicated in complex notation. In particular, a phase effect can simply be considered by a multiplication in complex notation. However, it should not be forgotten that only the real part is physically real. After separation of the time-dependent part, one obtains for the stationary solution of the Helmholtz equation for each vector component:

$$U(x, y, z) = A \cdot \cos(k_x x + k_y y + k_z z) = A \cdot \cos(\vec{k} \cdot \vec{r}). \tag{17-32}$$

Figure 17-5 illustrates two examples of plane waves in the x - z -plane (with $k_y=0$).



$$U(x, z) \sim \cos\left(n \frac{2\pi}{\lambda} z\right)$$

$$U(x, z) \sim \cos\left[n \frac{2\pi}{\lambda} (x \sin \alpha + z \cos \alpha)\right]$$

Figure 17-5: Plane waves and wave vectors in the x - z plane as stationary solutions of the Helmholtz equation.

17.3.2

Spherical Wave

In polar coordinates, the scalar wave equation can be written with the Laplace operator as

$$\Delta U = \frac{1}{r} \frac{d^2}{dr^2} (rU) \quad (17-33)$$

and the wave equation in polar coordinates is then obtained as

$$\frac{d^2(rU)}{dr^2} + k^2 rU = 0. \quad (17-34)$$

The elementary solutions of eq. (17-34) are given by spherical waves. The spherical wave has spherical equi-phase surfaces. It is scalar, obeying the formula

$$U(r, t) = \frac{e^{i(kr - \omega t)}}{r}. \quad (17-35)$$

With the radius $r = \sqrt{x^2 + y^2 + z^2}$ and the propagation constant $k = |\vec{k}|$. Figure 17-6 illustrates a spherical wave.

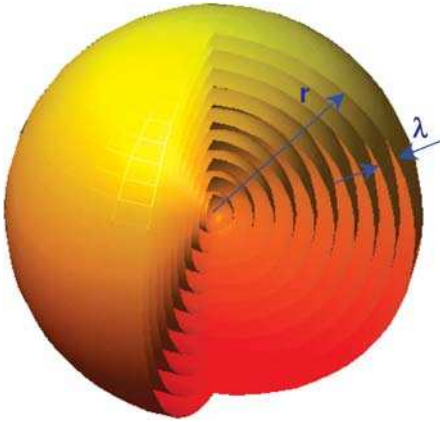


Figure 17-6: Illustration of a spherical wave.

17.3.3

Dipole Wave

A dipole wave is emitted by a dipole with an oscillating dipole moment \vec{p} . The radiated electric field is

$$\vec{E} = \left[k^2 \frac{(\vec{r} \times \vec{p}) \times \vec{r}}{r^2} + \left(\frac{3\vec{r}(\vec{r} \cdot \vec{p})}{r^2} - \vec{p} \right) \left(\frac{1}{r^2} - \frac{ik}{r} \right) \right] \frac{e^{ikr}}{r}. \quad (17-36)$$

The vector potential \vec{A} of a single radiating dipole in the origin, oscillating harmonically with frequency ω is given by

$$\vec{A}(r) = \frac{\mu_0 \vec{j}}{4\pi} \frac{e^{ikr}}{r} = -\frac{i\omega\mu_0}{4\pi} \frac{e^{ikr}}{r} \vec{p}. \quad (17-37)$$

Resulting in an electromagnetic field of

$$\begin{aligned} \vec{H}(\vec{r}) &= \frac{1}{\mu_0} \nabla \times \vec{A}(r) = -\frac{i\omega}{4\pi} \nabla \times \frac{e^{ikr}}{r} \vec{p}, \\ \vec{E}(r) &= \frac{i}{\omega\epsilon_0} \nabla \times \vec{H}(r) = \frac{1}{4\pi\epsilon_0} \nabla \times \left(\nabla \times \frac{e^{ikr}}{r} \vec{p} \right). \end{aligned} \quad (17-38)$$

A lengthy but straightforward calculation yields the dipole field in spherical coordinates (see figure 17-7 for coordinates).

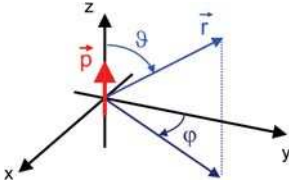


Figure 17-7 Spherical coordinates for dipole field.

$$\begin{aligned} E_r &= \frac{2p}{4\pi\epsilon_0} k_0^3 e^{ik_0 r} \left(\frac{1}{(k_0 r)^3} - \frac{i}{(k_0 r)^2} \right) \cos \vartheta, \\ E_\vartheta &= \frac{p}{4\pi\epsilon_0} k_0^3 e^{ik_0 r} \left(\frac{1}{(k_0 r)^3} - \frac{i}{(k_0 r)^2} - \frac{1}{k_0 r} \right) \sin \vartheta, \\ H_\varphi &= -\frac{icp}{4\pi} k_0^3 e^{ik_0 r} \left(\frac{1}{(k_0 r)^2} - \frac{i}{k_0 r} \right) \sin \vartheta, \\ E_\varphi &= H_r = H_\vartheta = 0. \end{aligned} \quad (17-39)$$

The radiation field separates into three zones:

- A near field, with an amplitude proportional to $1/r^3$.
- An intermediate field, with an amplitude proportional to $1/r^2$.
- A radiation field, with an amplitude proportional to $1/r^1$.

The far field intensity becomes

$$I \sim \left| E_\vartheta^{\text{far}} \right|^2 = \frac{p^2 k_0^4}{16\pi^2 \epsilon_0^2} \frac{\sin^2 \vartheta}{r^2}. \quad (17-40)$$

yielding doughnut-shaped radiation characteristics as depicted in figure 17-8.

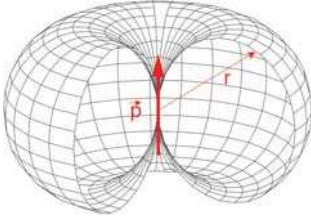


Figure 17-8: Far field radiation characteristic of a dipole.

17.3.4

Radiated Field of a Harmonic Current Distribution

Solutions of the Maxwell equations for time-dependent charge and current distributions are the retarded potentials [17-4]. For a harmonic current distribution and vanishing charge distribution

$$\begin{aligned}\vec{j}(\vec{r}, t) &= \vec{j}(\vec{r})e^{-i\omega t}, \\ \rho(\vec{r}, t) &= 0.\end{aligned}\quad (17-41)$$

The electromagnetic potentials become

$$\begin{aligned}\Phi(\vec{r}, t) &= 0, \\ \vec{A}(\vec{r}, t) &= \frac{\mu_0}{4\pi} \iiint_V \vec{j}(\vec{r}') \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|} d^3\vec{r}' e^{-i\omega t}.\end{aligned}\quad (17-42)$$

Computing the electric field from the electromagnetic potentials from

$$\vec{E} = -\frac{\partial \vec{A}}{\partial t} - \nabla \Phi \quad (17-43)$$

gives the result, for $\Phi = 0$ and harmonic currents, as

$$\vec{E}(\vec{r}) = \frac{i\omega\mu_0}{4\pi} \iiint_V \vec{j}(\vec{r}') \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|} d^3\vec{r}'. \quad (17-44)$$

17.3.5

A Note on Plane and Spherical Waves

As will be discussed in more detail in chapter 18, non-vanishing solutions of the wave equation need real light sources, which are not at infinity. Real light sources may further only emit a limited amount of energy. These necessary conditions for real solutions of the wave equation are summarized in the Sommerfeld conditions:

$$\begin{aligned} \lim_{r \rightarrow \infty} r \left(\frac{\partial U}{\partial r} + ikU \right) &= 0, \\ \lim_{r \rightarrow \infty} U &= 0. \end{aligned} \quad (17-45)$$

A consequence of the Sommerfeld condition is the uniqueness law: two solutions of the wave equation, which obey the Sommerfeld condition and have identical singularities, such as sources and sinks, are necessarily identical. According to the Sommerfeld condition, an everywhere regular solution of the wave equation is only given by $U = 0$.

Elementary solutions of the wave equation are thus only emitting spherical waves $\exp(ikr)/r$, but, for example, not $\sin(kr)/r$. In particular, plane waves do not obey the Sommerfeld condition. Nevertheless, it is allowed and often advantageous to expand the solution in a series expansion of plane waves with different propagation angles (see section 17.5).

17.4

Energy, Irradiance and Intensity

Adding Maxwell's equations (17-1) after scalar multiplication with \vec{E} or \vec{H} , and using the vector identity

$$\vec{H} \cdot (\vec{\nabla} \times \vec{E}) - \vec{E} \cdot (\vec{\nabla} \times \vec{H}) = \vec{\nabla} \cdot (\vec{E} \times \vec{H}) \quad (17-46)$$

we obtain

$$\vec{\nabla} \cdot (\vec{E} \times \vec{H}) = -\vec{j} \cdot \vec{E} - \left(\vec{E} \cdot \frac{\partial}{\partial t} \vec{D} + \vec{H} \cdot \frac{\partial}{\partial t} \vec{B} \right). \quad (17-47)$$

The last term on the right-hand side can be interpreted as the time derivative of the energy density ρ , to which the electric and magnetic fields contribute:

$$\rho = \rho_E + \rho_H, \quad (17-48)$$

The energy densities of the electric and magnetic field are consequently given by

$$\rho_E = \frac{1}{2} \vec{E} \cdot \vec{D} = \frac{1}{2} \epsilon_0 \epsilon_r \vec{E}^2, \quad \rho_H = \frac{1}{2} \vec{H} \cdot \vec{B} = \frac{1}{2} \mu_0 \mu_r \vec{H}^2, \quad (17-49)$$

in units of $\text{J/m}^3 = \text{Ws/m}^3$. After introduction of the Poynting vector \vec{S} with

$$\vec{S} = \vec{E} \times \vec{H}, \quad (17-50)$$

the continuity equation for electromagnetic fields is finally obtained:

$$\vec{\nabla} \cdot \vec{S} = -\vec{j} \cdot \vec{E} - \frac{\partial}{\partial t} \rho . \quad (17-51)$$

Integration over an arbitrary volume V with surface Σ and the application of the Gauss theorem yields

$$\oint_{\Sigma} \vec{S} \cdot d\vec{a} = - \int_V \vec{j} \cdot \vec{E} dV - \frac{d}{dt} \int_V \rho dV = - \int_V \vec{j} \cdot \vec{E} dV - \frac{d}{dt} W . \quad (17-52)$$

The first term on the right-hand side represents the total dissipated power within the volume V , with electric currents as sources. The second term represents the change over time of the complete electromagnetic energy W inside a volume V . The left-hand side of eq. (17-52) thus represents the total power flow through the boundary Σ of volume V . Since the time frequency of electromagnetic light waves is usually too high to be resolved by detectors, the power of the electromagnetic wave propagating through a surface Σ is given by the average rate of the energy flow [17-5]

$$\Phi = \left\langle \int_{\Sigma} \vec{S} \cdot d\vec{a} \right\rangle . \quad (17-53)$$

The power Φ is also called radiant flux. The radiant flux in physical units of Watts [W] is frequently given in physiological units – as luminous flux – in lumen [lm]. In source-free media ($\vec{j} = 0$), the right-hand side of eq. (17-52) vanishes. Thus the radiant flux through a closed surface and any closed surface integral over the time-averaged Poynting vector must vanish. The Poynting vector \vec{S} introduced above, thus represents the flux density of energy per second and unit area. According to the definitions of radiometry, the time average of the Poynting vector therefore corresponds to the irradiance or flux density I of a surface element perpendicular to the Poynting vector:

$$I = \left| \frac{d\Phi}{d\vec{a}} \right| = \left\langle \left| \vec{S}(r, t) \right| \right\rangle . \quad (17-54)$$

The irradiance I is given in physical units W/cm^2 or, in physiological units as illuminance in lux [lx] or footcandle [ft-c]. After insertion the continuity equation for source-free media is obtained:

$$\int_{\Sigma} \left\langle \vec{S} \right\rangle \cdot d\vec{a} = \int_{\Sigma} I \vec{s} \cdot d\vec{a} = \frac{1}{n} \int_{\Sigma} I \vec{p} \cdot d\vec{a} = 0 . \quad (17-55)$$

The right-hand side term in eq. (17-55) has a formal analogy to the continuity equation of mechanics with the direction cosine vector \vec{p} interpreted as velocity and the irradiance I as flux density.

For a plane wave in an isotropic medium the propagation direction of the wave \vec{k} and the direction of the energy flow \vec{S} are parallel to each other and perpendicular to

the fields \vec{E} and \vec{H} (figure 17-2 and 17-9). The Poynting vector, after insertion of eq. (17-20) into eq. (17-50), becomes

$$\vec{S} = \vec{E} \times \vec{H} = \frac{1}{\omega\mu_r\mu_0} |\vec{E}|^2 \vec{k} = \sqrt{\frac{\epsilon_0\epsilon_r}{\mu_0\mu_r}} |\vec{E}|^2 \vec{s} = \frac{c}{n} \cdot \rho \cdot \vec{s}. \quad (17-56)$$

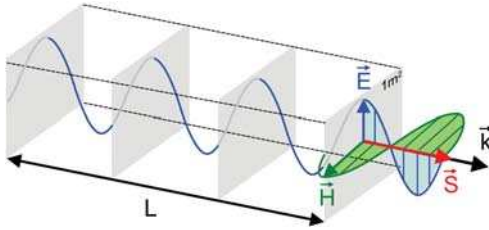


Figure 17-9: The energy in the sketched pile flows per second through 1 m^2 .

In Figure 17-9 the energy flow of a plane wave through a unit surface of 1 m^2 perpendicular to its propagation vector \vec{k} is illustrated. For a plane wave the energy flow per second through the surface is given by the total electromagnetic power W in a volume of 1 m^2 cross-section and length $L = c \cdot 1 \text{ s}$. In general, the volume can be considered as a light tube. Through the sidewalls, parallel to the wavefront normal \vec{s} , the power flow vanishes.

Considering that the mean value of the \cos^2 function is 0.5, we obtain the irradiance on a surface perpendicular to the wave vector for non-magnetic media:

$$I = \frac{\Phi}{1 \text{ m}^2} = \langle |\vec{S}(r, t)| \rangle = \frac{1}{2} \sqrt{\frac{\epsilon_0\epsilon_r}{\mu_0\mu_r}} |\vec{E}_0|^2 = \frac{1}{2} n Z_0 E_0^2 \quad (17-57)$$

with the free space impedance Z_0 (eq. (17-31)). As a consequence of (17-57), the irradiance of an electromagnetic wave is not proportional to the squared electric field alone. It is proportional to the refractive index multiplied by this value. Accordingly we get the irradiance of a scalar wave as

$$I(r) = \frac{1}{2} n |U(r)|^2. \quad (17-58)$$

A note of caution: the irradiance, proportional to the time average of the squared amplitude of the electrical field, has traditionally been called the intensity – as it will also be called in this volume. Care has to be taken not to confuse it with the radiant intensity, which is given by the power flux per solid angle $d\Phi/d\Omega$ (see volume 1, chapter 6). Furthermore, it should be noted that the energy flow density is not usually perpendicular to the surface under consideration. Obliquity factors, according to the projection of the direction cosine on the surface normal (eq. 17-55), have to be considered in order to compute the flow of energy through the surface.

17.5 The Angular Spectrum

17.5.1 Spatial Frequency Representation

Plane waves are the elementary solutions of the wave equation:

$$U(x, y, z) = A(x, y, z) \cdot e^{\pm i(k_x x + k_y y + k_z z)}. \quad (17-59)$$

For propagating solutions, the length of the wave vectors is given by the Ewald equation:

$$|k| = \sqrt{k_x^2 + k_y^2 + k_z^2} = n \cdot \frac{2\pi}{\lambda}. \quad (17-60)$$

Since both the Maxwell and the wave equation are linear, every propagating solution of the wave equation can be represented by a superposition of elementary solutions. Any monochromatic electromagnetic field component can therefore be represented as a superposition of plane waves. The complex amplitude of the ‘angular spectrum’ or ‘plane wave spectrum’ is obtained from a Fourier transform of the field. For this representation, however, a representation in spatial frequencies is more convenient. The spatial frequency vector is proportional to the wave vector. It is defined by

$$2\pi\vec{v} = \vec{k}. \quad (17-61)$$

The spatial frequency spectrum representation of a stationary field distribution can thus be written as a three-dimensional Fourier integral:

$$U(x, y, z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u(v_x, v_y, v_z) e^{2\pi i(xv_x + yv_y + zv_z)} dv_x dv_y dv_z. \quad (17-62)$$

A plane wave can now be described in the spatial domain by its spatial periodical field distribution with

$$U(x, y, z) = A \cdot e^{i2\pi(v_x x + v_y y + v_z z)}. \quad (17-63)$$

Considering an instantaneous shot of the wave field at the time $t = t_0$, wave peaks and troughs can be observed (figure 17-10).

After application of the differentiation rule of Fourier theory

$$F\{\Delta U(x)\} = -(2\pi v)^2 u(v) \quad (17-64)$$

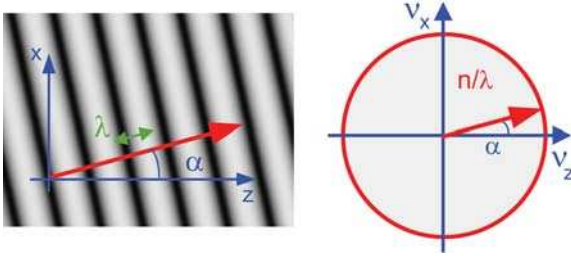


Figure 17-10: Plane wave in space and frequency domain.

the Ewald equation (17-60) can now be identified as the Helmholtz equation in the frequency domain. Since in monochromatic fields the length of the frequency vector is limited by the Ewald equation (17-60), the frequency spectrum is limited to a sphere in the frequency domain, the so-called Ewald sphere (figure 17-10, right-hand side). Due to the fact that the frequency vector is of length $1/\lambda$, any frequency-vector component can be expressed by the two others. In particular, the z component of the spatial frequency vector can be represented by

$$v_z = \pm \sqrt{\left(\frac{n}{\lambda}\right)^2 - v_x^2 - v_y^2}. \quad (17-65)$$

It should be noted that as well as propagating plane waves, in general, solutions of the wave equation are also given by evanescent waves (from the latin *evanescere*, to diminish), e.g.,

$$U(x, y, z) = A \cdot e^{i(k_x x + k_y y)} \cdot e^{-\chi z}. \quad (17-66)$$

Evanescent waves will be described in more detail in the next paragraph. Here it is to be noted that evanescent waves are given by frequencies not fulfilling the Ewald equation, i.e., imaginary solutions of the v_z -components according eq. (17-65):

$$\chi = \sqrt{k_x^2 + k_y^2 - \left(\frac{2\pi \cdot n}{\lambda}\right)^2}. \quad (17-67)$$

Evanescent fields are thus damped in the z direction and the range of evanescent fields is limited to small distances. Since only the propagating frequencies, according to eq. (17-65), will be present in free space far apart from diffracting apertures and sources, then eq. (17-65) allows for an interpretation of the Ewald sphere as the transfer function of free space. For those who are used to visual interpretations of physical phenomena, the Ewald sphere thus offers a powerful tool for the illustration of optical phenomena. Several examples will be given throughout this volume.

The wave vector \vec{k} and the spatial frequency vector \vec{v} differ only by a factor of 2π and are therefore used synonymously. However, the Fourier integral for a representation with wave vectors \vec{k} will differ from eq. (17-62) therefore by a factor of $(1/2\pi)^3$.

The wave vector \vec{k} of wave optics represents the normal vector, perpendicular to the wavefronts of constant phase φ with $\varphi = \vec{k} \cdot \vec{r} = 2\pi\vec{v} \cdot \vec{r}$. The wave vector is therefore parallel to the tangent vector \vec{s} at a light ray and is thus parallel to the optical direction cosine vector $\vec{p} = (p, q, m)$. The frequency vector is proportional to the direction cosine, too:

$$\lambda\vec{v} = \vec{p} = n\vec{s} = \vec{\nabla}W \quad (17-68)$$

with refractive index n , the tangential vector to the light ray \vec{s} and the wavefront W . Equation (17-68) is equivalent to the well known Eikonal equation. The transition to geometrical optics is discussed in more detail in section 17.7.1.

17.5.2

Transformation of the Three-dimensional Spectrum into Two Dimensions

According eq. (17-65), the frequency spectrum is restricted to the surface of a sphere and the volume integration (17-62) can be reduced to a surface integral. To perform the correct transition to the two-dimensional spectrum representation, eq. (17-62) is written in spherical coordinates. With

$$v_x = \rho \cos \vartheta \sin \alpha, \quad (17-69a)$$

$$v_y = \rho \sin \vartheta \sin \alpha, \quad (17-69b)$$

$$v_z = \rho \cos \alpha, \quad (17-69c)$$

the Jacobi determinant is written as $J = \rho^2 \sin \alpha$ and is

$$U(x, y, z) = \int_0^{2\pi} \int_0^{2\pi} \int_0^{\infty} u_{3D}(\rho, \vartheta, \alpha) e^{2\pi i \rho [(x \cos \vartheta + y \sin \vartheta) \sin \alpha + z \cos \alpha]} \rho^2 \sin \alpha \, d\rho d\vartheta d\alpha. \quad (17-70)$$

Applying the Ewald equation $\rho = n/\lambda$ in spherical coordinates, and after integration over ρ we obtain:

$$\begin{aligned} U(x, y, z) &= \int_0^{2\pi} \int_0^{2\pi} \int_0^{\infty} u_{3D}(\rho, \vartheta, \alpha) e^{2\pi i \rho [(x \cos \vartheta + y \sin \vartheta) \sin \alpha + z \cos \alpha]} \delta\left(\rho - \frac{n}{\lambda}\right) \rho^2 \sin \alpha \, d\rho d\vartheta d\alpha \\ &= \left(\frac{n}{\lambda}\right)^2 \int_0^{2\pi} \int_0^{2\pi} u_{3D}\left(\frac{n}{\lambda}, \vartheta, \alpha\right) e^{2\pi i \frac{n}{\lambda} [(x \cos \vartheta + y \sin \vartheta) \sin \alpha + z \cos \alpha]} \sin \alpha \, d\vartheta d\alpha. \end{aligned} \quad (17-71)$$

For convenience and ease of computation, a spectrum representation in cartesian coordinates is preferable. With $dv_x dv_y = \cos \alpha \sin \alpha \, d\alpha \, d\vartheta$ it follows that

$$U(x, y, z_1) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u_{2D}(v_x, v_y; z_0) e^{2\pi i \left(x v_x + y v_y \pm \delta z \sqrt{\left(\frac{n}{\lambda}\right)^2 - v_x^2 - v_y^2} \right)} dv_x dv_y \quad (17-72)$$

with

$$u_{3D} \left(\nu_x, \nu_y, \sqrt{\left(\frac{n}{\lambda}\right)^2 - \nu_x^2 - \nu_y^2} \right) = u_{2D}(\nu_x, \nu_y) \cdot \cos \alpha. \tag{17-73}$$

The 2D-spectrum $u_{2D}(\nu_x, \nu_y; z_0)$ of a monochromatic field $U(x, y, z_0)$ in the plane z_1 is thus given by a phase modulation from the spectrum in the plane z_0 . The phase corresponds to the shift phase according to the shift theorem.

For transformation of the transverse spectrum $u_{2D}(\nu_x, \nu_y)$ to three dimensions a cosine factor has to be considered. A homogeneous illumination $u_{2D}(\nu_x, \nu_y) = \text{constant}$ of the transverse pupil by a Lambertian source leads to a dampened amplitude $u_{3D}(\nu_x, \nu_y, \nu_z) \sim \cos \alpha$, while a homogeneous amplitude on the Ewald sphere, corresponding to, e.g., a spherical wave, leads to an amplitude distribution $u_{2D} \sim 1/\cos \alpha$ increasing with the aperture angle α .

17.5.3

Free-space Propagation of Transverse Fields

Equation (17-72) is obtained by the three-dimensional spectrum representation after integration over the delta-function according (17-65) with consideration of proper integration constants [17-1]. With

$$U(x, y, z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{u_{3D}(\nu_x, \nu_y, \nu_z)}{\sqrt{\left(\frac{n}{\lambda}\right)^2 - \nu_x^2 - \nu_y^2}} e^{2\pi i(x\nu_x + y\nu_y + z\nu_z)} \delta\left(\nu_z \pm \sqrt{\left(\frac{n}{\lambda}\right)^2 - \nu_x^2 - \nu_y^2}\right) d\nu_x d\nu_y d\nu_z \tag{17-74}$$

the transverse frequency spectrum $u_{2D}(\nu_x, \nu_y)$ is given by projection of the 3D spectrum onto the transverse frequency plane perpendicular to the ν_z -axis. Figure 17-11 illustrates the change from 3D- to 2D spectrum representation.

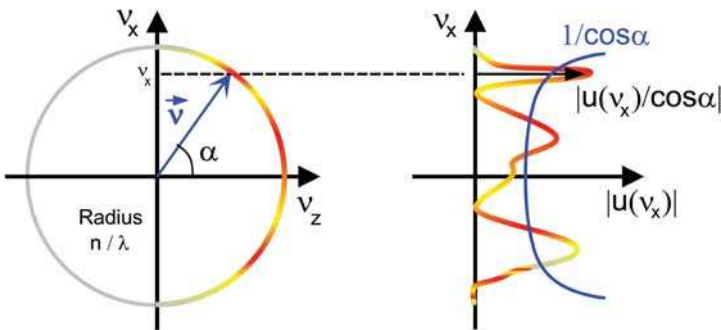


Figure 17-11: “Free-space propagation” of a plane wave.

In some textbooks eq. (17-72) is described as the free space propagation equation for electromagnetic fields [17-6], [17-7]. However, the ‘propagation’ of a plane wave

is given by the time dependency $e^{-i\omega t}$. Nevertheless, it is often advantageous to consider only transverse fields $U(x, y)$ with its spectra $u(\nu_x, \nu_y)$. In a stationary description the phase term in eq. (17-72) then describes the phase relation between different points spaced apart in an infinitely extended field. As illustrated in figure 17-12, the phase difference between two points with axial distance δz is given by

$$e^{i\varphi} = e^{\frac{i2\pi}{\lambda}\delta z \cdot n \cdot \cos \alpha} = e^{i2\pi\delta z \cdot \nu_z} = e^{i2\pi\delta z \cdot \sqrt{\left(\frac{n}{\lambda}\right)^2 - \nu_x^2 - \nu_y^2}} = h(\nu_x, \nu_y; \delta z) . \quad (17-75)$$

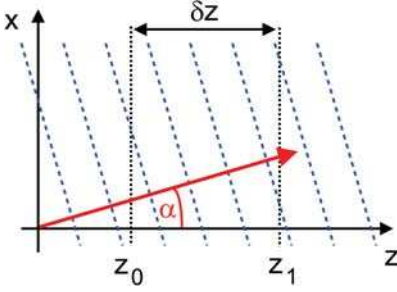


Figure 17-12: "Free-space propagation" of a plane wave.

The frequency spectrum in a plane $z_1 = z_0 + \delta z$ is thus given by the frequency spectrum in plane z_0 multiplied by the phase function (17-75):

$$u(\nu_x, \nu_y, z_1) = u(\nu_x, \nu_y, z_0) e^{2\pi i \delta z \nu_z} = h(\nu_x, \nu_y; \delta z) \cdot u(\nu_x, \nu_y, z_0) . \quad (17-76)$$

Multiplication in the frequency domain corresponds to a convolution in the space domain, thus the Fourier transform of the phase function $h(\nu_x, \nu_y; z)$ can be interpreted as the free space transmission function:

$$U(x, y, z_1) = H(x, y; \delta z) \otimes U(x, y, z_0) \quad (17-77)$$

with

$$H(x, y; \delta z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(\nu_x, \nu_y; \delta z) e^{2\pi i(x\nu_x + y\nu_y)} d\nu_x d\nu_y . \quad (17-78)$$

For small propagation angles, the free-space transfer function may be expanded by expansion of the hyperbolic phase into

$$\sqrt{1 - \lambda^2(\nu_x^2 + \nu_y^2)} = 1 - \frac{\lambda^2}{2}(\nu_x^2 + \nu_y^2) - \dots \quad (17-79)$$

and the parabolic or Fresnel approximation of the transfer function can be given analytically as:

$$H_p(x, y; z) = e^{\frac{i2\pi}{\lambda}z} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-i\pi\lambda z(\nu_x^2 + \nu_y^2)} e^{2\pi i(x\nu_x + y\nu_y)} d\nu_x d\nu_y = \frac{1}{i\lambda z} e^{\frac{i2\pi}{\lambda}z} e^{i\frac{\pi}{\lambda z}(x^2 + y^2)} . \quad (17-80)$$

After insertion into eq. (17-77), the Fresnel integral is obtained:

$$U_F(x, y; z) = \frac{1}{i\lambda z} e^{\frac{2\pi i}{\lambda} \left[z + \frac{1}{2z}(x^2 + y^2) \right]} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} U(x_0, y_0; 0) e^{\frac{\pi i}{\lambda z}(x_0^2 + y_0^2)} e^{-\frac{2\pi i}{\lambda z}(xx_0 + yy_0)} dx_0 dy_0. \quad (17-81)$$

Due to the increasing power of modern computers, the Fresnel approximation is of less importance. For large distances, when $z \gg D^2/\lambda$ where D is the lateral extension of the field under consideration, the first phase term of eq. (17-81) can be neglected and the far-field or Fraunhofer approximation is obtained:

$$u_F(x, y; z) = C \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u(x_0, y_0; 0) e^{-2\pi i \left(\frac{x}{\lambda z} x_0 + \frac{y}{\lambda z} y_0 \right)} dx_0 dy_0. \quad (17-82)$$

According to the Fraunhofer approximation, at large distances the field distribution is given by the transverse spectrum of the field distribution. Further details on approximate solutions to the wave equation and to diffraction equations will be given in section 17.7 and chapter 18.

17.5.4

Periodic Fields with Discrete Spectra

A transversally periodic wave field, as obtained, e.g., behind a periodic diffraction grating, is also periodic in z . This effect is called the Talbot effect after its discoverer [17-8]. A periodic field amplitude with period d is described by a discrete spatial frequency spectrum:

$$u(x, z = 0) = \sum_m a_m \cdot e^{i2\pi \frac{m}{d} x}. \quad (17-83)$$

A discrete spectrum in ν_x is also discrete in ν_z . For small frequencies the longitudinal frequency gap $\Delta\nu_z$ between $\nu_x = 0$ and the transverse frequency $\nu_x = m/d$ can be approximated by

$$\Delta\nu_z = \frac{n}{\lambda} - \sqrt{\left(\frac{n}{\lambda}\right)^2 - \left(\frac{m}{d}\right)^2} \approx \frac{1}{2} \frac{\lambda}{n} \left(\frac{m}{d}\right)^2. \quad (17-84)$$

With $m = 1$, the longitudinal or Talbot period z_T of the wave amplitude is thus given by

$$z_T = 2 \frac{d^2}{\lambda}. \quad (17-85)$$

However, as illustrated in figure 17-13, the longitudinal periodicity of a transversally periodic wave field is superposed by higher orders of the discrete spectrum, which give rise to different Talbot periods. The spatial frequency spectrum of a wave field strictly periodic in z , must therefore be discrete and equidistant in ν_z [17-9].

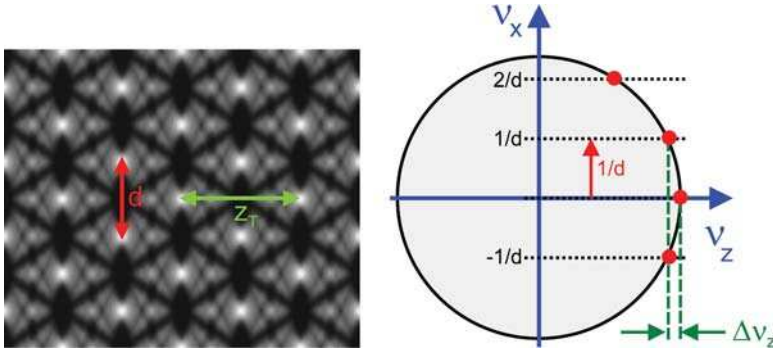


Figure 17-13: Talbot effect: a transversally periodic field is also longitudinally periodic due to discrete spectrum in ν_x and ν_z .

17.5.5

Boundary Conditions and the Spatial Frequency Spectrum

The spatial frequency representation is frequently applied to the solution of diffraction problems. The simplest case is given by a planar interface of two media with different refractive index. Since in general the electric field and its derivative have to be continuous everywhere in space, the transverse spatial frequencies of the fields on both sides of an interface must also be identical. The spectrum representations allow for a visual interpretation of the boundary conditions for electromagnetic fields at interfaces.

However, the concept is far more general. Frequently the elementary solutions of the wave equation of inhomogeneous or structured media can be calculated analytically or numerically, and are given by a series expansion. Those elementary solutions are called the modes. In homogeneous media, the plane waves represent the modes and there is a continuous spectrum of modes. In wave-guides, for example, there are discrete modes. In periodic media, like diffraction gratings, solutions of the wave equation are given by more complex modes. The solution of the diffraction problem is often reduced to the problem of fitting the mode spectra of the fields in the different regions. In the following, the simplest case at a planar interface will be considered. For convenience, the interface normal vector \vec{N} is chosen parallel to the z -direction. The transverse frequency spectrum has to be identical on both sides of the interface:

$$\nu_x = p_x/\lambda = n \sin \alpha/\lambda = \text{const.} \quad (17-86)$$

However, the radius of the Ewald sphere depends on the refractive index and thus the propagation direction of the plane waves on both sides of the interface differ in accordance with the difference of the refractive indices (figure 17-14, right-hand side). The change of the spatial frequency vector is in the direction of the surface normal and the transition equation for spatial frequencies is obtained [17-10] as:

$$\vec{v}_1 - \vec{v}_0 = \gamma \cdot \vec{N} \tag{17-87a}$$

with

$$\gamma = \vec{N} \cdot \vec{v}_1 - \vec{N} \cdot \vec{v}_0 . \tag{17-87b}$$

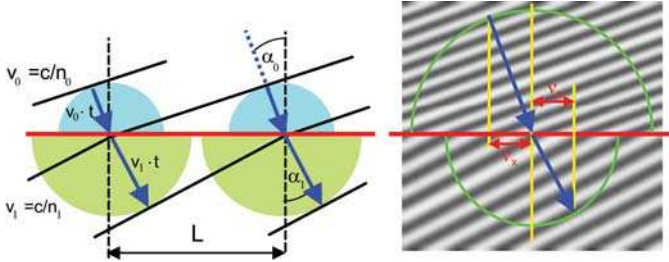


Figure 17-14: Frequency spectrum representation of boundary conditions at an interface and Huygens construction in space domain.

Equations (17-86) and (17-87) turn out to be equivalent to Snell’s law of refraction, which is frequently derived from Fermat’s principle for the minimum optical path length. The derivation is illustrated on the left-hand side of figure 17-14. According to the Huygens principle, two spherical waves are emitted at two points at the interface separated by a distance L with a phase difference given by $\varphi = v_0 \cdot t = c \cdot t / n_0$. In the medium with refractive index n_1 the Huygens elementary spherical waves propagate with phase velocity $v_1 = c / n_1$ and thus the following condition must be satisfied:

$$L = \frac{v_0 \cdot t}{\sin \alpha_0} = \frac{v_1 \cdot t}{\sin \alpha_1} . \tag{17-88}$$

With $t = T = 2\pi / \omega$ Snell’s law follows directly from eq. (17-88).

While the radius of the Ewald sphere is proportional to the refractive index n , the radius of the Huygens elementary spherical wave is proportional to $1/n$ (figure 17-14, left-hand side). The Ewald sphere in the frequency domain and the Huygens spherical wave in the space domain thus scale inversely with the refractive index. This corresponds to the principle of reciprocity.

17.5.6

Vector Field Representation by Spatial Frequencies

So far only the spatial frequency spectrum of scalar fields has been considered. Here the vector effects only are outlined, a detailed treatment of the vector effects is given in chapter 27.

The spectrum representation in three dimensions illustrates that, for each transverse spatial frequency, the propagation direction is given by the Ewald equation.

Since the electric field vector is transverse to the propagation vector, vector effects cannot be neglected when considering the vector characteristic of the electromagnetic field. In general it has to be considered, that

- the vector components of the spatial frequency spectrum are coupled;
- considering only one vector component, not all frequencies on the Ewald sphere are allowed solutions.

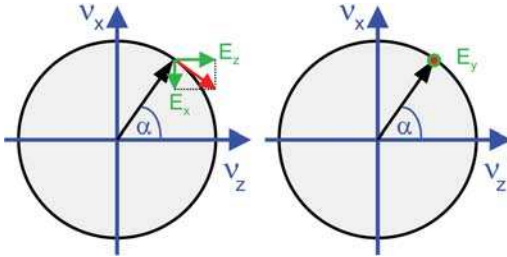


Figure 17-15: Vector representation of parallel (p) and perpendicular (s) polarization components.

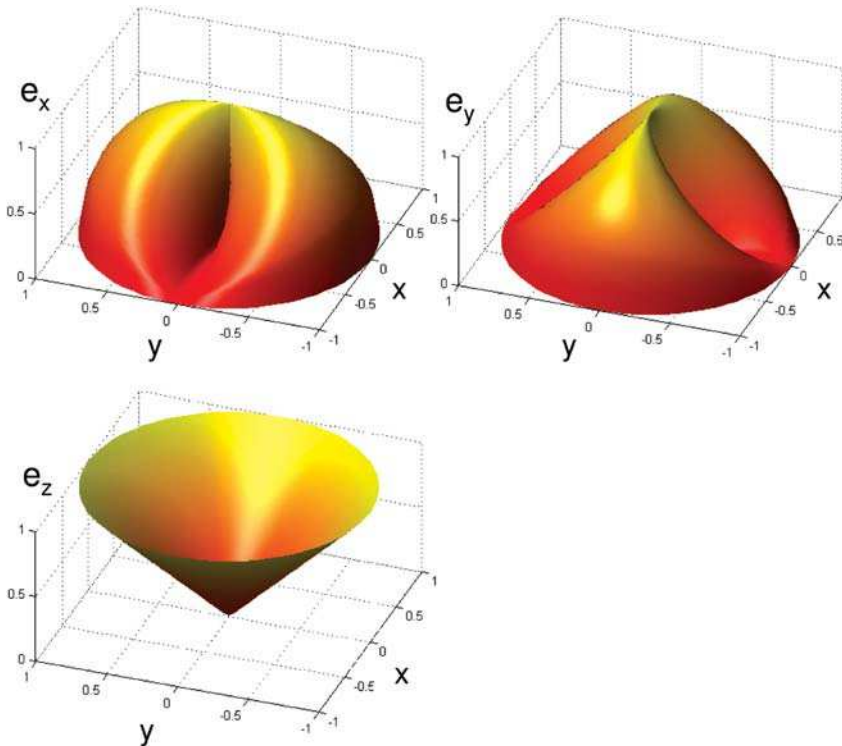


Figure 17-16: Magnitude of free-space transfer function for vector fields. Note that e_x and e_y have identical shape but they are rotated by 90° .

In general, the free-space transfer function of the vector components of the electromagnetic field has to be multiplied by an additional damping function. In the x - z plane, for example, the E_x -component is damped by a factor $\sim \cos \alpha$, while the E_z -component is damped by a factor $\sim \sin \alpha$ (Figure 17-15 left-hand side). Only the transverse component E_y is independent of the propagation angle. In general it follows in isotropic media that

$$e_x \cdot \sin \alpha \cos \varphi + e_y \cdot \sin \alpha \sin \varphi + e_z \cdot \cos \alpha = 0 \quad (17-89)$$

with the polar angle φ . The free-space vector transfer function is therefore proportional to

$$e_x = \cos \alpha \sin \varphi, \quad (17-90a)$$

$$e_y = \cos \alpha \cos \varphi, \quad (17-90b)$$

$$e_z = \sin \alpha. \quad (17-90c)$$

17.6

Evanescient Waves

So far only propagating solutions of the wave equation have been considered. In general, solutions of the stationary wave equation are written with a complex propagation vector as

$$U(x, y, z) = A \cdot e^{i(k_x x + k_y y + k_z z)} \quad (17-91)$$

with propagating solutions given by

$$k_z = \pm \sqrt{\left(\frac{2\pi \cdot n}{\lambda}\right)^2 - k_x^2 - k_y^2} \quad (17-92)$$

and non-propagating solutions as

$$k_z = \pm i \sqrt{\left(\frac{2\pi \cdot n}{\lambda}\right)^2 - k_x^2 - k_y^2}. \quad (17-93)$$

After introduction of the real damping constant χ for the complex propagation vector component k_z with

$$\chi = \sqrt{k_x^2 + k_y^2 - \left(\frac{2\pi \cdot n}{\lambda}\right)^2}, \quad (17-94)$$

the damped solution of the wave equation can be written as

$$U(x, y, z) = A \cdot e^{i(k_x x + k_y y)} \cdot e^{-\chi z} \tag{17-95}$$

As will be shown in chapter 18, evanescent waves play an important role in diffraction. However, evanescent waves have a very limited range and can therefore be neglected in most applications of optical imaging. Evanescent waves, however, may occur also at the refraction at an interface from larger refractive index n_1 to lower refractive index n_2 (total reflection, see figure 17-17). Beyond the so-called critical angle α_g given by

$$\sin \alpha_g = \frac{n_2}{n_1} \tag{17-96}$$

no propagating spatial frequency spectrum components exist in the medium with the lower refractive index. The boundary conditions require a continuous field in the weaker medium. Since the spatial frequency is larger than the allowed wavelength, the wave cannot get away from the boundary and so propagate into the medium.

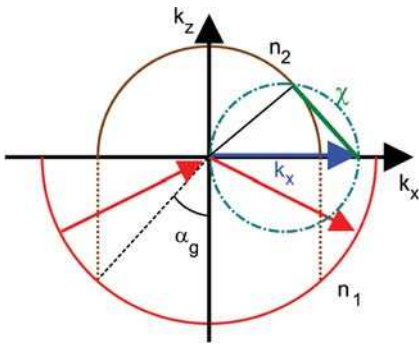


Figure 17-17: Evanescent fields at total reflection.

The range of the evanescent field is given by the inverse of the damping factor χ . The damping factor according to eq. (17-94) is shown in figure 17-18.

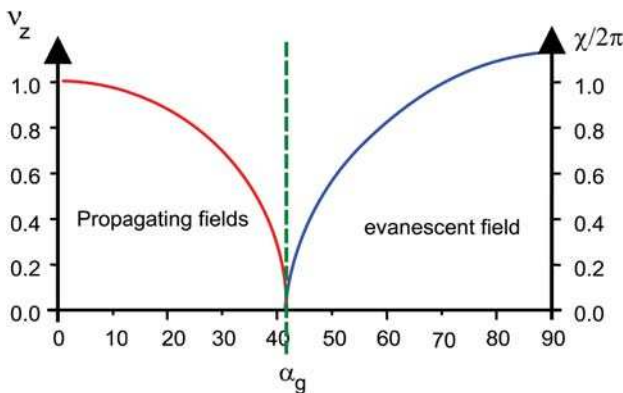


Figure 17-18: Propagating spatial frequency and damping factor for different angles of incidence for the example of refraction at an interface ($n_1 = 1.5$, $n_2 = 1.0$).

17.7

Approximative Solutions to the Wave Equation

17.7.1

Geometrical Optics and the Eikonal Equation

If the complex field amplitude U is separated into amplitude $A(x,y)$ and phase function $W(x,y)$ by

$$U(x, y) = A(x, y) \cdot e^{ik_0 W(x,y)} \quad (17-97)$$

one obtains after insertion into the Helmholtz equation (17-17)

$$\begin{aligned} \Delta U + k^2 U &= \Delta A \cdot e^{ik_0 W} + A \cdot \Delta e^{ik_0 W} + 2\vec{\nabla} A \cdot \vec{\nabla} e^{ik_0 W} + k^2 A \cdot e^{ik_0 W} \\ &= \left(\Delta A + ik_0 A \cdot \Delta W - k_0^2 A \cdot (\vec{\nabla} W)^2 + 2ik_0 \vec{\nabla} A \cdot \vec{\nabla} W + k^2 A \right) \cdot e^{ik_0 W} = 0, \end{aligned} \quad (17-98)$$

having used the equations

$$\vec{\nabla} e^{ik_0 W} = ik_0 \vec{\nabla} W e^{ik_0 W}, \quad (17-99a)$$

$$\Delta e^{ik_0 W} = ik_0 \Delta W e^{ik_0 W} - k^2 (\vec{\nabla} W)^2 e^{ik_0 W}. \quad (17-99b)$$

Both the real and imaginary parts of eq. (17-98) have to vanish separately, thus it follows that

$$\Delta A - k_0^2 A \cdot (\vec{\nabla} W)^2 + k^2 \cdot A = 0, \quad (17-100)$$

$$A \cdot \Delta W + 2\vec{\nabla} A \cdot \vec{\nabla} W = 0. \quad (17-101)$$

If rapid variations of the amplitude, A , can be neglected, i.e.,

$$\Delta A \ll k^2 \cdot A \quad (17-102)$$

the Eikonal equation is obtained from the real part (17-100):

$$(\vec{\nabla} W)^2 = n^2. \quad (17-103)$$

From the imaginary part of the solution of the wave equation (17-101), the continuity equation for the irradiance is derived. After multiplication of eq. (17-101) by the amplitude A , one obtains for the irradiance (or intensity) $I = A^2$ in analogy with eq. (17-55)

$$I \cdot \Delta W + \vec{\nabla} I \cdot \vec{\nabla} W = \vec{\nabla} (I \cdot \vec{\nabla} W) = \vec{\nabla} (I \cdot \vec{p}) = \oint_{\Sigma} I \vec{p} \cdot d\vec{a} = 0, \quad (17-104)$$

with the optical direction cosine \vec{p} (eq. (17-68)) and the integral over the closed surface Σ with surface elements $d\vec{a}$. By the continuity of irradiance, the integral of the irradiance I times the direction cosine vector over any closed surface Σ must vanish and so energy is conserved.

The Eikonal equation is the fundamental equation for a geometrical-optical description. As a consequence, the surfaces of constant phase can be interpreted as wavefronts with geometrical-optical light-rays travelling normal to them. Therefore the Eikonal equation can be written as a vector equation

$$\vec{\nabla}W = n\vec{s} = \vec{p} \quad (17-105)$$

with the wavefront normal \vec{s} as a tangent vector to the light ray and the optical direction cosine vector \vec{p} . In the slowly-varying amplitude approximation any wave can be interpreted locally as a plane wave travelling along the normal direction to the local wavefront, given by $\vec{\nabla}W$. After series expansion of the phase function in the neighbourhood of a point r_0 by

$$W = W_0 + r_0 \left. \frac{\partial W}{\partial r} \right|_{r_0} + \dots \quad (17-106)$$

one obtains, after insertion into (17-97)

$$U(\vec{r}) = A(\vec{r}_0) e^{ik_0(W_0 + \vec{r}_0 \cdot \vec{p}_0)}, \quad (17-107)$$

which is a plane wave of phase W_0 with wave vector $\vec{k} = k_0\vec{p}$. By a reverse argument, single rays are often interpreted as plane waves.

The restriction to slowly-varying amplitude A is in many cases insignificant. Only due to interference caused by the superposition of several waves may the amplitude vary significantly over a wavelength and eq. (17-102) is no longer applicable. Therefore the slowly-varying amplitude approximation breaks down in focal regions. It has to be noted that this restriction does not apply to the geometrical-optical description. In regions of interference it is only the transition from wave optics to geometrical optics, and vice versa, that is more complicated. For the derivation of the geometrical-optical light rays from the complex amplitude of an interference pattern of two waves, for example, the complex amplitude is practically described by two phase functions $W_1(x,y)$ and $W_2(x,y)$ instead of one and equation (17-103) may be applied to the phase functions W_1 and W_2 separately. More details on the transition between wave optics and geometrical optics will be given in chapter 20.

17.7.2

Paraxial Wave Equation

For small propagation angles α with respect to the z-axis, the fast variation with the coordinate z is separated by the factorization of the complex amplitude by

$$U(x, y, z) = A \cdot e^{i(k_x x + k_y y)} \cdot e^{ik_z z} = V(x, y, z) \cdot e^{ik_z z}. \quad (17-108)$$

Insertion into the Helmholtz equation (17-17) gives:

$$\begin{aligned} \Delta \bar{U} + k^2 \bar{U} &= \Delta \bar{V} \cdot e^{ik_z z} + V \cdot \frac{\partial^2}{\partial z^2} e^{ik_z z} + 2 \cdot \frac{\partial}{\partial z} V \cdot \frac{\partial}{\partial z} e^{ik_z z} + k^2 \cdot V \cdot e^{ik_z z} \\ &= \left[\Delta \bar{V} - k_z^2 \cdot V + 2ik_z \cdot \frac{\partial V}{\partial z} + k^2 \cdot V \right] \cdot e^{ik_z z} = 0. \end{aligned} \quad (17-109)$$

In the slowly-varying envelope limit, k_x^2 and k_y^2 can be neglected and $\vec{k} \approx k_z$. If the remaining z-dependence of the amplitude distribution V is weak, the second derivative of V can be neglected:

$$\left| \frac{\partial^2 V}{\partial z^2} \right| \ll k \cdot \left| \frac{\partial V}{\partial z} \right|. \quad (17-110)$$

As a result the paraxial wave equation is obtained:

$$\Delta_{\perp} V + 2ik \frac{\partial}{\partial z} V = 0 \quad (17-111)$$

with the transverse Laplace operator Δ_{\perp} given by

$$\Delta_{\perp} = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}. \quad (17-112)$$

The paraxial wave equation (17-111) is also called the slowly-varying envelope approximation (SVE) of the wave equation [17-11]. Since $k_z = 2\pi/\lambda \cos \alpha$, the scope of the SVE is equivalent to the scope of the approximation of $\cos \alpha$ by 1. Allowing for a maximum error of 10%, the propagation angle should not exceed $\approx 25^\circ$.

17.7.3

Transport of Intensity

The variation of the intensity with the z-coordinate is given by

$$\frac{\partial I}{\partial z} = \frac{\partial VV^*}{\partial z} = V \frac{\partial V^*}{\partial z} + V^* \frac{\partial V}{\partial z} = \frac{1}{2ik} (V \Delta_{\perp} V^* - V^* \Delta_{\perp} V) \quad (17-113)$$

where the last equality holds from eq. (17-111). After insertion of the separation of the complex field amplitude V in amplitude A and phase W by

$$V(x, y, z) = A(x, y, z) \cdot e^{ikW}, \quad (17-114)$$

one obtains for the paraxial transport equation of intensity [17-12], [17-13]:

$$\frac{\partial I}{\partial z} = -I \cdot \Delta_{\perp} W - \vec{\nabla}_{\perp} I \cdot \vec{\nabla}_{\perp} W = -\vec{\nabla}_{\perp} \left(I \cdot \vec{\nabla}_{\perp} W \right) \quad (17-115)$$

According to the paraxial transport of the intensity equation (TIE), the development of the intensity with the propagation coordinate z depends on the transverse gradients of the intensity and the phase. The shape of the phase surface, especially the gradient, which determines the normal to the wavefront, influences the form of the intensity profile at a distance z . Equation (17-115) offers a basis for the recovery of the phase of the field from pure intensity information. In a discrete solution of the phase retrieval problem, intensity information at several z -positions is required [17-14], [17-15], [17-16], [17-17].

17.7.4

Gaussian Beams

The paraxial wave equation (17-107) has a special set of solutions, given by the Hermite–Gaussian modes:

$$U_{n,m}(x, y, z) = A \frac{w_0}{w(z)} \cdot H_n \left(\sqrt{2} \frac{x}{w(z)} \right) \cdot H_m \left(\sqrt{2} \frac{y}{w(z)} \right) \cdot e^{-\frac{(x^2+y^2)}{w(z)^2}} \cdot e^{-i \left[kz - (m+n+1) \arctan \left(\frac{z}{z_0} \right) + \frac{\pi}{\lambda R(z)} (x^2+y^2) \right]} \quad (17-116)$$

Here m is the order of the mode, $w(z)$ is the beam radius, R is the radius of curvature for the spherical phase surface. The propagation distance z_0 , at which the beam radius $w(z)$ increases by a factor of $\sqrt{2}$ with respect to the minimum beam radius w_0 , is the so-called Rayleigh length or confocal parameter:

$$z_0 = \frac{\pi w_0^2}{\lambda} \quad (17-117)$$

H_m is the Hermite polynomial of degree m given by

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2} \quad (17-118)$$

Figure 17-19 illustrates the normalized Hermite–Gauss modes $U(x,0,0)$ of order n , $m=0$ for $n=0..5$ with $w=1$. The order number n determines the number of nodes of the mode. The Hermite–Gauss polynomials are usually applied to the description of coherent laser modes [17-11].

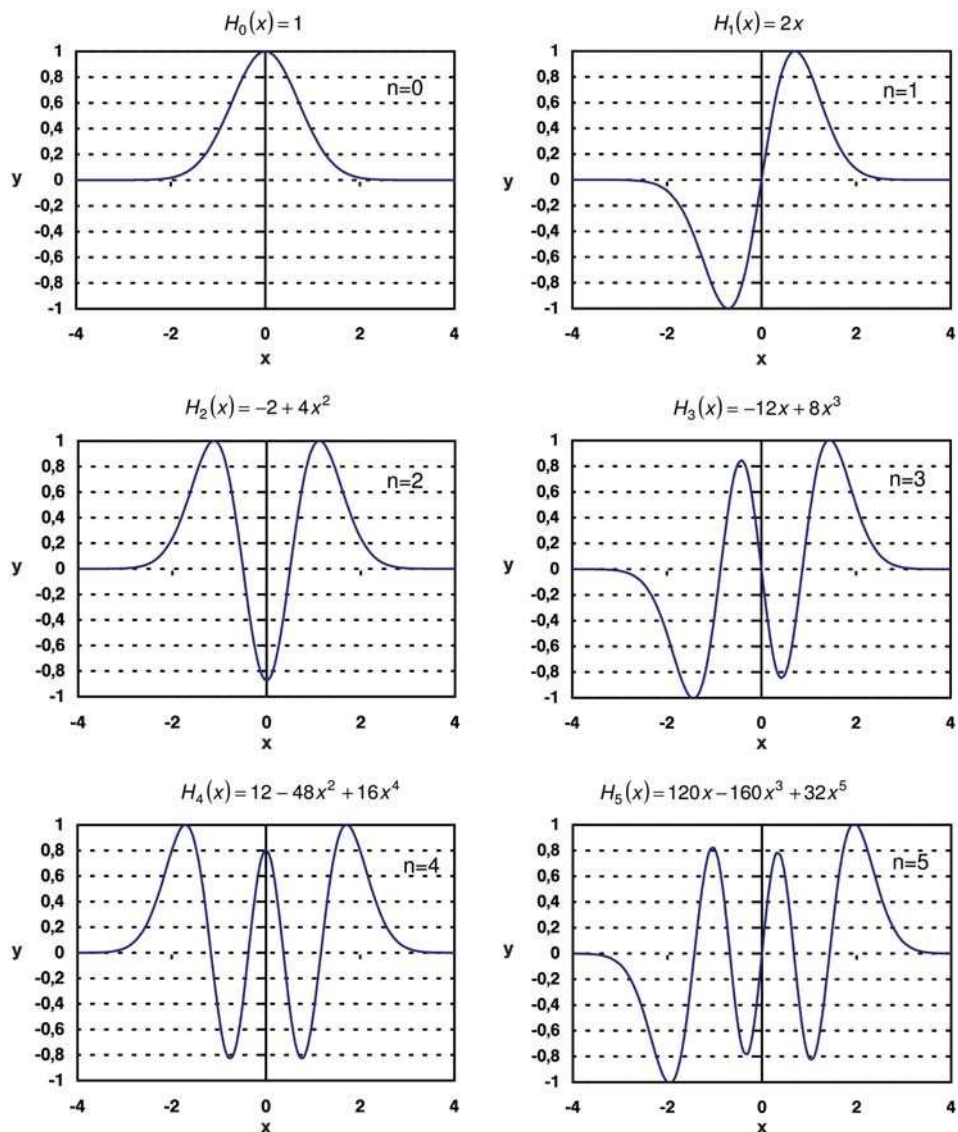


Figure 17-19: Hermite-Gauss polynomial of order $n=0$...5 with $w=1$.

In the lowest order $n=0$, the solutions of the paraxial wave equation are given by the Gaussian mode. Gaussian beams allow for a simplified treatment of optical systems and are discussed therefore in more detail. From eq. (17-118) one obtains for the Gaussian beam at the beam waist position in plane $z=0$ (with $R=\infty$):

$$U(x, y) = A \cdot \exp\left[-\frac{(x^2 + y^2)}{w_0^2}\right]. \quad (17-119)$$

After Insertion of eq. (17-119) into the Fresnel diffraction integral (see chapter 18), for the complex amplitude in the plane z we obtain

$$U(x, y, z) = \frac{iA\pi}{\lambda} \frac{2w_0^2}{2z + ikw_0^2} \cdot \exp\left[-\frac{(x^2 + y^2)}{w^2(z)}\right] \cdot \exp\left\{-ik\left[z + \frac{x^2 + y^2}{2R(z)}\right]\right\}. \quad (17-120)$$

Figure 17-20 illustrates the amplitude of a Gaussian beam of order $n = 0$. The beam has no exactly defined boundary.

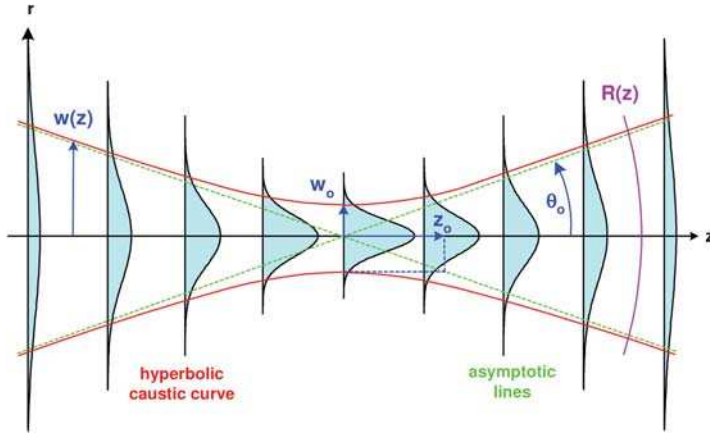


Figure 17-20: Gaussian beam profile.

The transverse coordinate $x = w(z)$ is defined as the beam radius. At its position the complex amplitude drops to $U_0(w) = A/e$, and the normalized intensity drops to $1/e^2 = 0.135$ relative to the centre of the profile for the basic mode ($n = m = 0$). It has to be noted, that there is still a remarkable portion of the energy of the beam for values $r > w$. The beam radius w is given by hyperbolic dependency on the propagation distance z

$$w(z) = w_0 \sqrt{1 + \left(\frac{z}{z_0}\right)^2} \quad (17-121)$$

where the beam waist w_0 is the smallest value of the beam radius at the beam waist position (here $z = 0$). For small z , eq. (17-121) approximates to a parabola, whereas for large z the asymptotic approximation holds:

$$w(z) \sim \frac{\lambda z}{\pi w_0} = \frac{w_0}{z_0} \cdot z = \tan \theta \cdot z \approx NA \cdot z \quad (17-122)$$

with the beam divergence angle θ given by

$$\theta = \frac{\lambda}{\pi w_0}. \quad (17-123)$$

In the region of the waist of length $2z_0$, diffraction dominates, while in the asymptotic approximation the geometric-optical approximation can be applied.

With eq. (17-122) the product of the beam waist and the beam divergence is proportional to the wavelength. The phase of the field distribution is always spherical due to the paraxial approximation, and its dependence on the propagation distance z is given by

$$R(z) = z + \frac{z_0^2}{z} . \tag{17-124}$$

The content of eq. (17-124) is illustrated in Figure 17-21. At the beam waist position the phase is constant, while far away from the waist the phase function is concentric to the waist position. The minimum radius of curvature is attained at the Rayleigh distance z_0 with $R_{\min} = 2z_0$.

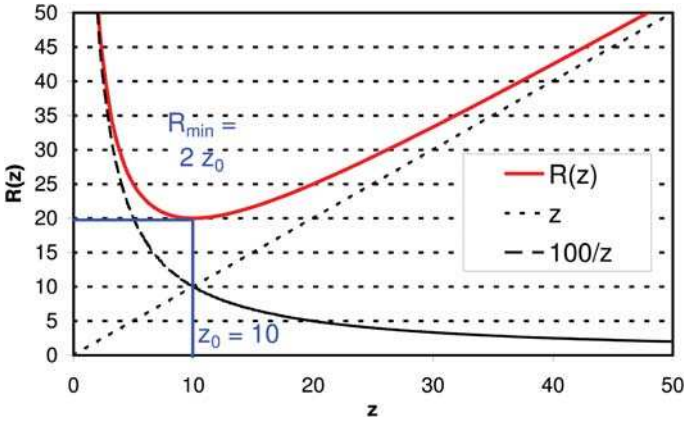


Figure 17-21: Radius of curvature in dependence on the propagation distance z for $z_0 = 10$.

Since a Gaussian beam is a solution of the wave equation, diffraction effects are incorporated. The minimum beam radius at the beam waist is given by eq. (17-122) with $w_0 \approx \lambda / (\pi \cdot NA)$. Furthermore, it is impossible to generate an ideal collimated beam, since the Gaussian beam always has a finite divergence angle and diverges under free space propagation. In total there are only two parameters defining the form of the beam caustic, two of the four parameters w_0 , z_0 , λ_0 or θ_0 can be selected arbitrarily. They are related by the equation

$$z_0 = \frac{\pi w_0^2}{\lambda} = \frac{w_0}{\theta_0} = \frac{\lambda_0}{\pi \theta_0^2} . \tag{17-125}$$

The intensity distribution of a Gaussian beam is given by

$$I(x, y, z) = \frac{2P}{\pi w^2} \cdot \exp\left[-\frac{2(x^2 + y^2)}{w^2}\right] \tag{17-126}$$

with the total power P of the beam

$$P = \int_0^{\infty} I(r) 2\pi r dr = \frac{\pi}{2} I_0 w^2. \quad (17-127)$$

In each plane z the transverse beam profile is again of Gaussian shape of width $w(z)$. The intensity profile on the z -axis with $x=y=0$ is given by a Lorentz function:

$$I(0, z) = \frac{2P}{\pi \cdot w_0^2} \cdot \frac{1}{1 + \left(\frac{z - z_T}{z_0}\right)^2} \quad (17-128)$$

Figure 17-22 illustrates the intensity distribution of a Gaussian beam through the beam waist position. The transverse and longitudinal profiles of the beam are illustrated in figure 17-23.

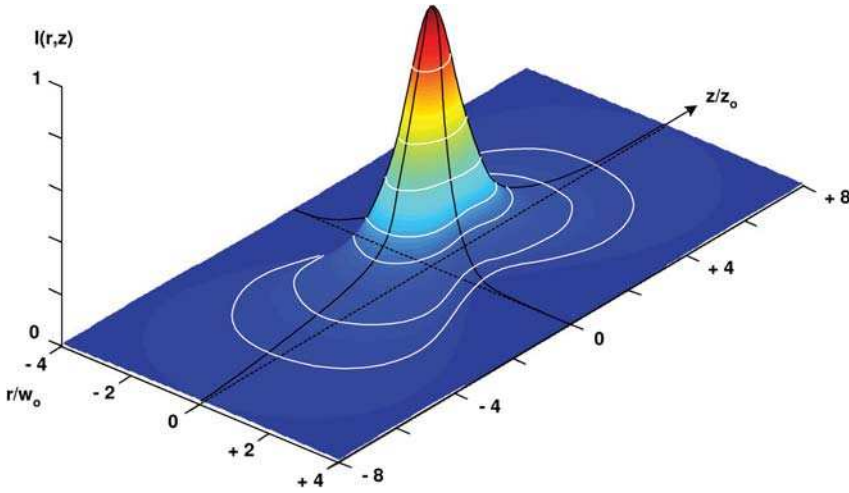


Figure 17-22: Intensity distribution of a Gaussian beam at its waist position.

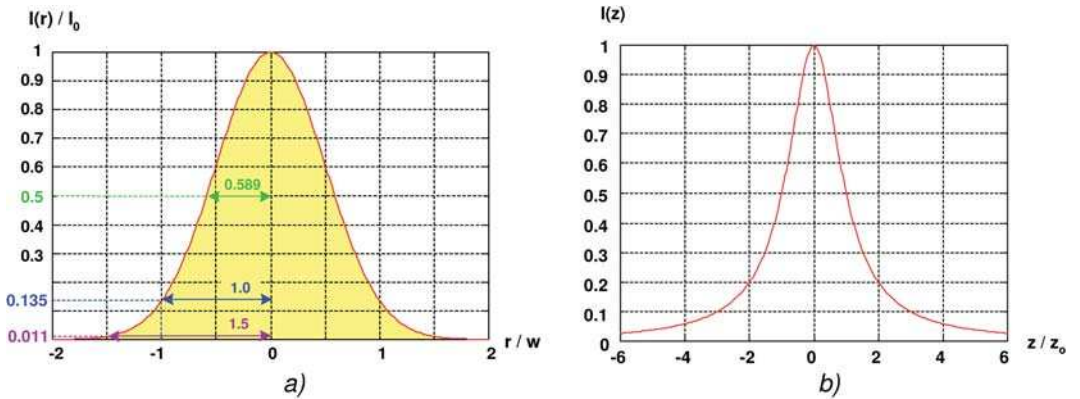


Figure 17-23: Transverse intensity profile (a) and longitudinal intensity profile (b) of a Gaussian beam.

The normalized encircled energy function of this beam profile is given by

$$EncE(r) = \frac{1}{P} \int_0^r I(r') 2\pi r' dr' = 1 - e^{-2\left(\frac{r}{w}\right)^2}. \quad (17-129)$$

The encircled energy is illustrated in figure 17-24.

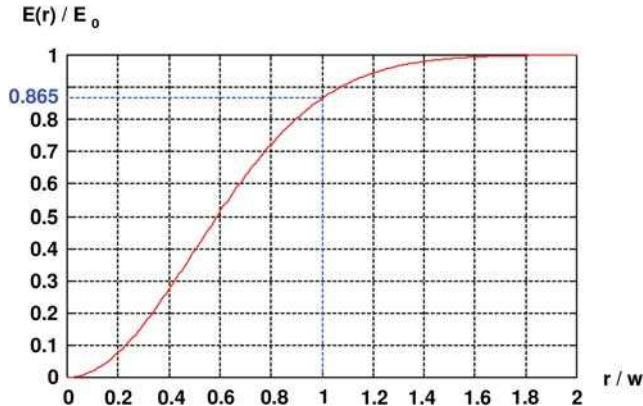


Figure 17-24: Encircled energy function of a Gaussian beam.

17.7.5

Ray Equivalent of Gaussian Beams

In a geometrical optical description using rays, Gaussian beams can be characterized by two rays, as illustrated in figure 17-25 [17-18], [17-19]. The first ray parallel to the axis of symmetry of the Gaussian beam is called the waist ray, while the divergence ray follows the asymptotic envelope of the Gaussian beam in the far field (compare figure 17-20). From the ray data in an arbitrary plane, the beam parameters can be calculated as follows;

$$w_0 = \frac{x_{wr} \cdot \tan \theta_0 - x_{dr} \cdot \tan \alpha}{\sqrt{\tan^2 \theta_0 + \tan^2 \alpha}}, \quad (17-130)$$

$$z_0 = \frac{x_{dr} \cdot \tan \theta_0 + x_{wr} \cdot \tan \alpha}{\tan^2 \theta_0 + \tan^2 \alpha}. \quad (17-131)$$

The beam radius can be evaluated at a position z , with the help of the two ray intersection heights, by

$$w = \sqrt{\alpha^2 + \theta_0^2}. \quad (17-132)$$

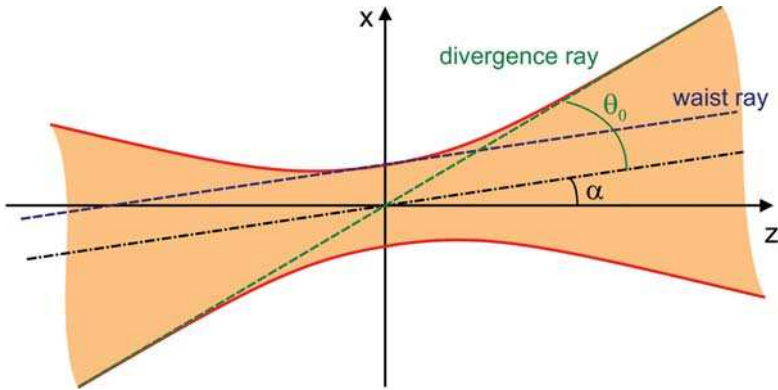


Figure 17-25: Ray equivalent description of a Gaussian beam by the waist and the divergence ray.

17.7.6

Gaussian Beams in Two Dimensions

The solution of the paraxial wave equation, given by (17-116), is separated into two orthogonal descriptions independent of each other. The general description of the two-dimensional Gaussian beams is thus straightforward, however, the beam parameters in different directions might be different, and astigmatic and anamorphic Gaussian beams may be obtained. In the simplest case, the two sections are orthogonal to each other, without any coupling between them. Figure 17-26 illustrates a beam of this kind for several different distances z , where the divergences and the waist positions are different. The transverse cross-section shows a rotating elliptical area of varying eccentricity. At two z -positions the cross-section of the beam ellipse becomes circular. In general the phase surface is not plane at any z -position, i.e., even at the 'waist' positions the wavefront might be described by a plane wave. There are two locations, however, where the phase is cylindrical, at all other z -locations the wavefront is given by a toroidal surface [17-20]. In more general astigmatic beams the two sections are coupled and there is a twist term in the mode function. Other types of generalized beams are:

1. decentred Gaussian beams [17-21],
2. non-paraxial Gaussian beams [17-22],
3. Gaussian beams with more general profile forms [17-23], [17-24],
4. Gaussian beams with more general phase surfaces [17-25].

These kinds of beams will be discussed in more detail in a later volume of this book.

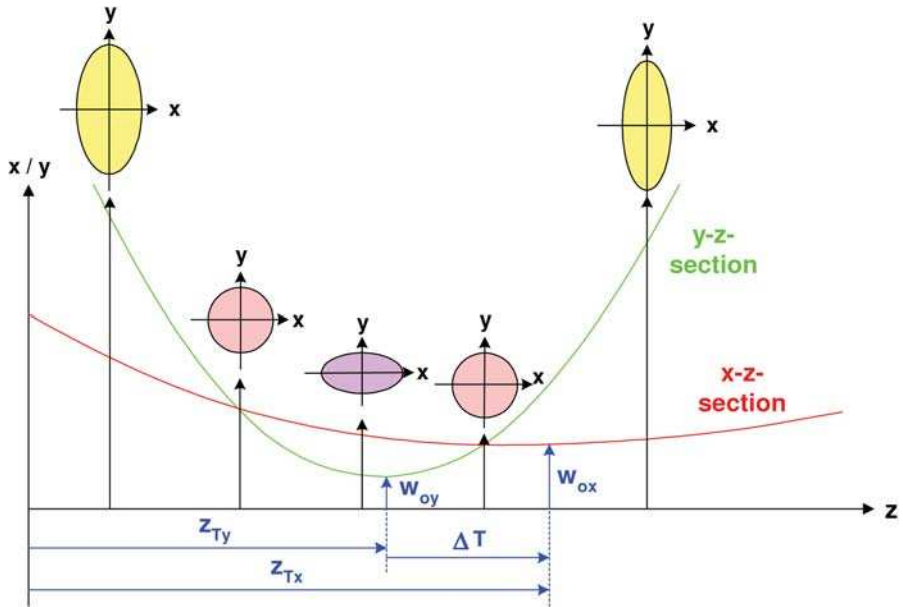


Figure 17-26: Diagram of the Gaussian beam transfer for an elliptical cross section.

17.8

Literature

- 17-1 J.D. Jackson, *Classical Electrodynamics*, 2nd ed. (Wiley, 1975).
- 17-2 D. Marcuse, *Light Transmission Optics* (Academic Press, New York, 1982).
- 17-3 A. Sommerfeld, *Vorlesungen über Theoretische Physik*, Bd. 3 "Optik", 3rd ed. (Leipzig, 1960).
- 17-4 R. Becker, F. Sauter, *Theorie der Elektrizität*, 21st ed. (Teubner, 1973).
- 17-5 A.V. Gitin, Radiometry. A comprehensive approach, *J. Opt. Technol.* **65**, 132–140 (1998).
- 17-6 W.T. Cathey, *Optical Information Processing and Holography* (Wiley, 1974).
- 17-7 J.W. Goodman, *Introduction to Fourier Optics* (McGraw-Hill, New York, 1968).
- 17-8 H.F. Talbot, *Phil. Mag.* **9**, No. 56, 401–407 (1836).
- 17-9 W. D. Montgomery, "Self-imaging Objects of Infinite Aperture", *J. Opt. Soc. Am.* **57**, 772–778 (1967).
- 17-10 R.K. Luneburg, *Mathematical Theory of Optics* (University of California Press, 1966).
- 17-11 A. E. Siegman, *Lasers* (Oxford University Press, Oxford, 1986).
- 17-12 M. R. Teague, Deterministic phase retrieval: a Green's function solution, *J. Opt. Soc. Am.* **73**, 1434 (1983).
- 17-13 M. R. Teague, Image formation in terms of the transport equation, *J. Opt. Soc. Am. A* **2**, 2019 (1985).
- 17-14 K. A. Nugent, T. E. Gureyev, D. F. Cookson, D. Paganin and Z. Barnea, Quantitative phase imaging using hard X rays, *Phys. Rev. Lett.* **77**, 2961 (1996).
- 17-15 T. E. Gureyev and K. A. Nugent, Rapid quantitative phase imaging using the transport of intensity equation, *Opt. Comm.* **133**, 339 (1997).
- 17-16 J. A. Quiroga, J. A. Gomez-Pedrero and J. C. Martinez-Anton, Wavefront measurement by solving the irradiance transport

- equation for multifocal systems, *Opt. Eng.* **40**, 2885 (2001).
- 17-17** V. V. Volkov, Y. Zhu and M. De Graef, A new symmetrized solution for phase retrieval using the transport of intensity equation, *Micron* **33**, 411 (2002).
- 17-18** R. Herloski, S. Marshall and R. Antos, Gaussian beam ray-equivalent modeling and optical design, *Applied Optics* **22**, 1168 (1983).
- 17-19** J. D. Zook and T. C. Lee, Geometrical interpretation of Gaussian beam optics, *Appl. Opt.* **11**, 2140 (1972).
- 17-20** J. A. Arnaud and H. Kogelnik, Gaussian Light Beams with General Astigmatism, *Appl. Opt.* **8**, 1687 (1969).
- 17-21** Y. Cai and Q. Lin, Decentered elliptical Gaussian beam, *Appl. Opt.* **41**, 4336 (2002).
- 17-22** S. Nemoto, Nonparaxial Gaussian beams, *Appl. Opt.* **29**, 1940 (1990).
- 17-23** R. Borghi, Elegant Laguerre-Gauss beams as a new toll for describing antisymmetric flattened Gaussian beams, *J. Opt. Soc. Am. A* **18**, 1627 (2001).
- 17-24** A. Belafhal and M. Ibnchaikh, Propagation properties of Hermite-cosh-Gaussian laser beams, *Opt. Comm.* **186**, 269 (2000).
- 17-25** S. Saghafi and C. J. Sheppard, Near field and far field of elegant Hermite-Gaussian and Laguerre-Gaussian modes, *J. Mod. Opt.* **45**, (1998).

18

Scalar Diffraction

- 18.1 Introduction 42
- 18.2 Kirchhoff Diffraction Integral 44
 - 18.2.1 Inconsistency of the Kirchhoff Diffraction Integral 48
- 18.3 1st and 2nd Rayleigh–Sommerfeld Diffraction Integral 48
- 18.4 Two-dimensional Diffraction 50
- 18.5 Huygens Principle 52
- 18.6 Fourier Space Formulation 54
- 18.7 Examples of Scalar Diffraction Patterns 57
 - 18.7.1 Diffraction Fields Behind Slits 57
 - 18.7.2 Diffraction by a Rectangular Aperture 59
- 18.8 Fresnel Diffraction 60
 - 18.8.1 Computation 61
 - 18.8.2 Validity 62
- 18.9 Collin’s Fresnel Diffraction Integral 64
 - 18.9.1 Definition 64
 - 18.9.2 Example 67
- 18.10 Fraunhofer Diffraction 69
- 18.11 Grating Diffraction 71
 - 18.11.1 Ronchi Grating 71
 - 18.11.2 The Sinusoidal Phase Grating and Surface Fabrication Errors 76
- 18.12 Scalar Diffraction at Dielectric Objects 79
- 18.13 Babinet’s Principle 82
- 18.14 Scalar Scattering 85
- 18.15 Boundary Diffraction Waves 89
 - 18.15.1 Geometrical Theory of Diffraction 90
 - 18.15.2 An Empirical Boundary Diffraction Wave 94
- 18.16 Literature 96

18.1

Introduction

Diffraction denotes the deviation of light from rectilinear propagation that is not due either to refraction or reflection (figure 18-1).

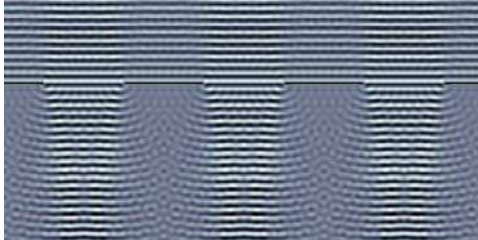


Figure 18-1: Diffraction of a scalar wave at a grating (numerical simulation, plane wave incident from the top).

From the viewpoint of an optics designer, diffraction has two aspects. On one hand it causes a diffraction-limited spot size and a finite resolution limit in optical far-field imaging, while on the other hand it enables the production of compact “diffractive” optics that permits, for example, the fast and convenient generation of user-defined, aspherical wavefronts or the reconstruction of 3D lightfields by means of holography.

There are several powerful methods used in the calculation of field transport and propagation problems as numerical solutions to the wave equation. The optimal choice of the algorithms depends on the special parameters of the task and therefore cannot be decided in a general way. In the following table the most important methods are listed for comparison.

As will be discussed in section 27, the general diffraction problem is an electromagnetic one. The task is to find the resulting electromagnetic field for a given distribution of the permittivity $\epsilon(r)$, the permeability $\mu(r)$, or the conductivity $\sigma(r)$. The present chapter presents scalar solutions where the vector character of light is neglected. Classical scalar diffraction theory (according to Kirchhoff, Rayleigh–Sommerfeld, Fresnel and Fraunhofer) does not solve a boundary *condition* problem but a boundary *value* problem. Furthermore, the scalar Kirchhoff diffraction integral solves an over-specified boundary value problem and leads frequently to inconsistent results. Inconsistent means that the boundary values are not reproduced. The first and second Rayleigh–Sommerfeld diffraction integrals are consistent but they are limited to boundary values on a plane. The Fresnel and Fraunhofer diffraction integrals are approximations for large distances and small diffraction angles. All these methods, their prerequisites and limitations will be discussed in detail in this chapter.

To end this introduction we give a brief outline of the subjects that are discussed here in addition to classical scalar diffraction theory.

Basic principle	Method	Application
Mode expansion	Plane wave expansion (Fourier method)	
	Gaussian modes	Laser beams
	Spherical waves	Scattering problems
	Bessel beams	Fibers, waveguides
Integral evaluation	Kirchhoff	
	Fresnel, convolution	
	Fraunhofer, Fourier integral	Far field
	Rayleigh–Sommerfeld	
	Debye integral	
Asymptotic solution of the wave equation	Watson transform	Cylinder geometry
	Geometrical theory of diffraction	
Analytical methods	Uniform asymptotic theory	
	Stationary phase	
	Cylinder diffraction	
Numerical solution of Maxwell's equations	Mie-scattering	Scattering by spheres
	Finite differences	Imaging, simulation
	Finite elements	Imaging, simulation
	Vector potential	
	Rigorous coupled waves	Imaging, simulation

Mathematically, scalar diffraction integrals are convolution integrals. They convolve the boundary values with an elemental excitation, the Huygens wavelet. In a more depictive way Huygens principle considers each point of the wavefront as the source of a Huygens wavelet. The interference of all wavelets produces the observed diffraction pattern. According to the convolution theorem of Fourier Theory, there exists a complementary description in terms of spatial frequencies: The convolution with the Huygens wavelets corresponds to multiplication of the Fourier transform of the boundary values by the Fourier transform of the wavelets. This leads naturally to the angular spectrum formulation of diffraction in terms of plane waves. These methods can be applied to aperture diffraction as well as to diffraction at phase objects. Numerous examples of computations for scalar diffraction will be given.

The well-known Babinet principle describes the similarity of the Fraunhofer diffraction patterns (apart from the central peak) behind complementary opaque obstacles in which the transparent and opaque parts are reversed. It will be discussed as a direct consequence of the linearity of the diffraction integrals. Following the far-field version we will discuss its less well-known extension to near-field diffraction as well as to diffraction at phase objects.

Based on the inclusion of a source term in the scalar wave equation, scattering theory can be formulated in scalar terms. The angular spectrum of the scattered field is given by the convolution of the incident angular spectrum with the three-dimensional Fourier transform of the inhomogeneity. Multiple scattering can be considered by the Born series expansion. A particularly nice feature of this theory is

that the Born approximation is also applicable for the approximate treatment of refraction, diffraction and scattering at surfaces.

In the theory of boundary diffraction waves, which is given at the end of this chapter, diffraction fields are divided into two parts: an undisturbed geometrical optics field and a boundary diffraction wave, originating from the rim of the diffracting object. The boundary diffraction wave is also the central concept of the geometrical diffraction theory of Keller (which becomes discontinuous at the shadow boundary) and its successors: the uniform asymptotic theory and the uniform theory of diffraction. The best-known example of a boundary diffraction wave is certainly Sommerfeld's exact solution of the diffraction by a perfectly conducting half-plane.

18.2

Kirchhoff Diffraction Integral

The Kirchhoff diffraction integral is a solution of the scalar, time-independent wave equation (Helmholtz equation, cf. chapter 17) using Green's relation [18-1]

$$\oint_S \left(U \frac{\partial G}{\partial \vec{s}} - G \frac{\partial U}{\partial \vec{s}} \right) d\sigma = - \iiint_V (U \Delta G - G \Delta U) dV, \quad (18-1)$$

where V denotes an arbitrary closed volume with surface S , \vec{s} is the inner normal vector of this surface, U and G are functions that have a second continuous derivative in V and $d\sigma$ and dV are infinitesimal surface and volume elements, respectively.

If U and G both satisfy the free-space Helmholtz equation

$$\begin{aligned} \Delta U + k^2 U &= 0, \\ \Delta G + k^2 G &= 0, \end{aligned} \quad (18-2)$$

the volume integral is zero and we obtain

$$\oint_S \left(U \frac{\partial G}{\partial \vec{s}} - G \frac{\partial U}{\partial \vec{s}} \right) d\sigma = 0. \quad (18-3)$$

Now consider G as an analogue to the "probe-charge" of electrostatic theory: A spherical wave (the free-space Green's function) that emanates from a point \vec{r}' inside the volume V

$$G = \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|}. \quad (18-4)$$

Inserting it into Green's relation we have to exclude the point of origin of the spherical wave because there it becomes singular. To do this a small sphere is drawn around \vec{r}' and Green's relation is applied to the volume that remains (figure 18-2):

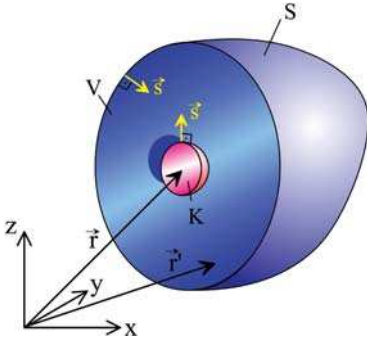


Figure 18-2: Slice through the volume V in which Green's relation is applied. A small sphere around the point of observation is excluded from the integration.

$$\oint_S \left(U \frac{\partial G}{\partial s} - G \frac{\partial U}{\partial s} \right) d\sigma + \oint_K \left(U \frac{\partial G}{\partial s} - G \frac{\partial U}{\partial s} \right) d\sigma = 0 \quad (18-5)$$

S is the outer surface of the volume and K is the surface of the recessed sphere.

The integral over the surface of the recessed sphere in the limit of vanishing radius, provides a constant contribution

$$P = -4\pi U . \quad (18-6)$$

To show this we write the limit of the integral as

$$P = \lim_{R \rightarrow 0} \oint_K \left(U \frac{\partial G}{\partial s} - G \frac{\partial U}{\partial s} \right) d\sigma \quad (18-7)$$

with

$$G = \frac{e^{ikR}}{R} . \quad (18-8)$$

The normal derivative of G on the surface K of the sphere is

$$\frac{\partial G}{\partial s} = \vec{\nabla} G \cdot \vec{s} = \left(ik - \frac{1}{R} \right) \frac{e^{ikR}}{R} . \quad (18-9)$$

According to (18.8) and (18.9) G and $\partial G/\partial s$ are constant on K . U is also constant ($U \rightarrow U(R=0)$) because for $R \rightarrow 0$ we have, in particular, $R \ll \lambda$. This simplifies (18.7) to

$$P = \lim_{R \rightarrow 0} \left\{ U \frac{\partial G}{\partial s} 4\pi R^2 \right\} - \lim_{R \rightarrow 0} G \oint_K \frac{\partial U}{\partial s} d\sigma . \quad (18-10)$$

The second term is zero because with the stereo angle Ω ($d\sigma = R^2 d\Omega$) and $U = \text{const.}$ we obtain

$$\lim_{R \rightarrow 0} G \oint_K \frac{\partial U}{\partial \vec{s}} d\sigma = \lim_{R \rightarrow 0} \frac{e^{ikR}}{R} \oint_K \frac{\partial U}{\partial \vec{s}} R^2 d\Omega = 0 . \tag{18-11}$$

This leaves the first term only in (18.10), which after inserting (18.9), becomes

$$P = \lim_{R \rightarrow 0} \left\{ U \left(ik - \frac{1}{R} \right) \frac{e^{ikR}}{R} 4\pi R^2 \right\} = -4\pi U . \tag{18-12}$$

Eventually the Kirchhoff diffraction integral becomes

$$U(\vec{r}) = \frac{1}{4\pi} \oint_S \left(U(\vec{r}') \frac{\partial}{\partial \vec{s}} \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|} - \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|} \frac{\partial}{\partial \vec{s}} U(\vec{r}') \right) d\sigma . \tag{18-13}$$

This describes the complex amplitude of light inside a closed volume V as a function of the complex amplitude on its surface S .

It should be remembered that Green's function G serves as a probing function which originates from the point \vec{r} under consideration, inside the volume. In the framework of Huygens principle its meaning can be reinterpreted. It now originates from the surface point \vec{r}' and propagates to the observed point \vec{r} . In this interpretation, Kirchhoff's diffraction integral describes the interference of Huygens wavelets which originate from the surface of the volume under consideration. Their excitation strength is given by the amplitude U and its normal derivative $\partial U/\partial \vec{s}$. Inside the volume they provide the field $U(\vec{r})$. Outside they sum to zero.

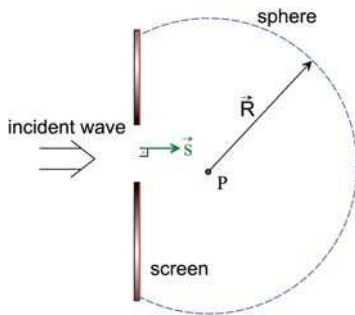


Figure 18-3: Application of the Kirchhoff diffraction integral to the diffraction produced by an aperture in a black screen. The enclosing surface is divided into three regions: back of screen, sphere, and aperture.

In the following we use the Kirchhoff integral to describe the diffraction by an aperture A in an infinitesimal thin, perfectly absorbing, screen. The surface of the closed volume consists of three parts:

1. the aperture;
2. the back of the screen;
3. a sphere of infinite radius around the observation point P enclosing the volume.

The surface integral over the infinite sphere with center r becomes with $R = r - r'$ and $d\sigma = R^2 d\Omega$ (in computing the normal derivative of $\exp\{ikR\}/R$ it must be observed that \vec{s} is the inner normal of the volume under consideration)

$$U(\vec{r}) = \lim_{R \rightarrow \infty} \frac{1}{4\pi} \iint_K \left(U(\vec{R}) \left(\frac{1}{R} - ik \right) \frac{e^{ikR}}{R} - \frac{e^{ikR}}{R} \frac{\partial}{\partial s} U(\vec{R}) \right) R^2 d\Omega \quad (18-14)$$

i.e.,

$$U(\vec{r}) = \lim_{R \rightarrow \infty} \frac{1}{4\pi} \iint_K \left(U(\vec{R}) - R \left(ikU(\vec{R}) + \frac{\partial}{\partial s} U(\vec{R}) \right) \right) e^{ikR} d\Omega. \quad (18-15)$$

The integral vanishes if

$$\lim_{R \rightarrow \infty} \left[U(\vec{R}) - R \left(ikU(\vec{R}) + \frac{\partial}{\partial s} U(\vec{R}) \right) \right] = 0. \quad (18-16)$$

If the source of $U(\vec{R})$ is a limited aperture, the finiteness condition

$$\lim_{R \rightarrow \infty} U(\vec{R}) = 0 \quad (18-17)$$

is fulfilled automatically. The remaining requirement for $U(\vec{R})$ is ‘‘Sommerfeld’s radiation condition’’

$$\lim_{R \rightarrow \infty} \left[R \left(ikU(\vec{R}) + \frac{\partial}{\partial s} U(\vec{R}) \right) \right] = 0. \quad (18-18)$$

This implies that, in the infinite limit, $U(\vec{R})$ becomes an outgoing spherical wave. i.e., $U(\vec{R})$ must vanish for $R \rightarrow \infty$ at least as fast as $1/R$.

In order to evaluate the remaining integrals over the aperture and the back of the screen, the Kirchhoff approximation is applied.

$$U(\vec{R}), \frac{\partial}{\partial s} U(\vec{R}) = \begin{cases} \text{undisturbed} & \text{in the aperture,} \\ 0 & \text{on the back of the screen.} \end{cases} \quad (18-19)$$

The Kirchhoff approximation indicates that the boundary values are obtained from geometrical optics, i.e., the boundary values are computed as if the light is propagating within the diffraction obstacle according to the laws of geometrical optics.

Using this approximation, the Kirchhoff diffraction integral consists of an integral over the aperture alone and becomes

$$\begin{aligned} U(\vec{r}) &= \frac{1}{4\pi} \iint_A \left(U(\vec{r}') \frac{\partial}{\partial s} \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|} - \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|} \frac{\partial}{\partial s} U(\vec{r}') \right) d\sigma \\ &= \frac{1}{4\pi} \iint_A \left(U(\vec{r}') \left(ik - \frac{1}{|\vec{r}-\vec{r}'|} \right) \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|} - \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|} \frac{\partial}{\partial s} U(\vec{r}') \right) d\sigma. \end{aligned} \quad (18-20)$$

Now it describes the scalar field U at a point \vec{r} behind an illuminated aperture A in a black (Kirchhoff) screen.

Inserting the directional derivative of the spherical wave along the normal \vec{s} to the aperture

$$\frac{\partial}{\partial \vec{s}} \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|} = \left(ik - \frac{1}{|\vec{r}-\vec{r}'|} \right) \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|^2} \cos \vartheta, \quad (18-21)$$

where ϑ denotes the angle enclosed by $\vec{r}-\vec{r}'$ and \vec{s} , results in

$$U(\vec{r}) = \frac{1}{4\pi} \iint_A \left(U(\vec{r}') \left(ik - \frac{1}{|\vec{r}-\vec{r}'|} \right) \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|^2} \cos \vartheta - \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|} \frac{\partial}{\partial \vec{s}} U(\vec{r}') \right) d\sigma. \quad (18-22)$$

18.2.1

Inconsistency of the Kirchhoff Diffraction Integral

The simultaneous and independent prescription of the field U and its normal derivative $\partial U/\partial \vec{s}$ over-specifies the boundary value problem. As a consequence, the appointed boundary values are not reproduced if the point of observation is chosen to be within the diffracting aperture.

Furthermore, it is mathematically incorrect to zero the field and its normal derivative simultaneously at the back of the screen. It can be shown that a three-dimensional potential that vanishes in a finite area, together with its normal derivative, must vanish everywhere [18-2]. However, despite all restrictions and inconsistencies, the Kirchoff diffraction integral frequently provides an excellent description of diffraction fields.

18.3

1st and 2nd Rayleigh–Sommerfeld Diffraction Integral

The Rayleigh–Sommerfeld diffraction integrals are obtained when the correct Green's function is inserted into the Kirchhoff diffraction integral. "Correct" means correct for the diffraction problem which is under consideration. This removes the aforementioned inconsistencies of the Kirchhoff integral and the boundary values are reproduced.

Have a second look at the Kirchhoff integral

$$U(\vec{r}) = \frac{1}{4\pi} \iint_A \left(U(\vec{r}') \frac{\partial}{\partial \vec{s}} G(\vec{r}-\vec{r}') - G(\vec{r}-\vec{r}') \frac{\partial}{\partial \vec{s}} U(\vec{r}') \right) d\sigma. \quad (18-23)$$

If U is prescribed as a boundary value on the screen, G must be zero there so that the second term in (18.23) vanishes. If instead $\partial U/\partial \vec{s}$ is the boundary value, $\partial G/\partial \vec{s}$ must be zero in order for the first term to vanish. For both conditions, G must also fulfil the Helmholtz equation and Sommerfeld's radiation condition.

In particular, for a plane screen, Green's function can be determined according to the mirror method [18-2], see also [18-3]: Green's function is formed from the superposition of two spherical waves (figure 18-4). One emanates from the point of observation P and one emanates from its mirror image P' .

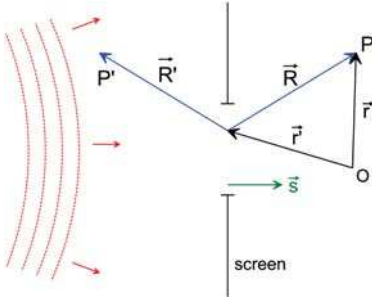


Figure 18-4: Construction of Green's function in a diffraction plane by means of the mirror method. P is the actual observation point and P' its mirror image.

Inside the plane of the screen the fields are identical (since $R = R'$), so

$$\frac{e^{ikR}}{R} = \frac{e^{ikR'}}{R'} \quad (18-24)$$

But the corresponding normal derivatives differ in sign because their sources are situated on opposite sides of the screen

$$\frac{\partial}{\partial s} \left(\frac{e^{ikR}}{R} \right) = - \frac{\partial}{\partial s} \left(\frac{e^{ikR'}}{R'} \right) \quad (18-25)$$

The principle of the construction of Rayleigh–Sommerfeld diffraction integrals is simple: The superposition of both spherical waves is chosen in such a way that the part that is not given as the boundary value, vanishes.

Boundary Value U: First Rayleigh–Sommerfeld Diffraction Integral

G is the difference of the spherical waves emanating from P and P'

$$G = \frac{e^{ikR}}{R} - \frac{e^{ikR'}}{R'} \quad (18-26)$$

Then the normal derivative is in the plane of the screen

$$\frac{\partial G}{\partial s} = 2 \frac{\partial}{\partial s} \frac{e^{ikR}}{R} \quad (18-27)$$

and the value in the plane of the screen

$$G = 0 \quad (18-28)$$

Inserting both relations into the Kirchhoff diffraction integral yields the first Rayleigh–Sommerfeld diffraction integral (RS1)

$$\begin{aligned}
 U_I(\vec{r}) &= \frac{1}{2\pi} \int_A U(\vec{r}') \frac{\partial}{\partial \bar{s}} \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|} d\sigma \\
 &= \frac{1}{2\pi} \int_A U(\vec{r}') \left(ik - \frac{1}{|\vec{r}-\vec{r}'|} \right) \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|} \cos \vartheta d\sigma .
 \end{aligned} \tag{18-29}$$

Boundary Value $\partial U/\partial \bar{s}$: Second Rayleigh–Sommerfeld Diffraction Integral

G is the sum of the spherical waves emanating from P and P'

$$G = \frac{e^{ikR}}{R} + \frac{e^{ikR'}}{R'} . \tag{18-30}$$

Then the normal derivative in the plane of the screen is

$$\frac{\partial G}{\partial \bar{s}} = 0 \tag{18-31}$$

and the value in the plane of the screen is

$$G = 2 \frac{e^{ikR}}{R} . \tag{18-32}$$

Inserting both equations into the Kirchhoff diffraction integral yields the second Rayleigh–Sommerfeld diffraction integral (RS2)

$$U_{II}(\vec{r}) = -\frac{1}{2\pi} \int_A \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|} \frac{\partial}{\partial \bar{s}} U(\vec{r}') d\sigma . \tag{18-33}$$

For a plane screen, the Kirchhoff diffraction integral is just the mean value of both Rayleigh–Sommerfeld diffraction integrals

$$U_K(\vec{r}) = \frac{1}{2} (U_{RS1}(\vec{r}) + U_{RS2}(\vec{r})) . \tag{18-34}$$

Comparisons of the presented formulations of scalar diffraction are given, for example, in [18-4]–[18-7].

18.4

Two-dimensional Diffraction

The diffraction problem is two-dimensional if the geometry of the diffracting object and the incident wave do not change along one dimension. The diffraction of a cylindrical wave by a slit in a screen is one example of this.

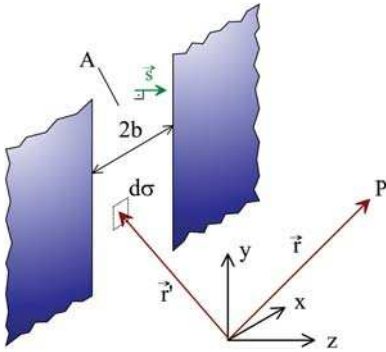


Figure 18-5: Diffraction by a slit.

For this problem the Kirchhoff diffraction integral can be integrated along the constant dimension

$$U(\vec{r}) = \frac{1}{4\pi} \int_{-\infty}^{\infty} \int_{-b}^b \left(U(\vec{r}') \frac{\partial}{\partial s} \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|} - \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|} \frac{\partial}{\partial s} U(\vec{r}') \right) dx' dy' . \quad (18-35)$$

The infinite integral can be computed analytically (again: if the incident wave is constant along this direction) by means of [18-8]

$$\int_{-\infty}^{\infty} \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|} dy' = i\pi H_0(k|\vec{\rho}-\vec{\rho}'|) . \quad (18-36)$$

H_0 is a Hankel function of the first kind, of zeroth order. ρ and ρ' are planar coordinates according to $\vec{\rho} = (x,0,z)$ and $\vec{\rho}' = (x',0,z')$. Inserting eq. (18.36) into the Kirchhoff integral yields

$$U(\vec{\rho}) = \frac{i}{4} \int_{-b}^b \left(U(\vec{\rho}') \frac{\partial}{\partial s} H_0(k|\vec{\rho}-\vec{\rho}'|) - H_0(k|\vec{\rho}-\vec{\rho}'|) \frac{\partial}{\partial s} U(\vec{\rho}') \right) dx' dy' . \quad (18-37)$$

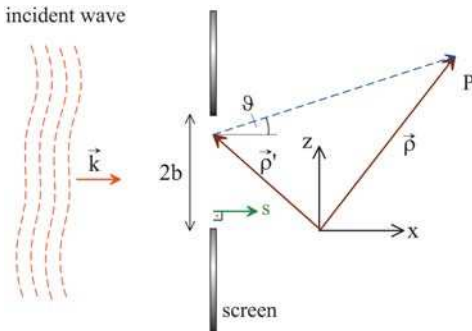


Figure 18-6: Two-dimensional diffraction by a slit.

The spherical wave as the Green's function of the three-dimensional diffraction problem is replaced, for the two-dimensional case, by the Hankel function. For two-

dimensional diffraction, Sommerfeld's radiation condition must also be replaced by its two-dimensional analogue

$$\lim_{R \rightarrow \infty} \left[\sqrt{R} \left(ikU(\vec{R}) + \frac{\partial}{\partial \vec{s}} U(\vec{R}) \right) \right] = 0 \quad (18-38)$$

i.e., U must vanish at least as an outgoing cylindrical wave in the limit of infinite radius.

By virtue of

$$\frac{d}{dx} H_0(x) = -H_1(x) \quad (18-39)$$

the normal derivative of the Hankel function becomes

$$\frac{\partial}{\partial \vec{s}} H_0(k|\vec{\rho} - \vec{\rho}'|) = kH_1(k|\vec{\rho} - \vec{\rho}'|) \cos \vartheta \quad (18-40)$$

where ϑ is the angle between $\rho - \rho'$ and \vec{s} .

Finally we obtain the following form of the two-dimensional Kirchhoff diffraction integral

$$U_K(\vec{\rho}) = \frac{i}{4} \int_{-b}^b \left(U(\vec{\rho}') kH_1(k|\vec{\rho} - \vec{\rho}'|) \cos \vartheta - H_0(k|\vec{\rho} - \vec{\rho}'|) \frac{\partial}{\partial \vec{s}} U(\vec{\rho}') \right) dx'. \quad (18-41)$$

The two-dimensional first Rayleigh–Sommerfeld diffraction integral becomes, accordingly

$$U_I(\vec{\rho}) = \frac{i}{2} \int_{-b}^b U(\vec{\rho}') kH_1(k|\vec{\rho} - \vec{\rho}'|) \cos \vartheta dx', \quad (18-42)$$

and the second Rayleigh–Sommerfeld diffraction integral becomes

$$U_{II}(\vec{\rho}) = -\frac{i}{2} \int_{-b}^b H_0(k|\vec{\rho} - \vec{\rho}'|) \frac{\partial}{\partial \vec{s}} U(\vec{\rho}') dx'. \quad (18-43)$$

18.5

Huygens Principle

To derive Huygens principle we have to reinterpret the spherical wave used in the derivation of the Kirchhoff diffraction integral (figure 18-7): By analogy with potential theory it is incidentally a “probe-wave” that originates from the point of observation (figure 18-7 left-hand side). According to Huygens principle it is the other way round: Each point of the aperture is the source of a spherical wave – the Huygens wavelet. The superposition of all Huygens wavelets provides the field at the observation point (figure 18-7 right-hand side).

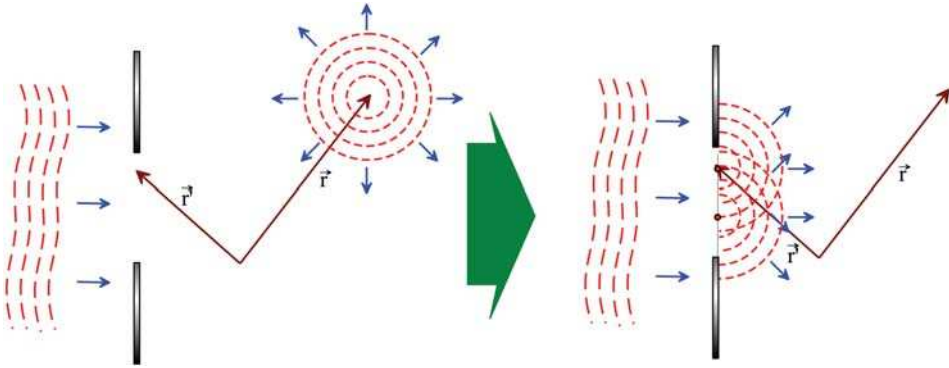


Figure 18-7: Huygens principle and the scalar diffraction integral.

In the following we give the form for the Huygens wavelets of the Rayleigh–Sommerfeld diffraction integrals.

The **first Rayleigh–Sommerfeld diffraction** integral can be written as

$$U_I(\vec{r}) = \iint_A U(\vec{r}') K_I(\vec{r} - \vec{r}') d\sigma' \quad (18-44)$$

with the Huygens wavelet

$$K_I(\vec{r} - \vec{r}') = \frac{1}{2\pi} \left(ik - \frac{1}{|\vec{r} - \vec{r}'|} \right) \frac{e^{ik|\vec{r} - \vec{r}'|}}{|\vec{r} - \vec{r}'|} \cos \vartheta . \quad (18-45)$$

The **second Rayleigh–Sommerfeld diffraction** integral can be written as

$$U_{II}(\vec{r}) = \iint_A \frac{\partial U(\vec{r}')}{\partial \vec{s}} K_{II}(\vec{r} - \vec{r}') dx' dy' \quad (18-46)$$

with the Huygens wavelet

$$K_{II}(\vec{r} - \vec{r}') = \frac{1}{2\pi} \frac{e^{ik|\vec{r} - \vec{r}'|}}{|\vec{r} - \vec{r}'|} . \quad (18-47)$$

The form of both wavelets is depicted in figure 18-8.

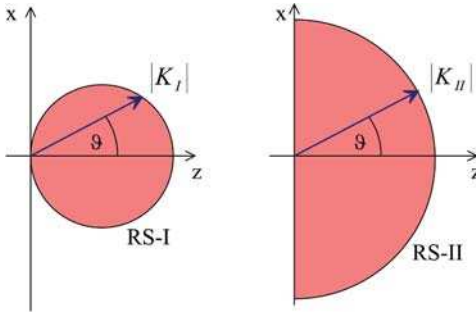


Figure 18-8: Directional characteristics of the Huygens wavelets for the first (left) and the second (right) Rayleigh–Sommerfeld diffraction integrals.

K_I has a cosine-shaped direction characteristic with a maximum value perpendicular to the aperture. There is no field radiated along $\vartheta = \pm 90^\circ$, i.e., along the aperture plane. This is a direct consequence of reproducing the assumed boundary values of the field in the plane of the aperture. If there were a radiated field in the aperture itself, the value would change there and the boundary values would not be reproduced.

The directional characteristic of K_{II} is isotropic because it is just a spherical wave, i.e., the field is radiated into the complete half-space. In particular there is also a field radiating along the aperture. Nevertheless, the boundary values are reproduced, because in the second Rayleigh–Sommerfeld diffraction integral, the boundary values are given by the normal derivative of the field. This is not changed by the spherical wave because the propagation direction is perpendicular to the normal vector of the aperture with the consequence that its normal derivative in the aperture plane is zero.

Both Rayleigh–Sommerfeld diffraction integrals can be written as a convolution

$$\begin{aligned} U_I(\vec{r}) &= U(\vec{r}) \otimes K_I(\vec{r}) , \\ U_{II}(\vec{r}) &= \frac{\partial U(\vec{r})}{\partial s} \otimes K_{II}(\vec{r}) . \end{aligned} \quad (18-48)$$

18.6 Fourier Space Formulation

Because the scalar diffraction integrals come in the form of convolution integrals, the convolution theorem can be applied to obtain a Fourier space, or angular spectrum, formulation.

$$\begin{aligned} F\{U_I(\vec{r})\} &= F\{U(\vec{r})\}F\{K_I(\vec{r})\} , \\ F\{U_{II}(\vec{r})\} &= F\left\{\frac{\partial U(\vec{r})}{\partial s}\right\}F\{K_{II}(\vec{r})\} . \end{aligned} \quad (18-49)$$

The Fourier transform $F\{\}$ of the Huygens wavelet acts as a propagation operator $P(\vec{k})$ describing the propagation of the field from the plane of the diffracting aperture to the point of observation. Having the aperture in the xy plane at $z = 0$ yields

$$\begin{aligned} U_I(x, y, z) &= F^{-1}\left\{F\{U(x, y, 0)\}P_I(\vec{k})\right\}, \\ U_{II}(x, y, z) &= F^{-1}\left\{F\left\{\frac{\partial U(x, y, 0)}{\partial z}\right\}P_{II}(\vec{k})\right\}, \end{aligned} \quad (18-50)$$

where we have placed the aperture perpendicular to the z -axis. The Fourier treatment of scalar diffraction theory is discussed in [18-3], [18-9]–[18-11].

The explicit form of the propagation operators follows from the angular spectrum of a diverging spherical wave (Weyl's representation [18-9])

$$F\left\{\frac{e^{ik_0R}}{R}\right\} = i2\pi \frac{1}{k_z} e^{ik_z z} \quad (18-51)$$

with

$$k_z = \sqrt{k_0^2 - k_x^2 - k_y^2} = 2\pi \sqrt{\frac{1}{\lambda^2} - \nu_x^2 - \nu_y^2}. \quad (18-52)$$

The propagation operator for the second Rayleigh–Sommerfeld integral becomes

$$P_I(\vec{k}, z) = i \frac{1}{k_z} e^{ik_z z}. \quad (18-53)$$

The propagation operator for the first Rayleigh–Sommerfeld integral is computed directly using the fact that the 2D Fourier transform is independent of z

$$P_{II}(\vec{k}, z) = F\left\{\frac{\partial}{\partial z} \frac{e^{ikR}}{R}\right\} = \frac{\partial}{\partial z} F\left\{\frac{e^{ikR}}{R}\right\} = e^{ik_z z}. \quad (18-54)$$

The exponential in the propagation operator is simply the phase shift of the corresponding plane wave for propagation over a distance z . This provides a simple physical explanation of scalar diffraction: The field (or its normal derivative) in the plane of the diffraction obstacle is expanded into plane waves (figure 18-9). The spatial frequencies of the field in the diffraction plane are just the projections of the corresponding propagation vectors onto that plane. The propagation of the plane waves over a distance z yields the angular spectrum of the field in the observation plane.

$$\begin{aligned} RS-I: \quad F\{U(x, y, z)\} &= e^{ik_z z} F\{U(x, y, 0)\}, \\ RS-II: \quad F\{U(x, y, z)\} &= \frac{i}{k_z} e^{ik_z z} F\{U(x, y, 0)\}. \end{aligned} \quad (18-55)$$

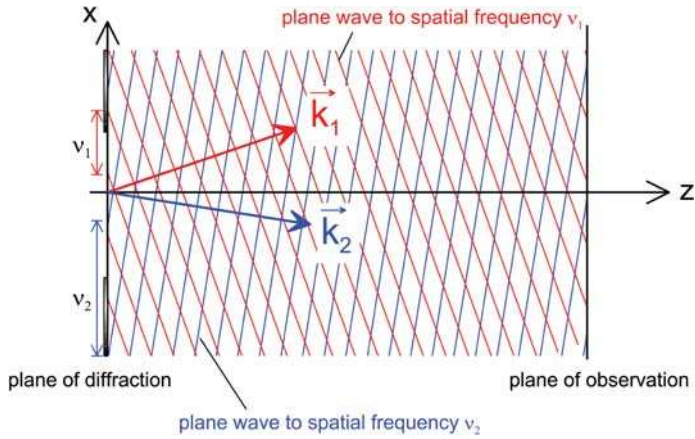


Figure 18-9: Angular spectrum representation of scalar diffraction. Two exemplary plane waves are shown ($|\vec{k}_1|=|\vec{k}_2|$).

The dependence of both propagation operators on the spatial frequency and the propagation distance is depicted in figure 18-10.

Note the exponential decay of the magnitude for $k > k_0$:

$$P_1(\vec{k}, z) = \begin{cases} e^{i\sqrt{k_0^2 - k_x^2 - k_y^2} z} & : k_x^2 + k_y^2 \leq k_0^2, \\ e^{-\sqrt{k_x^2 + k_y^2 - k_0^2} z} & : k_x^2 + k_y^2 > k_0^2. \end{cases} \quad (18-56)$$

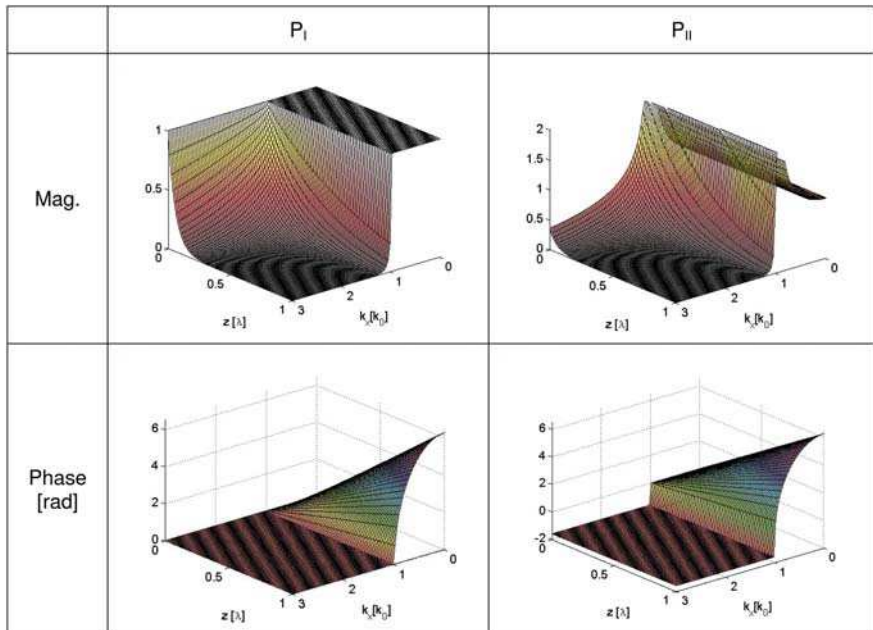


Figure 18-10: Magnitude and phase of propagation operators.

For $\sqrt{k_x^2 + k_y^2} \leq k_0$ the propagation operator describes an undisturbed propagation into free space. For $\sqrt{k_x^2 + k_y^2} > k_0$ the exponent becomes real, resulting in an exponential decay of the plane wave components with larger magnitude than the free-space propagation constant. In order to detect these components one has to place the detector at a distance $< \lambda$ from the structure. This is the physical basis for near-field optical microscopy.

The propagating plane wave components are phase-shifted according to their different angles to the z-axis while the phase of the evanescent components remains constant.

The magnitude of P_{II} (propagation operator of Rayleigh–Sommerfeld II) shows a singularity at $k_x = k_0$, i.e., at a spatial frequency of $\nu = 1/\lambda$ (figure 18-11). It stems from the plane wave component propagating along the plane of the aperture (cf. the angular spectrum of a spherical wave).

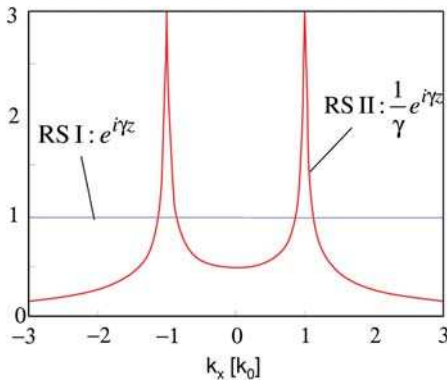


Figure 18-11: Magnitude of propagation operators at $z = 0$.

18.7

Examples of Scalar Diffraction Patterns

18.7.1

Diffraction Fields Behind Slits

In figure 18-12 the diffraction at slits of various widths in a black screen is compared according to the first and second Rayleigh–Sommerfeld diffraction integrals. The width of the slits increases from 0.5 to 10 wavelengths. The intensity is indicated by the color map. In addition, the iso-phase lines are plotted in blue.

The general appearance of the fields, according to the first and second Rayleigh–Sommerfeld diffraction integrals, is quite similar. Note the stronger confinement of the diffraction pattern for the smallest slit according to RS-I. This stems from the obliquity factor. The differences between RS-I and RS-II are most pronounced for distances from the slit which are within the slit width. Beyond this region, the results are almost indistinguishable.

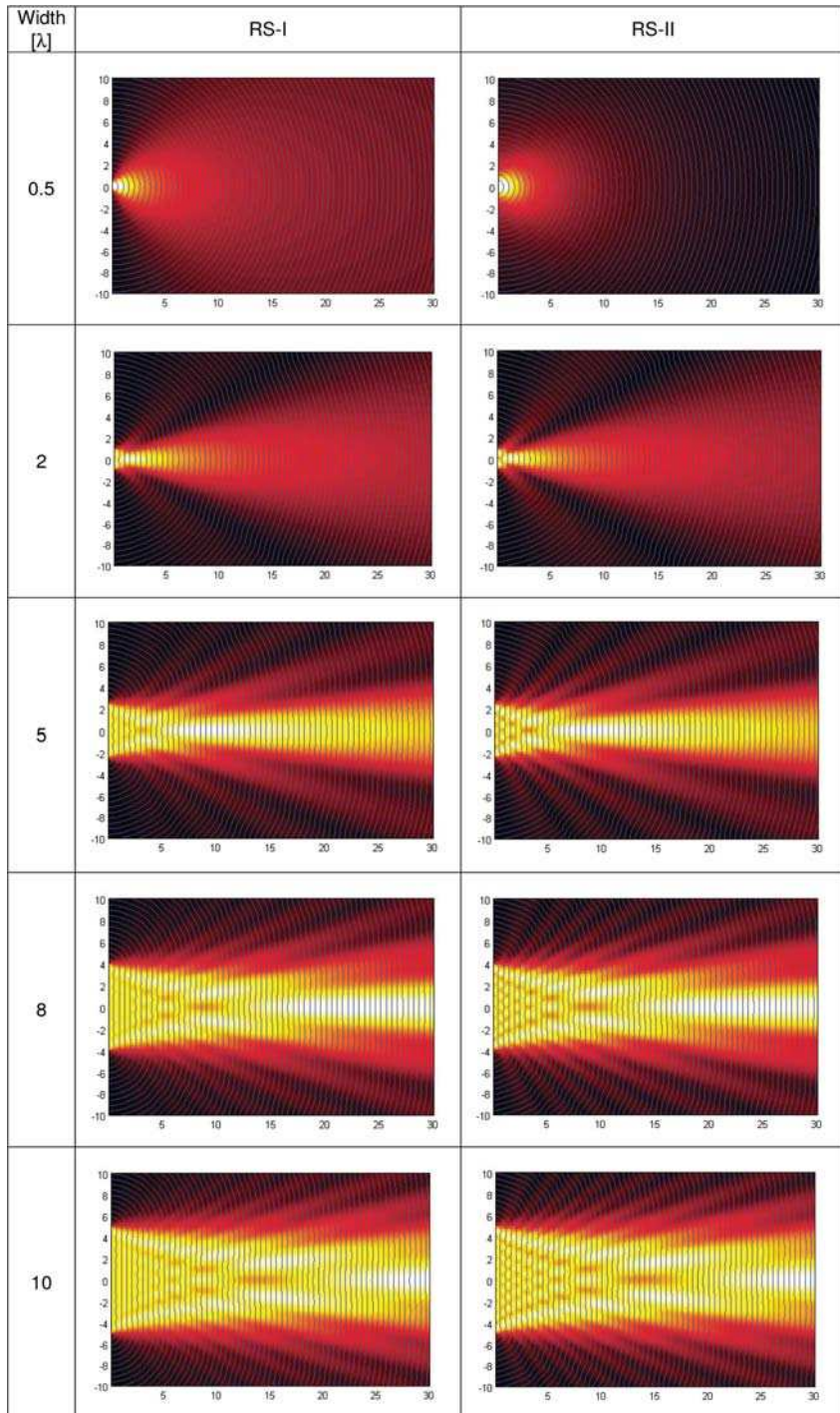


Figure 18-12: Examples of slit diffraction patterns for increasing slit width. Color map: field-amplitude, blue lines = equi-phase lines.

18.7.2

Diffraction by a Rectangular Aperture

Figure 18-13 shows the diffraction pattern behind a rectangular aperture of 5λ width and 3λ height computed according to the first Rayleigh–Sommerfeld diffraction integral. Note the continuous transition from the near-field region, where the shape of the aperture is reproduced, to the “far field” with distinguished diffraction lobes. We have chosen a rectangular shape because then the development of the far-field diffraction pattern out of the near-field pattern, can be observed. According to the inverse scaling law of Fourier theory the width of the diffraction pattern of a small aperture is larger than the width of the corresponding pattern behind a large aperture. So the intensity distribution of the Fraunhofer diffraction pattern of the horizontal rectangular aperture

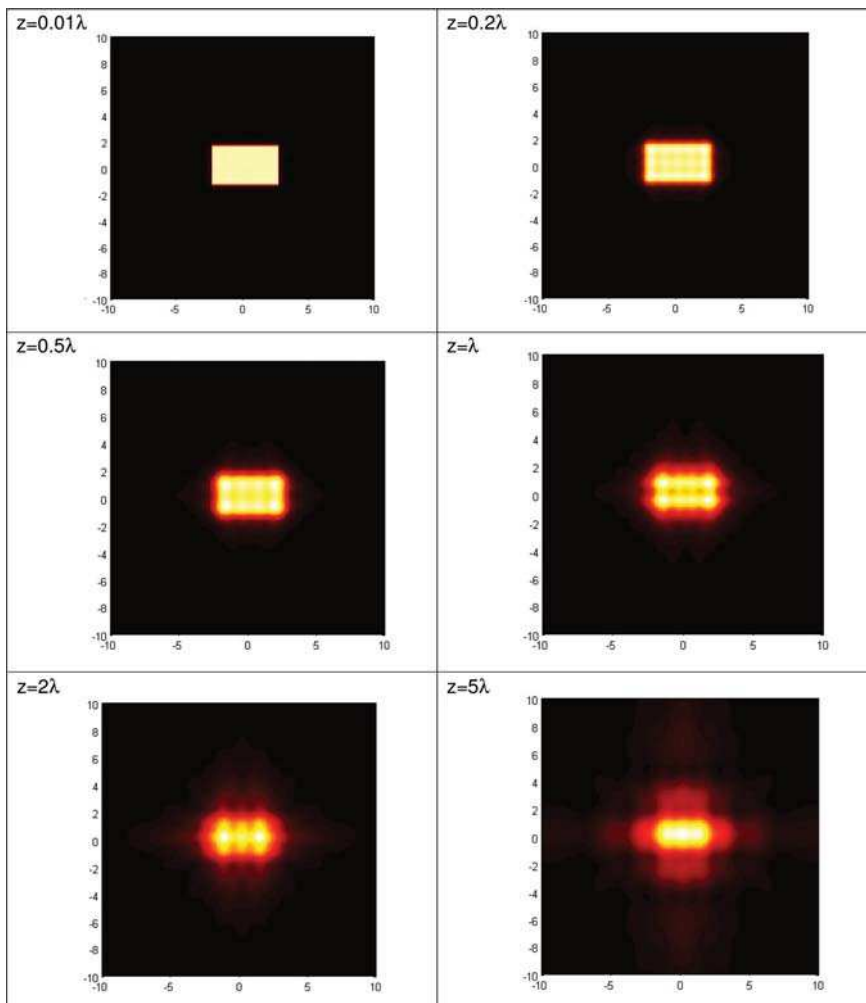


Figure 18-13: Diffraction pattern of a rectangular aperture for increasing distance according to the first Rayleigh–Sommerfeld diffraction integral.

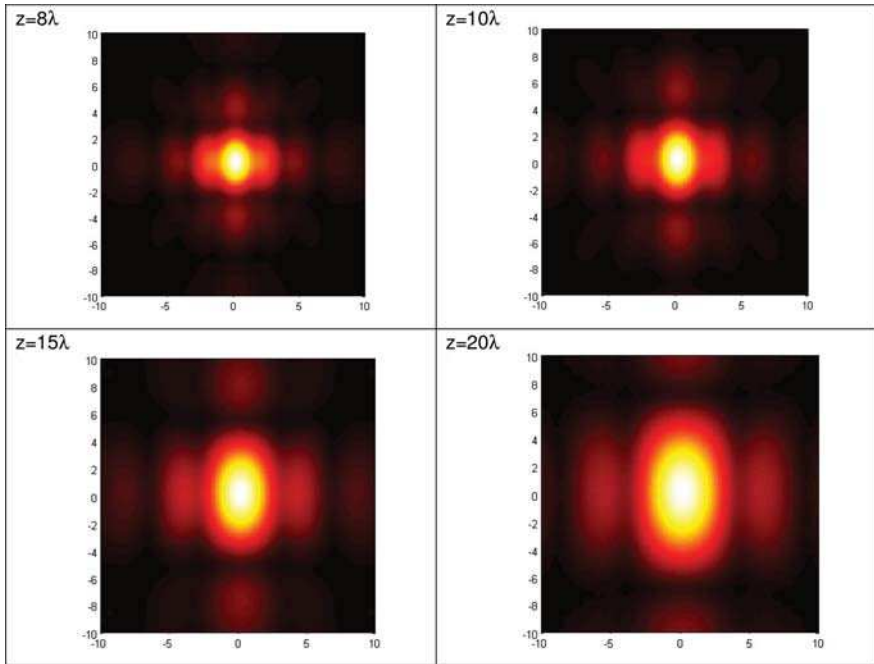


Figure 18-13: Continued.

in figure 18.13 should show a larger extent in vertical direction. The near field, however, has a larger extent in the horizontal direction. Traversing from a distance $z = 0.01\lambda$ to $z = 20\lambda$ we can see the gradual change in the diffraction pattern from a mainly horizontal orientation to a mainly vertical orientation with a break-even point around 10λ .

18.8 Fresnel Diffraction

The Fresnel approximation allows a simplified computation of scalar diffraction fields for large distances ($\gg \lambda$) and small diffraction angles. It describes the transition region in the z -direction between the extreme near field, that requires a direct evaluation of the diffraction integrals, and the far field as given by Fraunhofer diffraction.

Assuming the diffraction aperture is at $z' = 0$, the Fresnel approximation follows from an expansion of the distance of a point in the object plane $\vec{r}' = (x', y', 0)$ and the point of observation $\vec{r} = (x, y, z)$ into a Taylor series

$$\begin{aligned}
 |\vec{r} - \vec{r}'| &= z + \frac{1}{2} \frac{(x - x')^2 + (y - y')^2}{z} - \frac{1}{8} \frac{((x - x')^2 + (y - y')^2)^2}{z^3} + \dots \\
 &\approx z + \frac{1}{2} \frac{(x - x')^2 + (y - y')^2}{z}
 \end{aligned} \tag{18-57}$$

which is truncated after the quadratic term. Inserting the quadratic approximation into the first Rayleigh–Sommerfeld diffraction integral (RS1)

$$U(\vec{r}) = \frac{1}{2\pi} \iint_A U(\vec{r}') \left(ik_0 - \frac{1}{|\vec{r} - \vec{r}'|} \right) \frac{e^{ik_0|\vec{r} - \vec{r}'|}}{|\vec{r} - \vec{r}'|} \cos \vartheta d\sigma \quad (18-58)$$

and neglecting terms quadratic in $(r-r')$, a replacement of $(r-r')$ by z in the amplitude prefactor, and $\cos \vartheta \rightarrow 1$ yields the Fresnel diffraction integral

$$U(\vec{r}) = \frac{ik_0 e^{ik_0 z}}{2\pi z} \iint_A U(\vec{r}') e^{ik_0 \frac{(x-x')^2 + (y-y')^2}{2z}} d\sigma. \quad (18-59)$$

In terms of the Huygens principle, the Fresnel integral is a diffraction integral with parabolic Huygens wavelets. It can be written as a convolution

$$U(\vec{r}) = U_0(\vec{r}) \otimes K_F(\vec{r}) \quad (18-60)$$

with the convolution kernel

$$K_F(\vec{r}) = \frac{ik_0 e^{ik_0 z}}{2\pi z} e^{ik_0 \frac{(x-x')^2 + (y-y')^2}{2z}}. \quad (18-61)$$

18.8.1

Computation

Basically there are three different ways to evaluate the Fresnel integral: using

- a) the propagation operator;
 - b) the Fourier transform with quadratic prefactor;
 - c) Fresnel integrals.
- a) The propagation operator is obtained, as for the Rayleigh–Sommerfeld integrals, by Fourier transforming the convolution kernel

$$P_F(\vec{k}) = \frac{ik_0 e^{ikz}}{2\pi z} e^{i \left(k_0^2 - \frac{k_x^2 + k_y^2}{2} \right) z}. \quad (18-62)$$

It yields a Fresnel diffraction integral of

$$U(x, y, z) = \frac{ik_0 e^{ikz}}{2\pi z} F^{-1} \left\{ F\{U(x, y, 0)\} e^{i \left(k_0^2 - \frac{k_x^2 + k_y^2}{2} \right) z} \right\}. \quad (18-63)$$

Because of the underlying approximation, the propagation operator provides physically meaningless results for large spatial frequencies. This can be concluded from the fact that all spatial frequencies are propagating. Evanescent components do not exist in Fresnel approximation.

b) The exponential in (18.59) can be factorized further to yield

$$U(\vec{r}) = \frac{ie^{ikz}}{\lambda z} e^{ik\frac{x^2+y^2}{2z}} \iint_A U(\vec{r}') e^{ik\frac{x'^2+y'^2}{2z}} e^{-ik\frac{xx'+yy'}{z}} dx' dy' . \quad (18-64)$$

According to this evaluation, the Fresnel diffraction of a scalar field $U(x',y')$ is equivalent to the Fourier transform of the field multiplied by a quadratic phase factor (*within* the Fourier integral)

$$U(\vec{r}) = \frac{ie^{ikz}}{\lambda} e^{ik\frac{x^2+y^2}{2z}} F\left\{U(\vec{r}') e^{ik\frac{x'^2+y'^2}{2z}}\right\}_{v_x=\frac{x}{z}, v_y=\frac{y}{z}} . \quad (18-65)$$

c) The Fresnel integral is defined as (see appendix too)

$$F(w) = \int_0^w e^{i2\pi x^2} dx = C(w) + iS(w) \quad (18-66)$$

with

$$\left. \begin{aligned} C(w) &= \int_0^w \cos\left(\frac{\pi}{2} x^2\right) dx \\ S(w) &= \int_0^w \sin\left(\frac{\pi}{2} x^2\right) dx \end{aligned} \right\} \begin{aligned} C(0) &= S(0) = 0 \\ C(\infty) &= S(\infty) = \frac{1}{2} \\ C(-w) &= -C(w), \quad S(-w) = -S(w). \end{aligned} \quad (18-67)$$

It can be used to compute the Fresnel diffraction pattern behind an aperture.

18.8.2

Validity

The error of the Fresnel approximation is equal to the error of the quadratic approximation (18.57). It is smaller than the maximum value of the first neglected summand. If, in the same way as Papoulis [18-12], we allow an angle-error of $\pi/100$ in the exponent, then for the lateral range of validity of the Fresnel approximation in one dimension, for a diffraction aperture of width b we obtain

$$k \frac{(|x| + b)^4}{8z^3} < \frac{\pi}{100}, \quad (18-68)$$

and

$$\frac{|x|}{\lambda} < \frac{1}{\sqrt{5}} \left(\frac{z}{\lambda}\right)^{0.75} - \frac{b}{\lambda}. \quad (18-69)$$

So the minimum distance for the validity of the Fresnel approximation is

$$z_{\min} = 5^{2/3} \left(\frac{b}{\lambda} \right)^{4/3} \quad (18-70)$$

and for larger z , x must be below a curve depicted in figure 18-14.

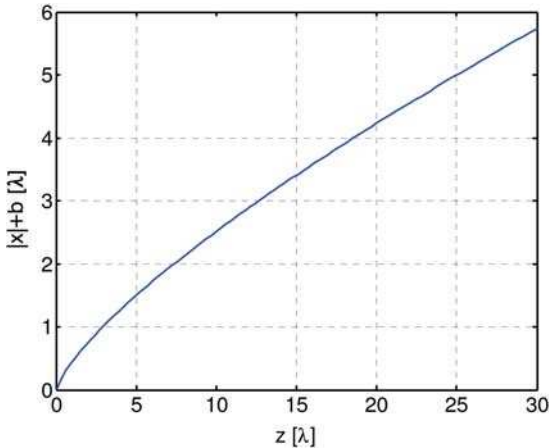


Figure 18-14 Lateral region of validity of the Fresnel approximation according to Papoulis [18-12].

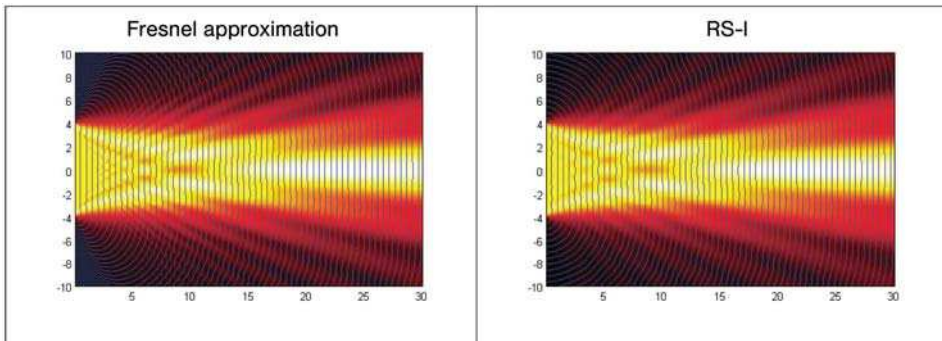


Figure 18-15: Slit diffraction pattern for Fresnel and Rayleigh–Sommerfeld diffraction. Color map: field amplitude, blue lines = equi-phase lines.

A comparison of the computed near-field of a slit of width 8λ is shown in figure 18-15, on the left-hand side, according to the angular spectrum formulation of the Fresnel integral (18.63) and on the right-hand side according to the first Rayleigh–Sommerfeld integral. Note the higher fringe density near the aperture according to the Fresnel integral and the contour of the phase lines near the screen. While they show a spherical wavefront computed from the first Rayleigh–Sommerfeld integral

(which makes sense) they are incredibly dense, according to the Fresnel approximation. One reason for this behaviour is that the evanescent components are not damped in the Fresnel approximation.

18.9

Collin's Fresnel Diffraction Integral

18.9.1

Definition

Consider the Fresnel diffraction integral (18.59) reduced to one transverse dimension

$$E(x_2, z_2) = \frac{\sqrt{i} \cdot e^{ikz}}{\sqrt{\lambda z}} \int_{-a}^a E(x_1, z_1) \cdot e^{-\frac{i\pi}{\lambda z} (x_2 - x_1)^2} dx_1. \quad (18-71)$$

In practical cases, there are very often optical systems which are located between the starting plane (index 1) and the receiving plane (index 2). If these components or subsystems can be considered as corrected and paraxial, the optical path length L between two corresponding planes can be easily written in terms of the elements of the paraxial system matrix (see vol 1, chapter 2)

$$\mathbf{M} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \quad (18-72)$$

in the form

$$L = \frac{Ax_1^2 - 2x_1x_2 + Dx_2^2}{2B}. \quad (18-73)$$

Using this, a generalized Fresnel integral is obtained which is named after Collins [18-13], [18-14]:

$$E(x_2) = \frac{\sqrt{i}}{\sqrt{\lambda B}} \int_{-a}^a E(x_1) \cdot e^{\frac{i\pi}{\lambda B} (Ax_1^2 - 2x_1x_2 + Dx_2^2)} dx_1. \quad (18-74)$$

Here the elements of the matrix occurring in the exponent constitute the eikonal function. Figure 18-16 shows the principle of the idea, the paraxial system is completely described by the three elements A , B and D of the matrix, the element C can be calculated with the help of the determinant

$$\det(\mathbf{M}) = 1. \quad (18-75)$$

There are only three independent parameters in the matrix, where identical refractive indices are assumed in the starting and the receiving space. The equation (18.74) for the Fresnel diffraction integral is a paraxial approximation, but it is able

to describe the propagation of a field with wave aberrations and special intensity profiles through nearly-perfect optical systems. Therefore its areas of application are mainly in the simulation of laser optical systems, where the radiation has a strong directionality and the Fresnel approximation gives very accurate results. The calculation of optical resonators and laser-beam delivery systems are mainly based on the above formulation of the Fresnel integral.

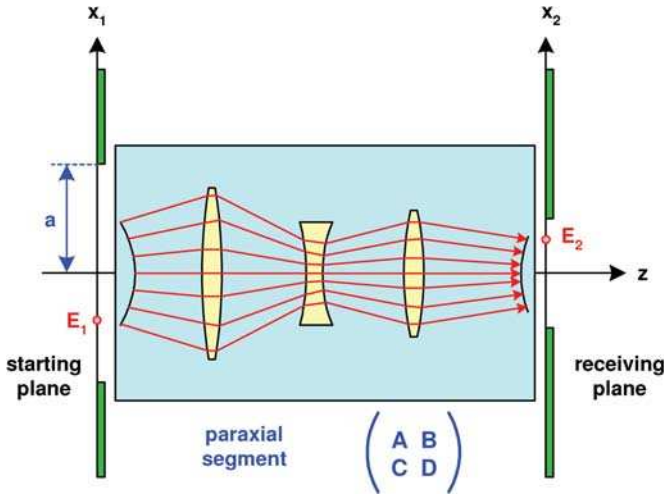


Figure 18-16: Propagation of the field through a paraxial ABCD system.

Collins Integral for Conjugate Planes

In the special case of conjugated planes, the matrix element B vanishes at $B = 0$ and the above formulation of the diffraction integral cannot be applied. But since eq. (18.75) still holds, the integral can be rearranged with the help of $D = 1/A$ in the form

$$E(x_2) = \frac{\sqrt{i}}{\sqrt{\lambda B}} \cdot e^{-\frac{ikCx_2^2}{2A}} \int_{-a}^a E(x_1) \cdot e^{\frac{i\pi A}{\lambda B} (x_1 - \frac{x_2}{A})^2} dx_1. \quad (18-76)$$

Using a representation of Dirac's delta function

$$\delta(x) = \frac{1}{\sqrt{i\pi\epsilon}} e^{-\frac{ix^2}{\epsilon}} \Big|_{\epsilon \rightarrow 0} \quad (18-77)$$

gives for the limiting case $B \rightarrow 0$ the equation for conjugate planes

$$E(x_2) = \frac{1}{A} \cdot e^{-\frac{i\pi C}{\lambda A} x_2^2} E(x_1) \Big|_{x_1 = \frac{x_2}{A}}. \quad (18-78)$$

It can be seen from this expression, that the transverse field is scaled but conserves its profile. But an additional quadratic phase is superposed on the complex field.

However, in real systems, there is always a limiting size of the components and the aperture cone has a finite angle, truncating the spectrum. The above form of the Fresnel integral cannot describe these effects. Because of its paraxial nature, transverse truncation effects are not incorporated. To do this, the system has to be divided into two subsystems with a limiting aperture stop of radius a inside. The paraxial matrix can be written as the product of four factors in the form

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} 1 & z_2 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ -1/f_1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ -1/f_2 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & z_1 \\ 0 & 1 \end{pmatrix}. \quad (18-79)$$

This composition can be interpreted as a free-space transition, a thin lens with focal length f_1 , a second lens with f_2 and an additional transition. Figure 18-17 shows the corresponding setup. The optical effect of the modified system is equivalent to the ABCD segment, if the relations

$$f_1 = z_1 = \frac{D-1}{C}, \quad (18-80)$$

$$f_2 = z_2 = \frac{A-1}{C} \quad (18-81)$$

are fulfilled. The calculation of the field propagation can now be described by two successive applications of the diffraction integral in the Fraunhofer approximation

$$E_p(x_p) = \int E_1(x_1) \cdot e^{-\frac{ik}{z_1}x_1 \cdot x_p} dx_1, \quad (18-82)$$

$$E_2(x_2) = \int E_p(x_p) \cdot e^{-\frac{ik}{z_2}x_2 \cdot x_p} dx_p, \quad (18-83)$$

and a clipping of the field E_p components outside the opening of the aperture between the two lenses for $|x_p| > a$.

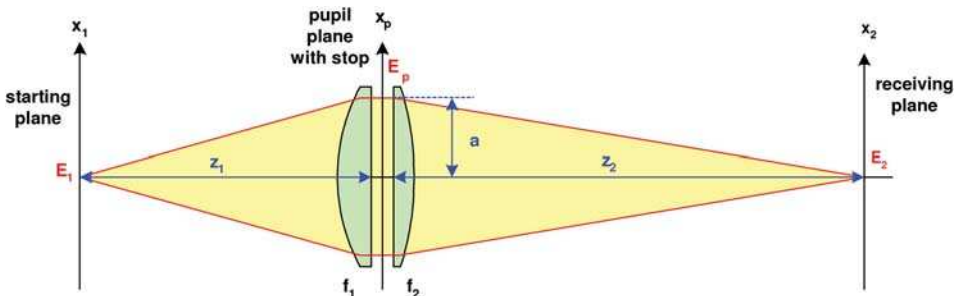


Figure 18-17: Propagation of the field through a paraxial ABCD system between conjugate planes, taking the finite aperture into account.

Collins Integral in the General Case

The extension of the above Collins integral (18.74) to two orthogonal transverse dimensions x and y is straightforward.

A more general paraxial three-dimensional system is described by a 4×4 matrix of the form

$$\vec{r}' = \mathbf{M} \cdot \vec{r} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix} \cdot \vec{r} = \begin{pmatrix} A_{xx} & A_{xy} & B_{xx} & B_{xy} \\ A_{yx} & A_{yy} & B_{yx} & B_{yy} \\ C_{xx} & C_{xy} & D_{xx} & D_{xy} \\ C_{yx} & C_{yy} & D_{yx} & D_{yy} \end{pmatrix} \cdot \vec{r} . \quad (18-84)$$

The optical path length is given by [18-15]

$$L = L_0 + \frac{1}{2} \vec{r}'^T \cdot \underline{\mathbf{R}} \cdot \vec{r} \quad (18-85)$$

with the combined position vector

$$\vec{r} = \begin{pmatrix} \vec{r}_1 \\ \vec{r}_2 \end{pmatrix} \quad (18-86)$$

and the matrix of the eikonal

$$\mathbf{R} = \begin{pmatrix} n_1 \mathbf{B}^{-1} \mathbf{A} & -n_1 \mathbf{B}^{-1} \\ n_2 (\mathbf{C} - \mathbf{D} \mathbf{B}^{-1} \mathbf{A}) & n_2 \mathbf{D} \mathbf{B}^{-1} \end{pmatrix} . \quad (18-87)$$

The Collins integral then takes the general form

$$\begin{aligned} E(\vec{r}_2) &= \frac{i e^{ikL_0}}{\lambda \sqrt{|\mathbf{B}|}} \iint E(\vec{r}_1) \cdot e^{-\frac{i\pi}{\lambda} \vec{r}_1^T \cdot \mathbf{R} \cdot \vec{r}_2} d\vec{r}_1 \\ &= \frac{i e^{ikL_0}}{\lambda \sqrt{|\mathbf{B}|}} \iint E(\vec{r}_1) \cdot e^{-\frac{i\pi}{\lambda} (\vec{r}_1 \mathbf{B}^{-1} \mathbf{A} \vec{r}_1 - 2\vec{r}_1 \mathbf{B}^{-1} \vec{r}_2 + \vec{r}_2 \mathbf{D} \mathbf{B}^{-1} \vec{r}_2)} d\vec{r}_1 \end{aligned} \quad (18-88)$$

with the determinant of the submatrix

$$|\mathbf{B}| = \det \mathbf{B} = B_{xx} B_{yy} - B_{xy} B_{yx} . \quad (18-89)$$

The application of this general form of the Collins integral is necessary in non-orthogonal systems with a coupling of the x and y -field components by rotated components or centering errors.

18.9.2

Example

As a concrete example of this calculation procedure, figure 18-18 shows the optical system of the phase space analyser [18-16], where the coherent field of a slit is propagated through a special system. A lens and a toroidal lens which is rotated by 45°

around the optical axis are located at appropriate distances. In the image plane of the slit, wave aberrations in the slit plane are transformed by the non-orthogonal system in transverse elongations. By analysing the deformation of the slit, this allows the determination of the wave aberration by geometrical means. Figure 18-19 shows the deformation of the slit at various distances z for a spherical aberration of $\lambda/2$. The changing form of the slit image and the effect of defocussing can be seen.

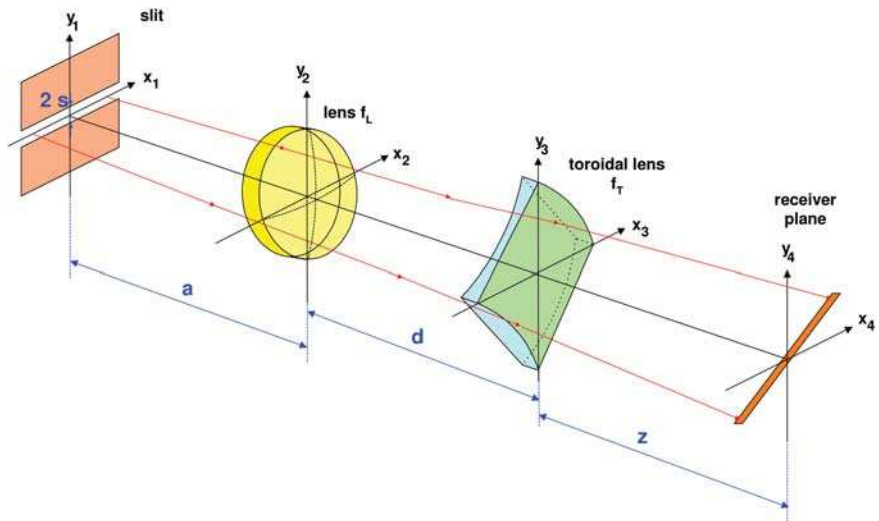


Figure 18-18: Propagation of the coherent field of a slit with spherical aberration through the paraxial non-orthogonal ABCD system of the phase-space analyser.

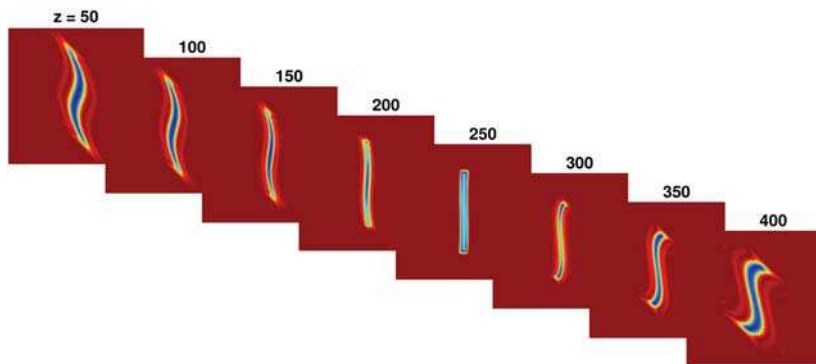


Figure 18-19: Deformation of the slit form, in the phase-space analyser, as a function of the distance z for a spherical aberration $\lambda/2$. The correct imaging condition lies at the location $z=210$.

18.10 Fraunhofer Diffraction

Neglecting terms that are quadratic in (x', y') in the Fresnel integral, i.e.,

$$|\vec{r} - \vec{r}'| \approx z + \frac{x^2 + y^2}{2z} - \frac{xx' + yy'}{z} \quad (18-90)$$

yields the Fraunhofer diffraction integral

$$U(\vec{r}) = \frac{ie^{ikz}}{\lambda|z|} e^{ik\frac{x^2+y^2}{2z}} \iint_A U(\vec{r}') e^{-ik\frac{xx'+yy'}{z}} dx' dy' \quad (18-91)$$

which is proportional to the Fourier transform of the field

$$U(\vec{r}) = \frac{ie^{ikz}}{\lambda|z|} e^{ik\frac{x^2+y^2}{2z}} F\{U(\vec{r}')\}_{v_x=\frac{x}{\lambda z}, v_y=\frac{y}{\lambda z}} \quad (18-92)$$

The intensity of the Fraunhofer diffraction pattern is equal to the squared magnitude of the field in the aperture plane.

Some examples of Fraunhofer diffraction patterns are shown in figure 18-20. To make the plots appear to the human eye (which is basically a logarithmic detector) as a diffraction pattern, the squared magnitude of the Fourier transform of the aperture is scaled to a single decade and the logarithm of the resulting quantity is plotted.

The range of validity of the Fraunhofer approximation is given by the magnitude of the neglected quadratic element in the expansion eq. (18.90). Its contribution to the exponential must be negligible, i.e.,

$$k_0 \frac{\max(x'^2 + y'^2)}{2z} \ll 1. \quad (18-93)$$

With $D = \max(x'^2 + y'^2)$ denoting the largest size of the diffracting aperture we can define the Fresnel number

$$N_F = \frac{\pi D^2}{\lambda z} \ll 1. \quad (18-94)$$

For $N_F \ll 1$ we are in the region of Fraunhofer diffraction. For larger values Fresnel diffraction has to be applied. The development of the Fresnel number with increasing distance for three aperture sizes is shown in figure 18-21.

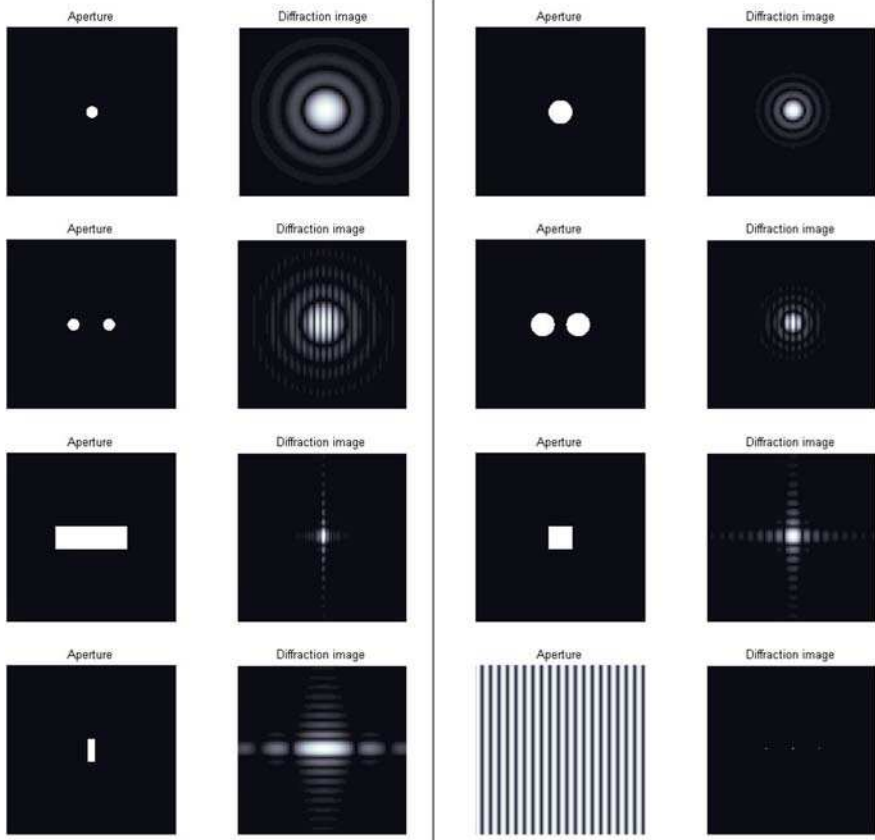


Figure 18-20: Examples of Fraunhofer diffraction patterns.

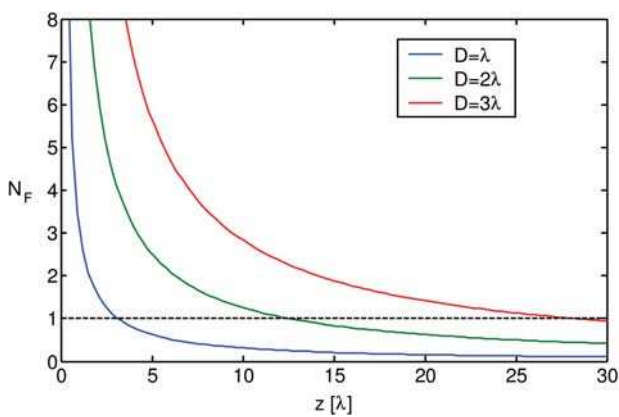


Figure 18-21: Development of Fresnel number N_F with increasing distance z for aperture sizes $D = \lambda, 2\lambda, 3\lambda$.

18.11 Grating Diffraction

From the point of view of diffraction theory, a grating is just a periodic object. Because of its periodicity its diffraction spectrum is discrete. If the size of the grating is limited the diffraction spectrum becomes more continuous. Designs, shapes and optical applications of gratings were presented in detail in chapter 14. Here we discuss the properties of gratings in terms of Fraunhofer diffraction.

18.11.1

Ronchi Grating

The most simple and frequently applied grating is an amplitude grating with an aspect ratio of 1:1 for transparent and opaque lines. It is called a Ronchi grating (section 14.6.5) in honour of the Italian physicist Ronchi. He determined the optical quality of optical instruments by inserting a line grating into the focus plane (the Ronchi-test, section 16.10.4)). The diffraction orders overlap, interfere and produce an interference pattern. From the interference pattern the quality of the wavefront can be deduced as will be discussed elsewhere. Here we discuss the Fourier transform of gratings which are important objects in optical imaging.

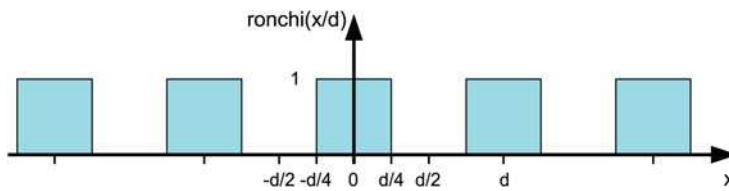


Figure 18-22: Schematic plot of a Ronchi grating.

The Ronchi grating is sketched in figure 18-22. Its mathematical form is

$$\text{ronchi}_A\left(\frac{x}{d}\right) = \begin{cases} 1 & |x - m \cdot d| < \frac{d}{2} \\ 0 & \text{else} . \end{cases} \quad (18-95)$$

Being a periodic function the Ronchi grating can be expanded into a Fourier series

$$\text{ronchi}_A\left(\frac{x}{d}\right) = \sum_m c_m \cdot e^{2\pi i \frac{m}{d} x} \quad (18-96)$$

with the Fourier coefficients

$$c_m = \frac{1}{d} \int_{-d/2}^{d/2} \text{ronchi}_A\left(\frac{x}{d}\right) \cdot e^{-2\pi i \frac{m}{d} x} dx = \frac{1}{d} \int_{-d/4}^{d/4} e^{-2\pi i \frac{m}{d} x} dx = \frac{1}{2} \text{sinc}\left(\frac{m}{2}\right) . \quad (18-97)$$

Inserting (18.97) into (18.96) we obtain the Fourier series as

$$\text{ronchi}_A\left(\frac{x}{d}\right) = \frac{1}{2} \sum_m \text{sinc}\left(\frac{m}{2}\right) \cdot e^{2\pi i \frac{m}{d} x} = \frac{1}{2} + \sum_{m>1} \text{sinc}\left(\frac{m}{2}\right) \cdot \cos\left(2\pi \frac{m}{d} x\right). \quad (18-98)$$

Finally, the sum can be computed yielding

$$\text{ronchi}_A\left(\frac{x}{d}\right) = \frac{1}{2} + \frac{2}{\pi} \cos\left(2\pi \frac{x}{d}\right) - \frac{2}{3\pi} \cos\left(2\pi \frac{3x}{d}\right) + \frac{2}{5\pi} \cos\left(2\pi \frac{5x}{d}\right) - \dots \quad (18-99)$$

Figure 18-23 shows the synthesis of the Ronchi grating by the cumulative sum of the Fourier coefficients up to seventh order. While the overall shape is reproduced more accurately with an increasing number of Fourier coefficients, the overshoot at the edges does not converge to a limit of 1 but to a limit of 1.179. This is called the Gibbs phenomenon after the American physicist J. Willard Gibbs who published a note on it in 1898 [18-17]. A discussion of the Gibbs phenomenon and a review of the literature about it can be found in [18-18]. The Gibbs phenomenon is a general type of behavior of the Fourier transform of discontinuous functions. Although the overshoot persists for any finite number of Fourier coefficients, the approximation of the discontinuous function in a least-squares sense, becomes increasingly accurate. The reason for this is that the period of the oscillations increases with the number of coefficients (this can be seen clearly in figure 18-23).

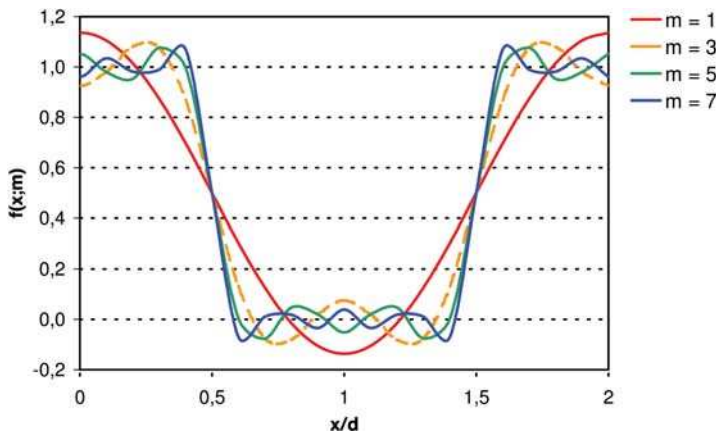


Figure 18-23: Contribution of the lowest order Fourier coefficients to the synthesis of the Ronchi grating. The plot shows the cumulative sum up to the first, third, fifth and seventh order.

An alternative and more versatile way of computing the Fourier transform of the Ronchi grating is based on its representation as a convolution of a grid of delta-pulses (comb-function) with a rect function. Furthermore, we can represent the finite extent of the grid if we multiply by a second rect function (figure 18-24 top)

$$\text{ronchi}_A\left(\frac{x}{d}\right) = \text{rect}\left(\frac{x}{d/2}\right) \otimes \text{comb}\left(\frac{x}{d}\right) \cdot \text{rect}\left(\frac{x}{D}\right). \quad (18-100)$$

The Fourier transform becomes (18-24 bottom)

$$F\left[\text{ronchi}_A\left(\frac{x}{d}\right)\right] = D \frac{d^2}{2} \left[\text{sinc}\left(\frac{d}{2} \nu\right) \cdot \text{comb}(d \cdot \nu) \right] \otimes \text{sinc}(D\nu). \quad (18-101)$$

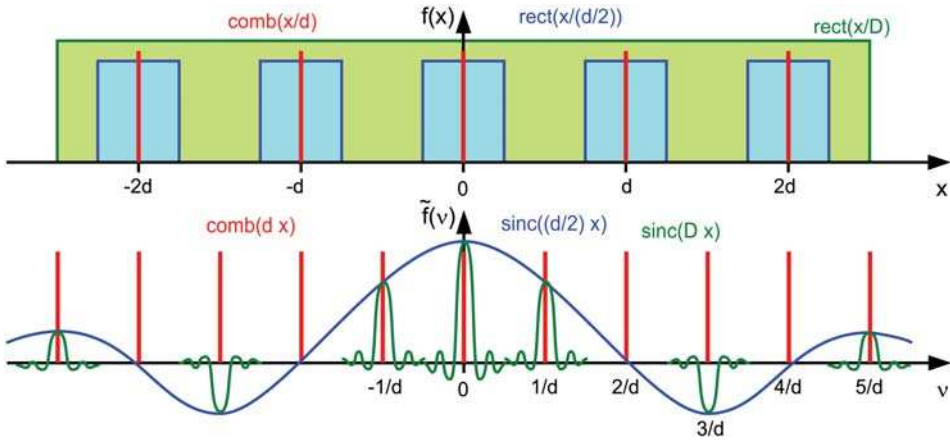


Figure 18-24: The finite Ronchi grating (top) and its frequency spectrum (bottom).

According to the convolution theorem of Fourier optics, the Fourier transform of the finite Ronchi grating consists of the convolution of the delta grid (red in figure 18-24) with the Fourier transform of the large rect function (sinc, green). The result is multiplied by the Fourier transform of the rect function of a single period (sinc, blue).

The computation of the Fourier transform of a pi-phaseshifting Ronchi grating is now straightforward. With the phase Ronchi grating

$$\text{ronchi}_P\left(\frac{x}{d}\right) = \begin{cases} -1 & |x - m \cdot d| < \frac{d}{2} \\ 1 & \text{else} \end{cases}. \quad (18-102)$$

the relation to the amplitude Ronchi grating becomes

$$\text{ronchi}_P(x) = 2 \cdot \text{ronchi}_A(x/P) - 1. \quad (18-103)$$

The spectrum follows from the linearity of the Fourier integral:

$$c_m^P = 2c_m^A - \delta_{m,0} = \text{sinc}\left(\frac{m}{2}\right) - \delta_{m,0}. \quad (18-104)$$

The diffraction efficiency η_m for the grating diffraction is given by the square of the coefficients of the series expansion with $\eta_m = |c_m|^2$. For the amplitude grating half of the incident power is absorbed by the grating and thus will be lost. In total the diffraction efficiency of the amplitude grating is therefore only 50% of the phase grating. The table 18-1 compares the diffraction efficiencies with the series expansion coefficients for amplitude and phase gratings of the Ronchi type.

Table 18-1: Diffraction efficiencies of amplitude and phase gratings.

	Order m	0	1	2	3
Amplitude	Amplitude c_m	0.5	$1/\pi = 0.32$	0.0	$-1/3\pi = -0.11$
	Efficiency η_m	25 %	10 %	0 %	1 %
Phase	Amplitude c_m	0.0	0.64	0.0	-0.22
	Efficiency η_m	0 %	41 %	0 %	5 %

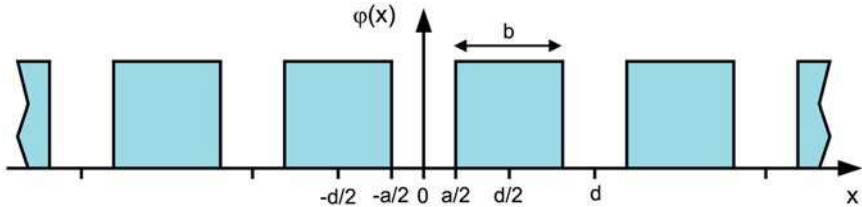


Figure 18-25: The generalized binary phase grating.

The diffraction orders, i.e., the direction of diffracted light is determined by the grating period, while, in the scalar approximation, the diffraction efficiency of a grating is solely determined by the shape of the period. In the simplest case of a binary grating, considered as a generalization of the Ronchi grating, the grating period is formed by a single phase step of width a and amplitude φ . The phase function of the grating is given by:

$$\varphi(x) = \begin{cases} 0 & |x - m \cdot d| < \frac{a}{2} \\ \varphi & \text{else} \end{cases} \tag{18-105}$$

The coefficients of the Fourier series expansion are thus given by:

$$\begin{aligned} c_m &= \frac{1}{d} \cdot \left[\int_{-d/2}^{-a/2} e^{i\varphi} e^{-2\pi i \frac{m}{d} x} dx + \int_{-a/2}^{a/2} e^{i\varphi} e^{-2\pi i \frac{m}{d} x} dx + \int_{a/2}^{d/2} e^{-2\pi i \frac{m}{d} x} dx \right] = \\ &= e^{i\varphi} \text{sinc}(m) + (1 - e^{i\varphi}) \frac{a}{d} \text{sinc}\left(m \frac{a}{d}\right) . \end{aligned} \tag{18-106}$$

For the zero'th diffraction order it follows that

$$c_0 = e^{i\varphi} + (1 - e^{i\varphi}) \frac{a}{d} = \frac{a}{d} + e^{i\varphi} \frac{b}{d} . \tag{18-107}$$

The condition for a vanishing zero diffraction order is therefore

$$e^{i\varphi} = \frac{-a}{b}. \quad (18-108)$$

With a and b both real eq. (18.108) can only be solved for phase differences φ of multiples of π . The zero order thus only vanishes, for example, for gratings with phase differences of $\varphi = \pi$ and $a = b$ (an ordinary Ronchi grating). In this special case it follows from eq. (18.106)

$$c_m = \operatorname{sinc}\left(\frac{m}{2}\right) - \operatorname{sinc}(m). \quad (18-109)$$

The diffraction efficiency of the m^{th} diffraction order follows after squaring and transformation from eq. (18.106) with

$$\begin{aligned} \eta_m &= \left| e^{i\varphi} \operatorname{sinc}(m) + (1 - e^{i\varphi}) \frac{a}{d} \operatorname{sinc}\left(m \frac{a}{d}\right) \right|^2 \\ &= \operatorname{sinc}^2(m) + 2 \frac{a}{d} \operatorname{sinc}\left(m \frac{a}{d}\right) \left[\operatorname{sinc}(m) - \frac{a}{d} \operatorname{sinc}\left(m \frac{a}{d}\right) \right] (\cos \varphi - 1). \end{aligned} \quad (18-110)$$

For $m = 0$, eq. (18.110) can be written as

$$\eta_0 = 1 - 2 \frac{a}{d} \left[1 - \frac{a}{d} \right] \cdot (1 - \cos \varphi) = 1 - 2 \frac{ab}{d^2} \cdot (1 - \cos \varphi). \quad (18-111)$$

In an analogous way, for $m \neq 0$ it follows that

$$\eta_m = 2 \left(\frac{a}{d} \right)^2 \operatorname{sinc}^2\left(m \frac{a}{d}\right) \cdot (1 - \cos \varphi) = A_m \cdot (1 - \cos \varphi). \quad (18-112)$$

A small change in the phase step by $\delta\varphi$ causes a change in the diffraction efficiency given by

$$\begin{aligned} \delta\eta_m &= A_m \cdot [1 - \cos(\varphi + \delta\varphi) - 1 + \cos \varphi] \\ &= A_m \cdot [\cos \varphi \cdot (1 - \cos \delta\varphi) + \sin \varphi \sin \delta\varphi] \approx A_m \cdot \left[\cos \varphi \cdot \frac{\delta\varphi^2}{2} + \sin \varphi \delta\varphi \right]. \end{aligned} \quad (18-113)$$

For typical phase gratings with $\varphi = \pi$, the diffraction efficiency of the m^{th} diffraction order with $|m| \neq 0$ is reduced by an amount proportional to the square of a phase error $\delta\varphi$, while the zero order increases by an amount proportional to $\delta\varphi^2$. Similarly, for deviations of the grating aspect ratio from an ideal value of $a = b = d/2$, the diffraction efficiency of the zeroth order increases by an amount proportional to the square of the phase step position error δx .

For a general treatment of binary and multilevel phase gratings the reader is referred to [18-19].

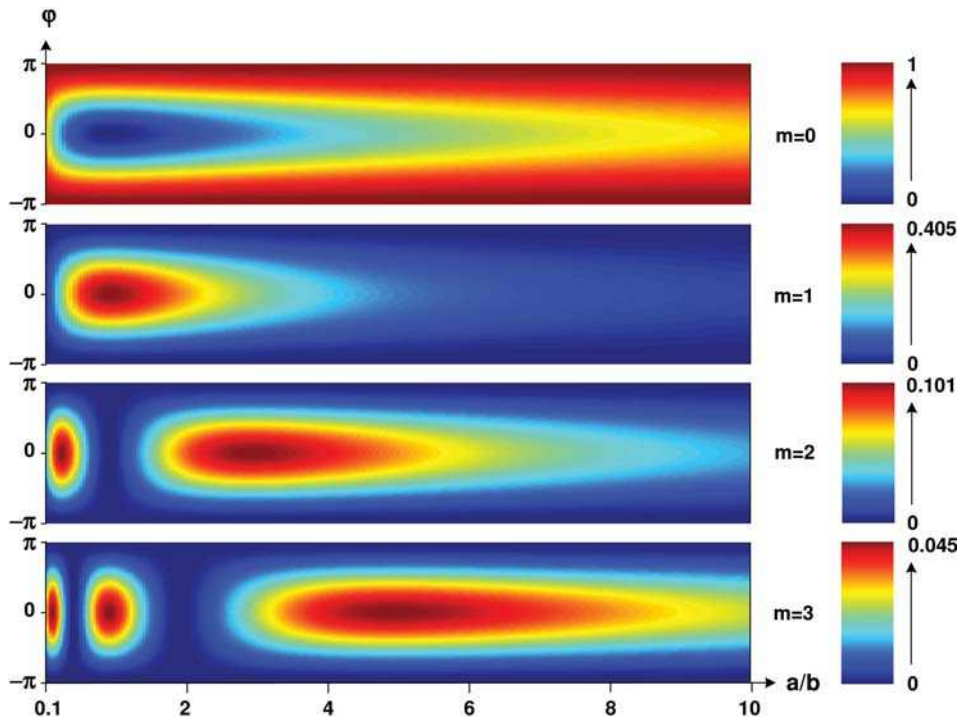


Figure 18-26: Diffraction efficiency of the first diffraction orders as a function of aspect ratio a/b and phase difference φ .

18.11.2

The Sinusoidal Phase Grating and Surface Fabrication Errors

The Fourier transform of an amplitude grating, with a transmission function given by the elementary harmonic functions cosine and sine, is evidently given by two delta-functions at the grating frequencies with

$$\sin\left(2\pi\frac{x}{d}\right) = \frac{1}{2i}\left(e^{i2\pi\frac{x}{d}} - e^{-i2\pi\frac{x}{d}}\right) \Leftrightarrow \frac{1}{2i}\left[\delta\left(\frac{1}{d}\right) - \delta\left(-\frac{1}{d}\right)\right], \quad (18-114)$$

$$\cos\left(2\pi\frac{x}{d}\right) = \frac{1}{2}\left(e^{i2\pi\frac{x}{d}} + e^{-i2\pi\frac{x}{d}}\right) \Leftrightarrow \frac{1}{2}\left[\delta\left(\frac{1}{d}\right) + \delta\left(-\frac{1}{d}\right)\right]. \quad (18-115)$$

The phase of a grating given by the sine-function with period d is

$$\varphi(x) = A \cdot \sin\left(\frac{2\pi}{d} \cdot x\right). \quad (18-116)$$

The Fourier transform of the sinusoidal phase grating is then given by

$$\begin{aligned}
 u(\nu) &= \int e^{iA \sin\left(\frac{2\pi}{d}x\right)} e^{-i2\pi\nu x} dx = \sum_m J_m(A) \int e^{im\frac{2\pi}{d}x} \cdot e^{-i2\pi\nu x} dx \\
 &= \sum_m J_m(A) \cdot \delta\left(\nu - m \cdot \frac{2\pi}{d}\right)
 \end{aligned}
 \tag{18-117}$$

where the Jacobi identity was applied:

$$e^{ix \sin \phi} = \sum_m J_m(x) \cdot e^{im\phi}.
 \tag{18-118}$$

The Bessel function is given by the well-known series expansion

$$J_m(x) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k! (m+k)!} \cdot \left(\frac{x}{2}\right)^{2k+m} = \frac{1}{m!} \left(\frac{x}{2}\right)^m - \frac{1}{(m+1)!} \cdot \left(\frac{x}{2}\right)^{2+m} + \dots
 \tag{18-119}$$

which can be approximated for small arguments as

$$J_m(x) = \frac{1}{m!} \left(\frac{x}{2}\right)^{|m|}.
 \tag{18-120}$$

Table 18-2: Small argument approximations for Bessel functions of order $m=0 \dots 3$.

m	0	1	2	3
$A \rightarrow 0$	$J_0 = 1$	$J_1(A) = \frac{A}{2}$	$J_2(A) = \frac{A^2}{8}$	$J_3(z) = \frac{A^3}{48}$

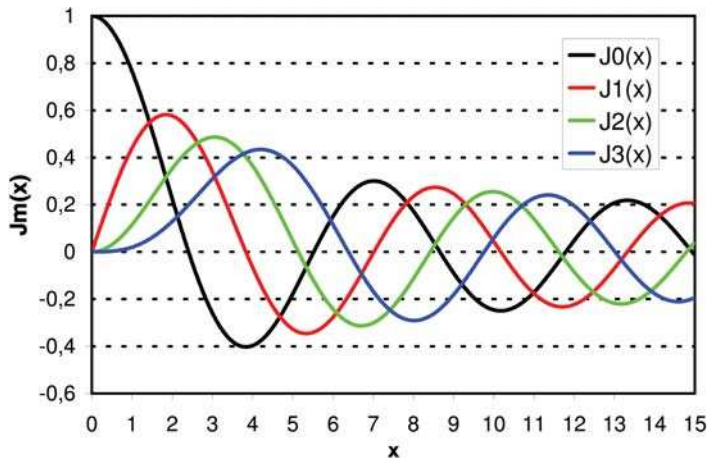


Figure 18-27: Bessel functions of orders $m=0 \dots 3$.

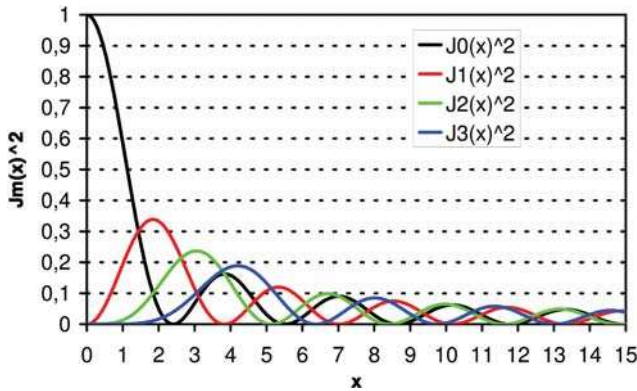


Figure 18-28: Square of Bessel functions of orders $m=0 \dots 3$.

The diffraction efficiency of the m^{th} diffraction order is thus given by

$$\eta_m = J_m^2(A). \quad (18-121)$$

Table 18-3 gives the values of the diffraction efficiencies of the phase grating with sinusoidal phase profile for diffraction orders $m=0 \dots 3$. The total diffracted light is approximated by the sum over the diffraction efficiencies under consideration of the \pm diffraction order, giving the total integrated scatter (TIS). From table 18-3 it can be seen that the simple scalar analysis according to the Fraunhofer approximation, gives considerably accurate results only for low amplitudes A of the sinusoidal-phase grating. With increasing phase differences of the amplitude A the total integrated scatter deviates rapidly from 100%. It is also remarkable that for the sinusoidal-phase grating it is not possible to optimize the depth for only one diffraction order. At a maximum efficiency of 33.8%, for the first diffraction orders at $A \approx 1.846$, about 10% is still diffracted into each of the zeroth and second orders.

Table 18-3: Maxima and minima of the diffraction efficiencies of the first orders of a sinusoidal grating.

A	$J_0^2(A)$	$J_1^2(A)$	$J_2^2(A)$	$J_3^2(A)$	TIS(O3)
0.000	1.0000	0	0	0	1.00
1.846	0.0982	0.3386	0.1006	0.0111	1.00
2.402	0	0.2701	0.1861	0.0394	0.99
3.047	0.0759	0.1033	0.2367	0.1007	0.96
3.838	0.1622	0	0.1612	0.1772	0.84
4.189	0.1429	0.0182	0.0983	0.1887	0.75
5.127	0.0183	0.1150	0	0.1165	0.48
5.332	0.0042	0.1198	0.0042	0.0884	0.43
5.508	0	0.1163	0.0143	0.0646	0.39

Sinusoidal-phase gratings are frequently applied to illustrate the effect of small surface errors due to fabrication errors. When only the ± 1 diffraction orders are considered, the total integrated scatter is then given after insertion of eq. (18.120) into eq. (18.121), by the square of the RMS value of the phase error

$$TIS^{\pm 1} = 2 \cdot J_1^2(A) \approx \left(\frac{A}{\sqrt{2}}\right)^2 = RMS^2 \quad (18-122)$$

with the root-mean-square average of the sinusoidal phase error given by

$$RMS = \sqrt{\frac{1}{d} \int_0^d A^2 \sin^2\left(\frac{2\pi}{d}x\right) dx} = \sqrt{\frac{A^2}{2d} \left[x + \frac{d}{4\pi} \sin\left(\frac{4\pi}{d}x\right)\right]_0^d} = \frac{A}{\sqrt{2}}. \quad (18-123)$$

For surface gratings, the phase difference is given by multiplication of the surface $z(x) = a \cdot \sin(2\pi/d \cdot x)$ with the propagation vector $k_z = 2\pi/\lambda \cdot n(x,z) \cdot \cos \alpha_i$ (figure 18-29).

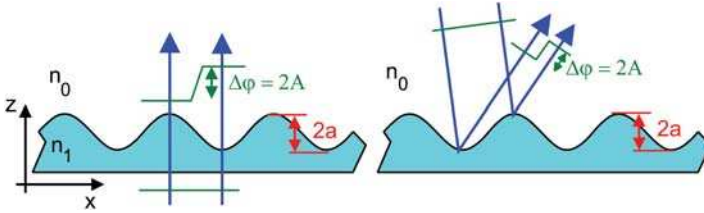


Figure 18-29: Relation between phase and surface error in transmission and reflection.

The transmission amplitude A_t and the reflection amplitude A_r of the sinusoidal phase is therefore related to the amplitude of the surface error by

$$A_t = (n_1 - n_0) \cdot \frac{2\pi}{\lambda} a, \quad (18-124)$$

$$A_r = 2n_0 \cos \alpha_i \cdot \frac{2\pi}{\lambda} a. \quad (18-125)$$

With $n_0 = 1$ and $n_1 - n_0 \approx 0.5$ the well-known TIS formulae for transmission (index t) and reflection (index r) follow:

$$TIS_t = \left[(n - 1) \frac{2\pi}{\lambda} \frac{a}{\sqrt{2}} \right]^2 = \left[\frac{\pi \cdot RMS}{\lambda} \right]^2, \quad (18-126)$$

$$TIS_r = \left[2 \cos \alpha_i \frac{2\pi}{\lambda} \frac{a}{\sqrt{2}} \right]^2 = \left[\frac{4\pi \cdot RMS \cdot \cos(\alpha_{in})}{\lambda} \right]^2. \quad (18-127)$$

18.12 Scalar Diffraction at Dielectric Objects

Incidentally, the scalar diffraction theories were applied to diffraction by apertures in black screens. Of course, the scalar diffraction integrals can also be applied to the propagation of arbitrary wave fields, with the effective aperture size becoming infi-

nite and the spatial modulation being imposed on the incident wave. According to the thin-element-approximation the transmitted field is computed according to the law of geometrical optics as if the phase object is an infinitely thin, phase-shifting sheet (figure 18-30)

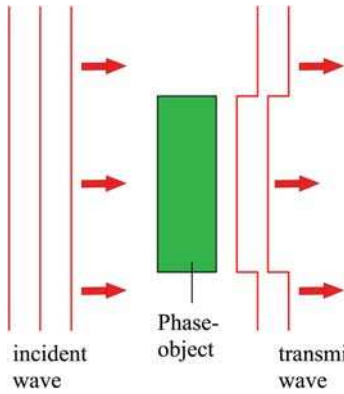


Figure 18-30: Transmission through a phase object: geometrical optics approximation.

$$T(x, y) = e^{ik_0(n(x,y)-1)d(x,y)} \quad (18-128)$$

where the local thickness is $d(x, y)$ and the local refractive index is $n(x, y)$.

The geometrical optic field in the plane of the sheet is conveniently divided into the incident field and a scattered field

$$\begin{aligned} U(x, y) &= \begin{cases} T(x, y) U_i(x, y) & : \text{ behind object} \\ U_i(x, y) & : \text{ else} \end{cases} \\ &= U_i(x, y) + \begin{cases} (T(x, y) - 1) U_i(x, y) & : \text{ behind object} \\ 0 & : \text{ else} \end{cases} \end{aligned} \quad (18-129)$$

Insertion into a scalar diffraction integral yields the diffracted field as the sum of the propagated incident field and a scattered field. Using the first Rayleigh–Sommerfeld integral we obtain, for instance,

$$U(\vec{r}) = U_i(\vec{r}) + \frac{1}{2\pi} \iint_A (T(\vec{r}') - 1) U(\vec{r}') \frac{\partial}{\partial \vec{s}} \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|} d\sigma. \quad (18-130)$$

The transmission according to geometrical optics, however, is only a very rough approximation of the real boundary values and therefore deviations from the actual fields are to be expected. Actually, the incident wave interacts with the object, producing a scattered wave. The resulting wave, that is the diffraction field, is the sum of the incident and the scattered field (cf. section 18.14). Intuitively one expects that the above approximation is actually valid for only thin phase objects ($d \ll \lambda$). It turns out, however, that the validity depends on the diffraction theory used, in combina-

tion with the polarization, the thickness of the sheet and the spatial frequency spectrum of the transmittance [18-20].

In order to compute the diffraction at thick phase objects, the object can be divided into a stack of N infinitely thin, phase-shifting sheets of individual transmittance (figure 18-31)

$$T_j(x, y) = e^{ik_0(n(x,y)-1)d_j(x,y)}. \quad (18-131)$$

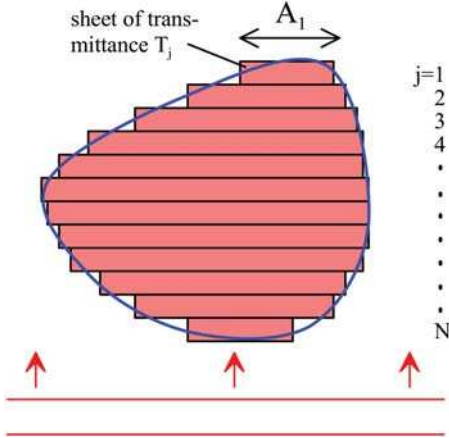


Figure 18-31: Subdivision of a phase object into thin phase-shifting sheets.

In the first Born approximation, multiple scattering is neglected. In this context this means that the diffraction of the undisturbed incident field is considered for each sheet separately, i.e., the primary scattered fields of all sheets are superposed without taking successive diffraction process into account. The diffracted field becomes

$$U(\vec{r}) = U_i(\vec{r}) + \frac{1}{2\pi} \sum_{j=1}^N \iint_{A_j} (T_j(\vec{r}') - 1) U_i(\vec{r}') \frac{\partial}{\partial s} \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|} d\sigma. \quad (18-132)$$

In a better approximation the successive diffraction is computed as in the beam propagation method, i.e., the incident field onto the j th sheet stems from the diffraction at the $(j-1)$ th sheet and the diffraction integral becomes

$$U_j(\vec{r}) = U_i(\vec{r}) + \frac{1}{2\pi} \iint_{A_{j-1}} (T_{j-1}(\vec{r}') - 1) U_{j-1}(\vec{r}') \frac{\partial}{\partial s} \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|} d\sigma, \quad (18-133)$$

$$j = 1, \dots, N, \quad U_0 = U_i.$$

This approximation neglects backward scattering.

For numerical implementation, the propagation operator formalism is particularly convenient because the Fourier transform of the transmitted field is readily com-

puted without having to cope with artificial external boundaries as in the space-evaluation of the scalar diffraction integrals. If the latter is to be applied, the decomposition of the transmitted field into the undisturbed incident wave and a scattered field is to be recommended.

18.13 Babinet's Principle

The far-field Babinet principle indicates the similarity of the Fraunhofer diffraction patterns (except for the central peak) behind complementary opaque obstacles in which transparent and opaque parts are reversed [18-21].

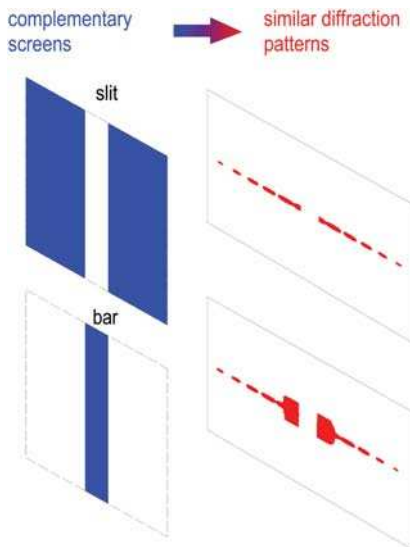


Figure 18-32: Illustration of Babinet's principle from the measurements of D.W. Pohl [18-22].

This original formulation of Babinet's principle is a special case of a more general scalar Babinet principle [18-2] which states that the superposition of the complex scalar diffraction fields $U_A(r)$ and $U_{\bar{A}}(r)$ behind two complementary black screens A and \bar{A} equals the non-diffracted incident wave

$$U_A(\vec{r}) + U_{\bar{A}}(\vec{r}) = U_i(\vec{r}). \quad (18-134)$$

This is a direct consequence of the superposition principle of scalar diffraction (stating that the diffracted field of two disjunctive obstacles is the sum of the individual diffraction fields). Calculating the diffraction field of a weak phase object, however, also implies an undisturbed superposition of the scattered fields of different parts of the object. Multiple scattered fields are neglected. For this reason an extension of Babinet's principle to phase objects is possible as this is valid within the scope of the

first Born approximation. A rigorous form of the scalar Babinet principle is provided by Bouwkamp [18-23] for acoustics. Here complementary objects satisfy different boundary conditions: The complement of a soft screen which satisfies $U = 0$, i.e., the Dirichlet condition is a hard obstacle which satisfies $dU/d\vec{s} = 0$.

The electromagnetic Babinet principle for diffraction fields behind complementary perfectly conducting objects is frequently examined in the literature. A list of early authors dealing with this principle can be found in [18-23]. The electromagnetic Babinet principle relies on the duality of Maxwell's equations [18-24]. In the present chapter, we restrict ourselves to the approximate (i.e., not rigorous) scalar Babinet principle, which can be extended to thick, transparent diffracting objects in a much more straightforward way [18-25].

Consider two complementary phase objects P and \bar{P} as depicted in figure 18-33, where the phase-shifting parts are reversed so that the superposed object has constant transmission.

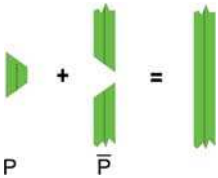


Figure 18-33: Complementary phase objects.

The diffracted field of objects P and \bar{P} are written as the sum of the incident and the respective scattered fields

$$\begin{aligned} U_P(\vec{r}) &= U_i(\vec{r}) + U_P^S(\vec{r}) , \\ U_{\bar{P}}(\vec{r}) &= U_i(\vec{r}) + U_{\bar{P}}^S(\vec{r}) . \end{aligned} \quad (18-135)$$

Due to the superposition

$$\begin{aligned} U_P(\vec{r}) + U_{\bar{P}}(\vec{r}) &= 2U_i(\vec{r}) + (T - 1)U_i(\vec{r}) \\ &= TU_i(\vec{r}) + U_i(\vec{r}) , \end{aligned} \quad (18-136)$$

the superposition of the complex scalar diffraction fields behind two complementary phase objects equals the undiffracted incident field, plus the field TU_i transmitted by the superposed dielectric object, that is, an infinite phase-shifting sheet. The consequences of the extended Babinet principle for far-field diffraction patterns are the same as those of the original one for black screens. The Fraunhofer diffraction patterns are identical except for the central peak.

Since the extended Babinet principle is a consequence of the first Born approximation, it can be applied if the first Born approximation is valid – for instance for weakly absorbing objects. The infinitely thin black screen does not affect the scattered field in its vicinity and is therefore also included ($T = 0$).

Another consequence of this form of Babinet's principle is that complementary objects yield complementary near-fields. For the deviation of the diffraction field

from the incident field, i.e., the amplitude $\Delta a = (|U|/|U_i|) - 1$ and the phase $\Delta\phi = \arg(U) - \arg(U_i)$, the extended Babinet principle becomes

$$[1 + \Delta a(\vec{r})]e^{i\Delta\phi(\vec{r})} + [1 + \overline{\Delta a(\vec{r})}]e^{i\overline{\Delta\phi(\vec{r})}} = T + 1. \quad (18-137)$$

Using $e^{i\Delta\phi} \approx 1 + i\Delta\phi$ we obtain

$$\begin{aligned} \Delta a(\vec{r}) + \overline{\Delta a(\vec{r})} &= |T| - 1, \\ \Delta\phi(\vec{r}) + \overline{\Delta\phi(\vec{r})} &= \arg(T). \end{aligned} \quad (18-138)$$

The principle was verified by near-field measurements with 3 cm microwaves (figure 18.34).

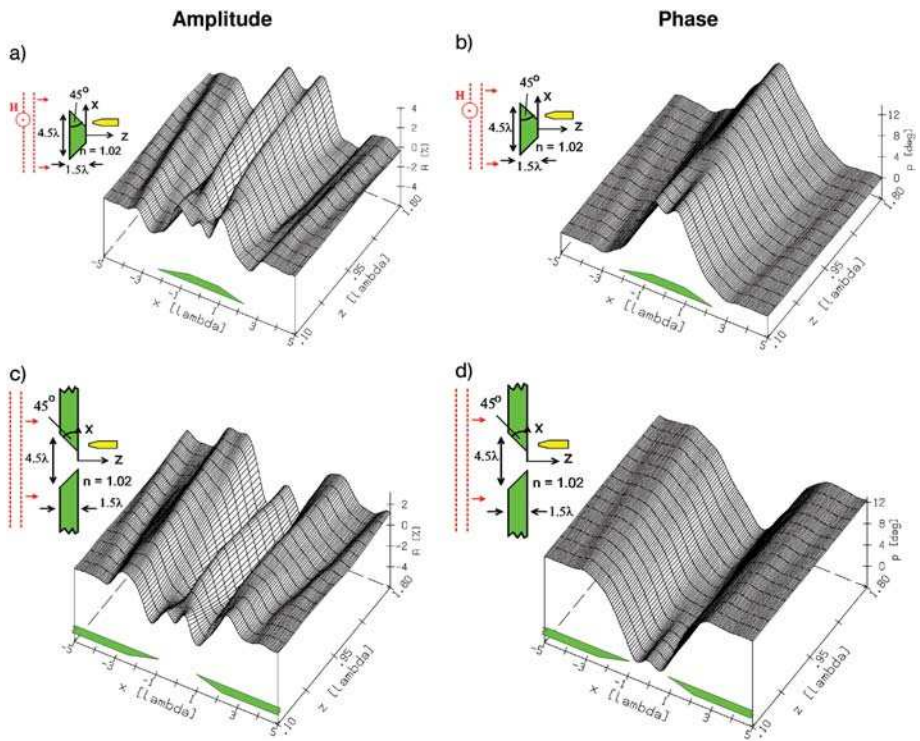


Figure 18-34: Amplitude (a,c) and phase-distribution (b,d) for increasing distance ($z = 0.1\lambda - 1.8\lambda$) behind complementary weak phase objects ($n = 1.02$) measured with 3 cm microwaves [18-25].

18.14 Scalar Scattering

In inhomogeneous media, the refractive index can be split into a constant part and a spatially varying part $\delta n(r)$. In particular, for the case of weakly inhomogeneous media with $n(\vec{r}) = \bar{n} + \delta n(\vec{r})$, the space-dependent terms of the wave equation can be combined and treated as a source term [18-26]:

$$\Delta U(\vec{r}) + \bar{n}^2 k_0^2 U(\vec{r}) = -(2\delta n(\vec{r})\bar{n} + \delta n^2(\vec{r})) k_0^2 U(\vec{r}) = F(\vec{r}) k_0^2 U(\vec{r}) . \quad (18-139)$$

The inhomogeneity $F(r)$ can also be interpreted as the scattering potential. In the case of weak scattering or Rayleigh scattering, the solution of eq. (18.139) is given by the incident field U_i and a weak perturbation, the scattered field U_s

$$U(\vec{r}) = U_i(\vec{r}) + U_s(\vec{r}) \quad (18-140)$$

where the scattered field U_s given by the implicit Fredholm integral equation

$$U_s(\vec{r}) = k_0^2 \iiint_V F(\vec{r}') U(\vec{r}') G(r - \vec{r}') d^3 r' . \quad (18-141)$$

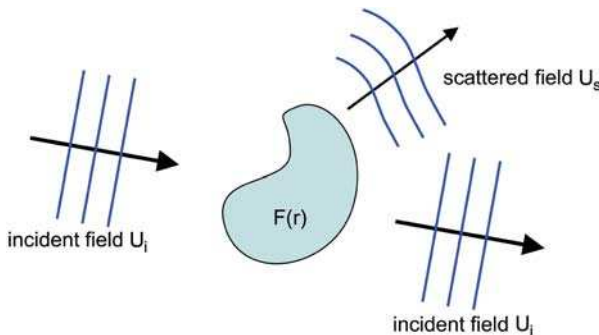


Figure 18-35: Scattering in the first Born approximation.

This is a three-dimensional convolution integral of the source term FU with Green's function. Taking a spherical wave as the Green's function together with the spectrum is, according to eq. (18.51)

$$F \left\{ \frac{e^{ikR}}{R} \right\} = i \frac{2\pi}{k_z} e^{ik_z z} \quad (18-142)$$

which yields the plane wave expansion

$$G(\vec{r} - \vec{r}') = \frac{e^{ik_0|\vec{r}-\vec{r}'|}}{|\vec{r} - \vec{r}'|} = \frac{i}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{e^{i[k_x(x-x') + k_y(y-y') + k_z(z-z')]} }{k_z} dk_x dk_y . \quad (18-143)$$

Inserting the plane wave expansion into the Fredholm equation and performing the 2D Fourier transformation we obtain

$$\begin{aligned} u_s(k_x, k_y, z) &= k_0^2 \iiint F(r) \cdot U_i(r) \cdot \frac{1}{k_z} \cdot e^{-i(k_x x + k_y y + k_z(z-z'))} dx' dy' dz' \\ &= \frac{k_0^2}{(2\pi)^3} \cdot \frac{e^{ik_z z}}{k_z} \iiint f(k - k_i) \cdot u_i(k_i) d^3 k_i. \end{aligned} \quad (18-144)$$

Accordingly, the angular spectrum of the scattered field is given by the convolution of the incident angular spectrum with the three-dimensional Fourier transform of the inhomogeneity $F(r)$:

$$u_s(k_x, k_y, z) = \frac{k_0^2}{k_z} e^{ik_z z} \cdot f(\vec{g}) \otimes u_i(\vec{g}). \quad (18-145)$$

Multiple scattering can be considered by the Born series expansion. In general the $(p+1)^{\text{th}}$ -order scattered field from the p^{th} -order, according to eq. (18-141), is given by

$$U_{p+1}(r) = k_0^2 \cdot \iint F(r') \cdot U_p(r') \cdot G(r' - r) d^2 r'. \quad (18-146)$$

For the case of a single incident plane wave, which is described in the frequency domain by a delta function

$$u_i(\vec{g}) = \delta(\vec{g} - \vec{k}_i), \quad (18-147)$$

eq. (18.145) reduces to

$$u_s(k_x, k_y, k_z = \sqrt{k^2 - k_x^2 - k_y^2}; z) = \frac{i k_0^2}{k_z} e^{ik_z z} \cdot f(\vec{k} - \vec{k}_i). \quad (18-148)$$

A consequence of eq. (18.148) is the Laue equation [18-27], which states that scattering is described by the spectrum of the scattering potential, taken at those object frequencies

$$\vec{g} = \vec{k} - \vec{k}_i \quad (18-149)$$

which correspond to possible frequency transfers from the incident k_i -vector to the Ewald sphere of the scattered k -vector (note that the angular components k of the field distribution are restricted by the Ewald equation) [18-28]–[18-30].

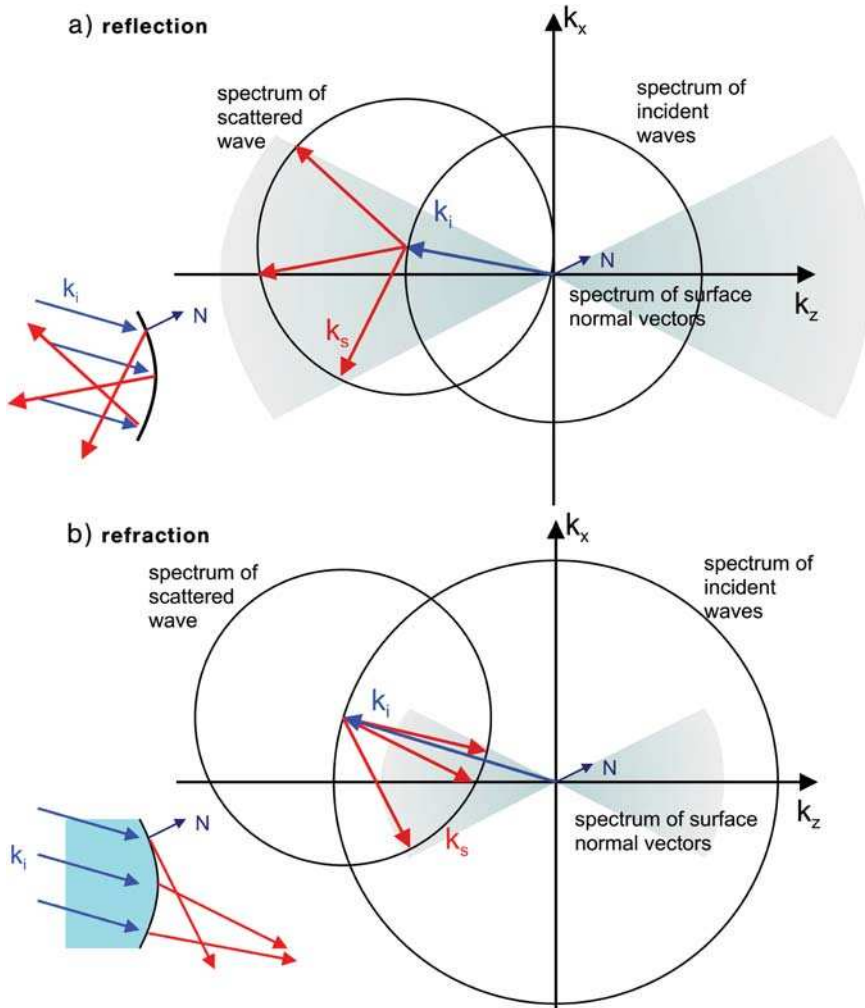


Figure 18-36: a) Laue construction of reflection at an interface.

b) Laue construction of refraction at an interface from $n_0 = 1.5$ to $n_1 = 1.0$.

According to the Born approximation of first order, the solution of eq. (18.141) is approximately given by a first-order series expansion [18-31]. As was pointed out by Tatarski, this first-order approximation is only valid for small phase changes S_1 relative to the phase S_1 of the incident wave U_i [18-32]

$$|S_1| \ll 1 \quad \text{with} \quad U_s = A \cdot e^{i(S_0 + S_1)}. \quad (18-150)$$

For large, but smooth variations of the refractive index as in gradient index media, the Rytov approximation is preferable [18-32].

It has been shown that the Born approximation is also applicable for the approximate treatment of refraction, diffraction and scattering at surfaces. For a surface given by $z=h(x,y)$ the scattering potential $f(r)$ can be written as

$$F(\vec{r}) = (n_0^2 - n_1^2) \cdot \Theta(z - h(x, y)) \quad (18-151)$$

where the Heavyside step function is defined by $\Theta(z) = 1$ for $z > 0$ and $\Theta(z) = 0$ otherwise. The frequency spectrum of the scattering potential is thus given by

$$f(\vec{g}) = C \iint \left[\delta(g_z) + \frac{i}{g_z} \right] \cdot \exp[-ih(x, y)g_z] \cdot e^{-i(g_x x + g_y y)} dx dy. \quad (18-152)$$

The δ -function at $g_z = 0$ will not contribute to the refracted spectrum and can be omitted. Insertion of eq. (18-144) yields

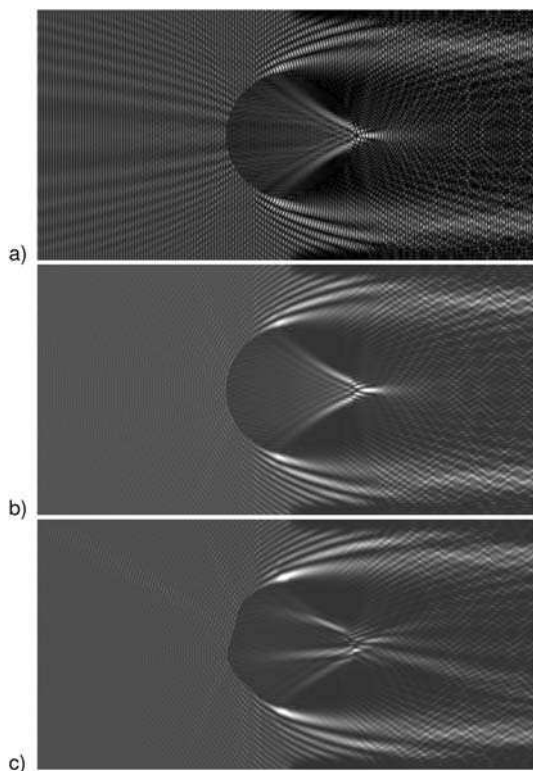


Figure 18-37: a) Amplitude distribution of a plane wave scattered at a dielectric cylinder with index 1.5 and a diameter of 21 wavelengths (image extension 85×42 wavelengths; b) Intensity distribution corresponding to a); c) Intensity distribution of the scattering of a plane wave at a deformed dielectric cylinder with index 1.5 [18-33].

$$U(\vec{r}) = \frac{\chi^2 k_0^2}{2(2\pi)^3} \int_{x', y'} \int_{k_x, k_y} \int_{k_z^i} \frac{e^{-ih(x') \cdot (k_z - k_z^i)}}{k_z (k_z - k_z^i)} \cdot U_i(x', y'; k_z^i) \cdot e^{i[k_x(x-x') + k_y(y-y') + k_z z]} dx' dy' dk_x dk_y dk_z^i . \quad (18-153)$$

Using the Born series expansion, the scattered field at refracting surfaces can be computed in a relatively simple manner especially when no analytical solutions can be found [18-33]. Figure 18-36 shows examples of simulation results at a perfect and at a disturbed cylinder.

For an analytical solution of eq. (18-153), the first exponential can be written as a Fourier series or the interface function $h(x)$ can be expanded into a polynomial. For small angles, the difference between the refracted and incident k_z -vector can be approximated by the difference between the refractive indices Δn multiplied by $2\pi/\lambda$. At $z=0$, the first exponential can be interpreted as the transmission function $T(x)$ of the phase object as given in the Kirchhoff approximation for infinitesimal thin objects:

$$T(x, y) = e^{i\frac{2\pi}{\lambda}\Delta n \cdot h(x, y)} . \quad (18-154)$$

Therefore the Kirchhoff approximation for thin phase gratings is in the paraxial approximation:

$$u(x, y, z = 0) \approx T(x, y) u_i(x, y, z = 0) . \quad (18-155)$$

The Kirchhoff approximation considers only transverse frequency transfers and neglects the longitudinal frequency transfer [18-34].

18.15 Boundary Diffraction Waves

In the theory of boundary diffraction waves, diffraction fields $U(r)$ are divided into two parts. An undisturbed geometrical optics field $U_g(\vec{r})$ and a boundary diffraction wave $U_{BDW}(\vec{r})$ originating from the rim of the diffracting object (figure 18-38). For scalar waves it is

$$U(\vec{r}) = U_g(\vec{r}) + U_{BDW}(\vec{r}) \quad (18-156)$$

with

$$U_{BDW}(\vec{r}) = \int_C K(\vec{r}, \vec{r}', U_i(\vec{r}')) ds . \quad (18-157)$$

The boundary diffraction wave is discontinuous when compensating for the discontinuity of the geometrical optics field.

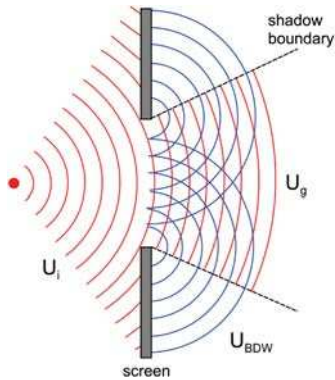


Figure 18-38: Diffraction with a boundary diffraction wave.

The idea of calculating diffraction fields using a boundary diffraction wave dates back to Young (1773–1829) [18-35]. The idea was later revived by Maggi [18-36] and by Rubinowicz [18-37], who showed that the Kirchhoff diffraction integral splits into two parts, namely a geometrical optics field and a boundary diffraction wave. Its existence in the Rayleigh–Sommerfeld diffraction integrals was confirmed by Marchand and Wolf [18-38].

The boundary diffraction wave is also the central concept in the geometrical diffraction theory (GTD) of Keller [18-39] – which becomes discontinuous at the shadow boundary – and its successors, the uniform asymptotic theory (UAT) [18-40] and the uniform theory of diffraction (UTD) [18-41]. However, the best known example of a boundary diffraction wave is definitely Sommerfeld’s exact solution of the diffraction by a perfectly conducting half-plane [18-2]. The geometrical theory of diffraction is of interest because the fields are electromagnetic but the mathematical framework is nevertheless comparatively simple. So it can be used for quick estimations of near-fields [18-45]. For this reason, we discuss them in the present chapter, which is concerned with scalar diffraction, although they are, strictly speaking, electromagnetic.

18.15.1

Geometrical Theory of Diffraction

The Geometrical Theory of Diffraction (GTD) developed by J.B. Keller [18-39] applies the concept of optical rays to the calculation of diffraction fields. In addition to the geometrical optics rays, diffracted rays are introduced which emanate from interfaces, edges and vertices and behave and propagate inside a homogeneous medium in the same way as ordinary rays.

Applying Fermat’s principle to edge diffraction, it can be shown that the diffracted rays in conjunction with the reflected rays, represent the main part of the scattered field, since all other rays interfere destructively (by the method of stationary phase). Hence in this high-frequency approximation, the scattering process is treated as a local phenomenon. Summing up, the diffraction field is divided into a (discontinuous) geometrical optics field and an (also discontinuous) boundary diffraction wave emanating from the edges.

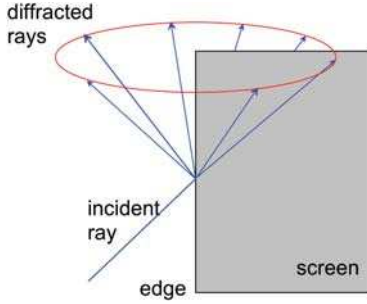


Figure 18-39: Diffracted rays.

Consider a perfectly conducting half-plane. The electromagnetic field behind it is described as the superposition of a geometrical optics field (E_g, H_g) and a boundary field (E_{BDW}, H_{BDW}), i.e., by analogy with (18.156)

$$\begin{aligned}\vec{E}(\vec{r}) &= \vec{E}_g(\vec{r}) + \vec{E}_{BDW}(\vec{r}), \\ \vec{H}(\vec{r}) &= \vec{H}_g(\vec{r}) + \vec{H}_{BDW}(\vec{r}).\end{aligned}\quad (18-158)$$

According to the geometry shown in figure 18-39, the edge of the half-plane coincides with the y -axis and the electromagnetic field will be constant in this direction (two-dimensional diffraction).

The boundary diffraction wave is identical for the GTD and the UAT. It is discontinuous at the reflection and transmission shadow boundary. The geometrical optics field is

$$\vec{E}_g(\vec{r}) = \begin{cases} 0 & : \text{shadow region} \\ \vec{E}_i(\vec{r}) & : \text{lit region,} \end{cases}\quad (18-159)$$

where $\vec{E}_i(\vec{r})$ denotes the incident field. While in the GTD the geometrical optics field remains unchanged, in the UAT it is modified for correction of the GTD singularity at the shadow boundary:

$$\vec{E}(\vec{r}) = \vec{E}_{mg}(\vec{r}) + \vec{E}_{BDW}(\vec{r}).\quad (18-160)$$

The modified geometrical optics field can be written as the sum of two fields emanating from the real source and a virtual source [18-42]

$$E_{mg}^y(\vec{r}) = \Delta F(\zeta_i) E_i^y - \Delta F(\zeta_r) E_r^y\quad (18-161)$$

with

$$\Delta F(\zeta) = F(\zeta) - \hat{F}(\zeta).\quad (18-162)$$

$F(\zeta)$ is the Fresnel function and $\hat{F}(\zeta)$ is the asymptotic Fresnel function with

$$F(\zeta) = \frac{e^{-i\pi/4}}{\sqrt{\pi}} \int_{\zeta}^{\infty} e^{it^2} dt, \tag{18-163}$$

$$\hat{F}(\zeta) = \frac{1}{2\zeta\sqrt{\pi}} e^{i(\zeta^2 + \pi/4)} .$$

$E_i(r)$ in (18.159) denotes the incident field at the point of observation and $E_s(r)$ denotes the field emanating from the virtual source that corresponds to an image source.

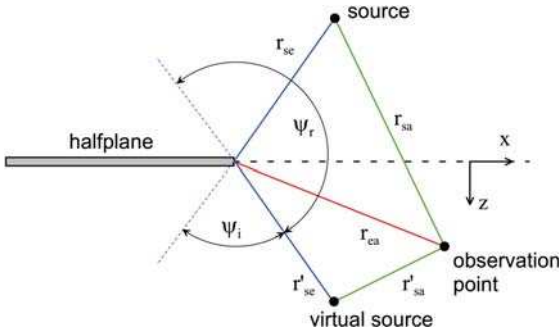


Figure 18-40: Geometry for UAT computations.

The geometrical arrangement is shown in figure 18-40. The virtual source has the same distance to the object plane as the physical source. r_{sa} denotes the distance between the source and the point of observation, and r_{se} denotes the distance between the source and the edge of the half-plane. The corresponding primed quantities have the same meaning but they refer to the virtual source.

The detour parameters ζ_i and ζ_r are given by

$$\zeta_i = \varepsilon_i \sqrt{k_0(r_{se} + r_{ea} - r_{sa})},$$

$$\zeta_r = \varepsilon_i \sqrt{k_0(r'_{se} + r_{ea} - r'_{sa})}. \tag{18-164}$$

where ε_i is the shadow indicator with

$$\varepsilon_i = \begin{cases} -1 & : \text{lit region} \\ 1 & : \text{shadow region} . \end{cases} \tag{18-165}$$

If the cylindrical wave emanating from the source is normalized to “1” at a point on the edge, the modified geometrical optics field becomes

$$E_{mg}^y(\vec{r}) = \Delta F(\zeta_i) - \sqrt{\frac{r_{sa}}{r'_{sa}}} e^{ik_0(r_{sa} - r'_{sa})} \Delta F(\zeta_r) . \tag{18-166}$$

With the same normalization, the boundary diffraction wave becomes

$$E_{BDW}^y(\vec{r}) = D_{TM} \frac{e^{ik_0 r_{ca}}}{\sqrt{r_{ca}}} \sqrt{\frac{r_{sa}}{r_{se}}} e^{ik_0(r_{se} - r_{sa})} \Delta F(\zeta_r) , \tag{18-167}$$

with the diffraction coefficient (ψ_i and ψ_r are defined in figure 18-40)

$$D_{TM} = \frac{e^{i\pi/4}}{\sqrt{8\pi k_0}} \left[\csc\left(\frac{\psi_i}{2}\right) - \csc\left(\frac{\psi_r}{2}\right) \right]. \quad (18-168)$$

As a consequence of the simple form of these equations, the calculation of the diffraction field by means of the UAT can be performed extremely quickly.

The diffraction coefficient D_{TM} is determined by comparison with the asymptotic extension of the well known exact solution of A. Sommerfeld [18-2]. D_{TM} becomes singular for $\psi_i = 0$ and $\psi_r = 0$, and consequently the boundary diffraction wave becomes discontinuous here. In the UAT these singularities are corrected by the singularities of ΔF , while they are not corrected in the GTD.

Using eqs (18.166)–(18.168), the diffraction field behind a perfectly conducting half-plane can be calculated for TE polarization. For TM polarization the y-component of the magnetic field can be calculated in a very similar way if (18.166) and (18.167) are replaced by

$$H_{mg}^y(\vec{r}) = \Delta F(\zeta_i) + \sqrt{\frac{r_{sa}}{r_{sa}'}} e^{ik_0(r_{sa}' - r_{sa})} \Delta F(\zeta_r) \quad (18-169)$$

and

$$H_{BDW}^y(\vec{r}) = D_{TE} \frac{e^{ik_0 r_{ea}}}{\sqrt{r_{ea}}} \sqrt{\frac{r_{sa}}{r_{se}}} e^{ik_0(r_{se} - r_{sa})} \Delta F(\zeta_r) \quad (18-170)$$

with the diffraction coefficient

$$D_{TE} = \frac{e^{i\pi/4}}{\sqrt{8\pi k_0}} \left[\csc\left(\frac{\psi_i}{2}\right) + \csc\left(\frac{\psi_r}{2}\right) \right]. \quad (18-171)$$

The change from TE to TM polarization is equivalent to a change in the boundary conditions which are given for a perfectly conducting screen in the xy -plane as

$$E_{x,y} = 0, \quad \frac{\partial E_z}{\partial z} = 0, \quad \frac{\partial H_{x,y}}{\partial z} = 0, \quad H_z = 0, \quad (18-172)$$

i.e., the boundary condition for E_y in TE polarization is $E_y = 0$ and for H_y in TM polarization it is $\partial H_y / \partial z = 0$.

The formalism of the GTD/UAT can be extended to wedges [18-43] and to opaque objects with limited conductivity as well as to phase objects [18-44]. For this, the appropriate diffraction coefficient has to be determined [18-45] which requires objects of simple geometry (half-planes, wedges, etc.). A simple extension of the UAT to phase-shifting half-planes is derived in [18-46].

18.15.2

An Empirical Boundary Diffraction Wave

Using Sommerfeld's rigorous solution for diffraction at a perfectly conducting half-plane [18-2], Andrews and Margolis [18-47] derived an empirical expression for the boundary diffraction wave at such a half-plane in TE polarization (E parallel to the edge): For a linearly polarized, normally incident plane wave, the boundary diffraction wave becomes

$$\frac{E_y^{BDW}(x, z)}{E_y^i(0, 0)} = -\sqrt{\frac{1 + \sin \phi}{(1 + g) \frac{\rho}{\lambda}}} \frac{e^{i(\beta + \beta_g)}}{2\pi \sin \phi} \quad (18-173)$$

with the polar coordinates (cf. figure 18-41)

$$\rho = \sqrt{x^2 + z^2}, \quad \tan \phi = \frac{z}{x} \quad (18-174)$$

with

$$g = \frac{1 + 1.5\sqrt{\beta}}{\pi\beta(1 + \beta)} \quad (18-175)$$

and

$$\beta_g = \frac{\pi}{4} (1 - e^{-\sqrt{\beta}}) \quad (18-176)$$

where

$$\beta = k_0(\rho - z). \quad (18-177)$$

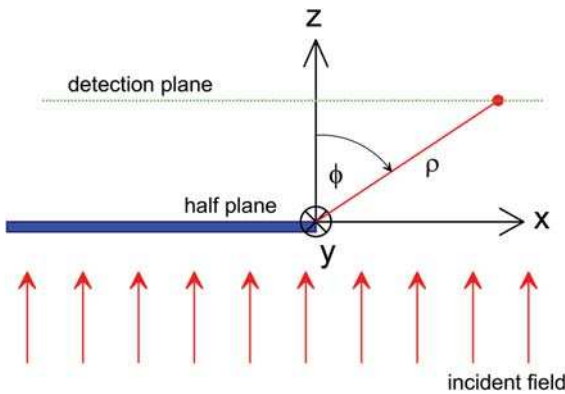


Figure 18-41: Polar coordinates for empirical boundary diffraction wave at a half-plane.

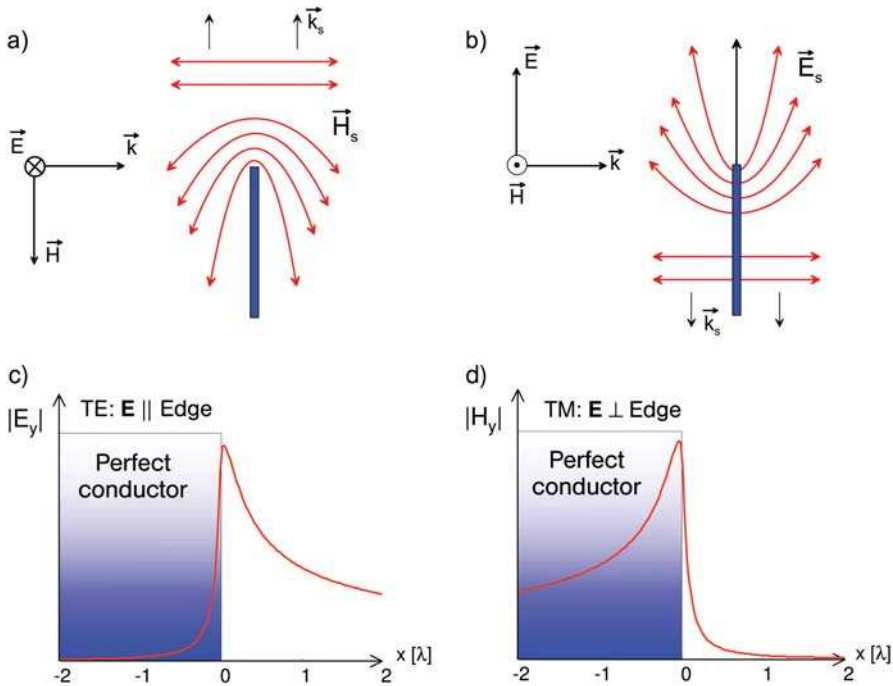


Figure 18-42: The empirical boundary diffraction wave. Top: field lines (after [18-47]). Bottom: amplitude at a distance of 0.1λ behind the object plane.

Figure 18-42 shows, on the left-hand side, the amplitude of the boundary diffraction wave for TE polarization ($E \parallel$ edge). Note the strong asymmetry of the amplitude which is due to the fact that, for $x < 0$, the boundary diffraction wave propagates in the vicinity of a perfect conductor with the boundary condition $E_y = 0$ at its surface. With increasing distance the field becomes more symmetrical.

With the electromagnetic form of Babinet's principle, the TM polarized version of the boundary diffraction wave is easily obtained. In fact, Andrews and Margolis used this boundary diffraction wave for a re-derivation of the electromagnetic Babinet principle. A consequence of this principle, is that the electric amplitude for TE polarization and the magnetic amplitude for TM polarization are mirror-symmetric to each other. The field lines plotted at the top of figure 18-42 illustrate this.

However, in the last two sections we have clearly left the field of scalar optics. The methods were nevertheless presented in this (scalar) section because the considered diffraction problems are essentially two-dimensional, thus reducing the vector diffraction problem to a scalar one. The discussion of actual vector diffraction theories will be presented in chapter 27.

18.16

Literature

- 18-1 I.N. Bronstein, K.A. Semendjajew, Taschenbuch der Mathematik, 19th ed. (Leipzig, 1979), p. 496, 494.
- 18-2 A. Sommerfeld, Vorlesungen über Theoretische Physik, Bd.3, Optik, 3rd ed. (Leipzig, 1960).
- 18-3 J. W. Goodman, Introduction to Fourier Optics, 2nd. ed. (McGraw Hill, New York, 1996).
- 18-4 G.C.Heurtley, Scalar Rayleigh–Sommerfeld and Kirchhoff diffraction integral: A comparison of exact evaluations for axial points, *J. Opt. Soc. Am.* **63**, 1003–1008 (1973).
- 18-5 E.Wolf, E.W.Marchand, Comparison of the Kirchhoff and Rayleigh-Sommerfeld theories of diffraction at an aperture, *J. Opt. Soc. Am.*, **54**, 587–594 (1964).
- 18-6 E.W.Marchand, E. Wolf, Consistent formulation of Kirchhoffs diffraction theory, *J. Opt. Soc. Am.*, **56**, 1712–1722 (1966).
- 18-7 E.W.Marchand, E. Wolf, Diffraction at small apertures in black screens, *J. Opt. Soc. Am.*, **59**, 79–90 (1969).
- 18-8 L. Eyges, The Classical Electromagnetic Field, (Adison-Wesley, 1972) p. 263.
- 18-9 J.J. Stamnes, Waves in Focal Regions, Adam Hilger Series on Optoelectronics.
- 18-10 J.E.Harvey, Fourier treatment of near-field scalar diffraction theory, *American Journal of Physics*, **47**, 974–980 (1979).
- 18-11 J.E.Harvey, R.V.Shack, Aberrations of diffracted wave fields, *Appl. Opt.*, **17**, 3003–3009 (1978).
- 18-12 A. Papoulis, Systems and Transforms with Applications in Optics (McGraw Hill, New York, 1968) p. 316f.
- 18-13 S. A. Collins, Lens-system diffraction integral written in terms of matrix optics, *J. Opt. Soc. Am.* **60**, 1168 (1970).
- 18-14 P. Baues, The connection of geometrical optics with the propagation of Gaussian beams and the theory of optical resonators, *Opto-Electronics* **1**, 103 (1969).
- 18-15 Y. A. Ananev and A. Y. Bekshaev, Matrix method of analysis of misaligned optical systems with astigmatic elements, *Opt. Spectrosc.* **66**, 531 (1989).
- 18-16 H. Weber, Wave optical analysis of the phase space analyser, *J. Mod. Opt.* **39**, 543 (1992).
- 18-17 J.W. Gibbs, Fourier's series, *Nature* **59**, 200 (1898).
- 18-18 W. J. Thomson, Fourier series and the Gibbs phenomenon, *Am. J. Phys.* **60**, 425–429 (1992).
- 18-19 G.J. Swanson, Binary optic technology: the theory and design of multi-level diffractive optical elements, MIT Technical Report 854 (1989).
- 18-20 M. Totzeck, B. Kuhlow, Validity of the Kirchhoff approximation for diffraction by weak phase objects, *Opt. Commun.* **78**, 13–19 (1990).
- 18-21 M. Babinet, Memoires d'optique météorologique, *C.R. Acad. Sci.* **4**, 638–648 (1837).
- 18-22 R.W. Pohl, *Optik*, 10th ed. (Springer, 1958).
- 18-23 C.J. Bouwcamp, Diffraction theory, *Rep.Prog. Phys.* **17**, 35–100 (1954).
- 18-24 T.B A. Senior, Some extensions of Babinet's principle in electromagnetic theory, *IEEE Trans. Antennas Propag.* **AP-25**, 417–420 (1977).
- 18-25 M. Totzeck, M.A. Krumbügel, Extension of Babinet's principle and the Andrew boundary diffraction wave to weak phase objects, *J. Opt. Soc. Am.* **11**, 3235–3240 (1994).
- 18-26 D. S. Jones, *Acoustic and Electromagnetic Waves* (Oxford University Press, Oxford, New York, 1986) Chap. 6, p. 321.
- 18-27 A. Sommerfeld, *Mathematische Theorie der Diffraktion*, *Math. Ann.* **47**, 317–374 (1896).
- 18-28 R. Dändliker, K. Weiss, Reconstruction of the three-dimensional refractive index from scattered waves, *Opt. Commun.* **1**, 323–328 (1970).
- 18-29 A. W. Lohmann, Three-dimensional properties of wave-fields, *Optik* **51**, 105–117 (1978).
- 18-30 W. Singer and K.-H. Brenner, Transition of the scalar field at a refracting surface in the generalized Kirchhoff diffraction theory, *J. Opt. Soc. Am. A* **1913–1919** (1995).
- 18-31 D.A. DeWolf, Wave propagation through quasi-optical irregularities, *J. Opt. Soc. Am.* **55**, 812–817 (1965).
- 18-32 V. I. Tatarski, *Wave Propagation in a Turbulent Medium* (McGraw-Hill Inc., New York, 1961).
- 18-33 A. Rohrbach und W. Singer, Scattering of the scalar field at a dielectric surface by Born se-

- ries expansion, *J. Opt. Soc. A* **15**, 2651–2659 (1998).
- 18-34** W. Singer und H.J. Tiziani, The Born-approximation for the non-paraxial scalar treatment of thick phase gratings, *Appl. Opt.* **37**, 1249–1255 (1998).
- 18-35** T. Young, On the theory of light and colors, *Philosophical transactions of the Royal Society* 1802, Reprint in G. Peacock, ed., *Miscellaneous Works of the late Thomas Young*, M.D., F.R.S.&C, Vol. 1 (John Murray, London, 1855) which is reprinted in *The Sources of Science*, No. 67, (Johnson Reprint Corp., New York, 1972).
- 18-36** G. A. Maggi, Sulla propagazione libera e perturbata delle onde luminose in un mezzo isotropo, *Ann. Matemat. Pura Applic. II* **16**, 21–48 (1888).
- 18-37** A. Rubinowicz, *Die Beugungswelle in der Kirchhoffsche Theorie der Beugung*, 2nd ed. (Springer Verlag, Berlin, 1966).
- 18-38** E.W. Marchand, E. Wolf, Boundary diffraction wave in h domain of the Rayleigh-Kirchhoff diffraction theory, *J. Opt. Soc. Am.* **52**, 761–767 (1962).
- 18-39** J.B. Keller, Geometrical theory of diffraction, *J. Opt. Soc. Am.* **52**, 116–130 (1962).
- 18-40** R.M. Lewis, J. Boersma, Uniform asymptotic theory of edge diffraction, *J. Math. Phys* **10**, 2291–2305 (1969).
- 18-41** R.G. Kouyoumjian, P.H. Pathak, Uniform geometrical theory of diffraction for an edge in a perfectly conducting surface, *Proc. IEEE* **62**, 1448–1461 (1974).
- 18-42** R.C Menendez, S.W. Lee, On the role of the geometrical optics field in aperture diffraction, *IEEE Transact. Ant. & Prop.* **AP-25**, 688–695 (1977).
- 18-43** R. Tiberio, G. Pelosi, G. Manara, A uniform GTD formulation for the diffraction by a wedge with impedance faces, *IEEE Trans. Ant. & Prop.* **AP-33**, 867–873 (1985).
- 18-44** J.L. Volakis, T.B.A. Senior, Diffraction by a thin dielectric half-plane, *IEEE Trans. Ant. & Prop.* **AP-35**, 1483–1487 (1987).
- 18-45** G.L. James, *Geometrical Theory of Diffraction for Electromagnetic Waves*, IEE *Electromagnetic Waves Series 1*, (Peter Peregrinus Ltd., 1975).
- 18-46** M.A. Krumbügel, M.Totzeck, Calculation of the diffraction near-field of weak phase objects by a simple adaptation of the Uniform Asymptotic Theory, *Optik*, **96**, 69–77 (1994). M.A. Krumbügel, M.Totzeck, Improved adaption of the Uniform Asymptotic Theory for calculation of the diffraction nearfield of weak phase objects, *Optik*, **97**, 165–168 (1994).
- 18-47** C.L. Andrews, D. P. Margolis, Elemental derivation of the Babinet principle in electromagnetic form, *Am. J. Phys.* **43**, 672–676 (1975).

19

Interference and Coherence

- 19.1 Basic Principles 100
- 19.1.1 Introduction 100
- 19.1.2 Two-beam Interference and Double Slit Diffraction 102
- 19.1.3 Contributions of Different Points of the Light Source 105
- 19.1.4 The High-frequency Term 107
- 19.1.5 The Low-frequency Term 108
- 19.1.6 Different Light Source Points with Statistical Phase 109
- 19.2 Mathematical Description of Coherence 113
- 19.2.1 Coherence Function 113
- 19.2.2 Wigner Distribution Function 116
- 19.2.3 Moments of the Wigner Distribution Function 120
- 19.2.4 Smoothing of the Wigner Distribution Function and Diffraction Focus 121
- 19.2.5 Wigner Distribution Function of Coherent Fields 122
- 19.2.6 Ambiguity Function 123
- 19.2.7 The Characterizing Functions in their Context 125
- 19.3 Temporal Coherence 126
- 19.3.1 Superposition of Signals with Different Frequency 126
- 19.3.2 Spectral Distribution of a Light Source 127
- 19.3.3 Bandwidth-limited Signals 128
- 19.3.4 Axial Coherence Length 130
- 19.3.5 Thermal Light Sources 133
- 19.3.6 Temporal Coherence in the Michelson Interferometer 134
- 19.4 Spatial Coherence 135
- 19.4.1 Introduction 135
- 19.4.2 Propagation of the Coherence Function 138
- 19.4.3 Van Cittert-Zernike Theorem 140
- 19.4.4 The Coherence Function of a Circular Source 140
- 19.4.5 Coherence Function behind a Double Slit 143
- 19.4.6 Propagation of the Wigner Distribution Function 146
- 19.5 Gaussian Schell Beams 149
- 19.5.1 Definition of Gaussian Schell Beams 149
- 19.5.2 Coherence and Wigner Functions of Gaussian Schell Beams 154

19.5.3	Basis Mode Expansion of Partial Coherent Fields	156
19.6	Statistical Optics and Speckle	159
19.6.1	Photon Statistics	159
19.6.2	The Speckle Effect	161
19.6.3	Speckle Parameters and Surface Structure	163
19.6.4	Computation of Speckle Effects	165
19.6.5	Speckle Reduction	169
19.7	Array Homogenizer	172
19.7.1	Setup of the System	172
19.7.2	Pupil Filling	175
19.7.3	Coherence Effects	176
19.7.4	Example Calculation	177
19.8	Miscellaneous	179
19.8.1	General Coherence Length	179
19.8.2	General Degree of Coherence	182
19.8.3	Coherence and Polarization	183
19.9	Literature	184

19.1 Basic Principles

19.1.1 Introduction

Usually only the light intensity can be measured in optics, i.e., the complex field amplitude is not measurable at optical frequencies of the order of 10^{14} Hz. Generally the absolute value of the amplitude of the electrical field can only be derived from the measured intensity, and those effects, in which the phase of the field amplitude is involved are actually not very common in practice. In special experimental arrangements and in some cases, the phase relationships of the field amplitudes are manifested in the occurrence of characteristic intensity patterns known as interference effects. In interferometry, e.g., these interference effects are used for relative measurements of the path length. According to the definition given by Zernike, coherence should be understood as a measure for the observability of interference effects, whereas the degree of coherence in a given intensity pattern depends in general on the light source and the viewing apparatus [19-1].

In some cases, the purpose is to utilize coherence effects: examples are the previously mentioned interferometry for relative measurements of the path length or optical image formation. According to the theory of Abbe, image formation is a result of the interference of the diffraction orders occurring through diffraction at the object from a light source. However, when operating with coherent light, unwanted interference effects might occur. Sometimes the term “coherent light” is defined by the occurrence of such effects. Thus, optical systems based on laser light sources, e.g., often suffer from such undesired or parasitic interference effects, caused, e.g., by scattering at rough surfaces, atmospheric turbulences or contaminations, which are known as speckle effects.

The concept of coherence is therefore closely related to the phenomenon of interference. Since at least two interfering “partners” are always involved, one often speaks of mutual coherence. In general there are three basic properties which affect the mutual coherence of interfering electromagnetic fields: the frequency or color spectrum, the spatial distribution or shape and size of the source and the polarization. The summation of the light amplitude can produce interference phenomena only for the same vector components of the electromagnetic field; different vector components have no effect on each other. Hence the discussion in this chapter will be restricted to the scalar case.

Coherence in a more general sense is a statistical effect in wave optics and is discussed as such in numerous references [19-2] through to [19-11]. If there is no fixed coupling of the phases of the complex field components, the superposition or the interference of the waves results in reduced contrast of the interference patterns and phase changes will have a minor impact. In the limit of totally statistical behavior of the phases of the field components, down to dimensions of the wavelength, a wave field is identified with perfect incoherence. If, on the other hand, all phases are strictly coupled, the resulting intensity pattern will be very sensitive on phase

changes, one speaks of ideal or complete coherence. Thermal light sources are, to a large extent, incoherent and lasers operating in their fundamental mode can be considered to be coherent. The question of coherence, from a statistical point of view, is thus also connected to the amount of information in a wave field. A coherent wave field may be described by a single mode, while incoherent wave fields need a superposition of at least several independent modes.

The coherence properties of optical imaging consequently are determined by:

1. Statistical initial phases of the emitted waves.
2. The finite length of the wave train and its finite spectral bandwidth.
3. Large differences in the optical path lengths of the transfer system as a result of the finite size of the light sources.

Coherence is a spatial-temporal phenomenon, and may be separated in purely spatial and purely temporal coherence. The parameters for the description of coherence depend on the position in space and vary with the light propagation, but can be recognized also at a definite position by observing the time dependence, e.g. These two views on coherence properties can be observed in interferometric arrangements by splitting the input wave either by “amplitude” or “phase” splitting:

- In the Michelson interferometer the amplitude is split and one observes the temporal coherence.
- In the Young interferometer the wave is geometrically split and one observes the spatial coherence.

In the so called phase space or, more general, Wigner space, fields with different spatial coherence can be distinguished by their characteristic features. For a coherent field, a well-defined coherent field strength exists at each point and consequently can interfere everywhere. In phase space, one may obtain a narrow line whose width is given by the diffraction limit. The Wigner distribution function, discussed later in this chapter, for coherent wave fields will also have negative values, by which diffraction and interference effects are considered. On the contrary, when the light field is perfectly incoherent, no destructive interference can take place, and one obtains in the phase space, in the frame of the geometrical boundary conditions, a distribution with a wide angular spectrum for each spatial position. In general the distribution in the phase space exhibits an intermediate size between these two limiting cases and, depending on geometry and coherence properties, a specific angular distribution at each point. Figure 19-1 illustrates the different spatial-angular radiation distributions and the representations in the phase space. This indicates that the description of partially coherent light contains a substantially larger number of degrees of freedom and correspondingly the numerical simulations of the light propagation or imaging will require substantially more effort.

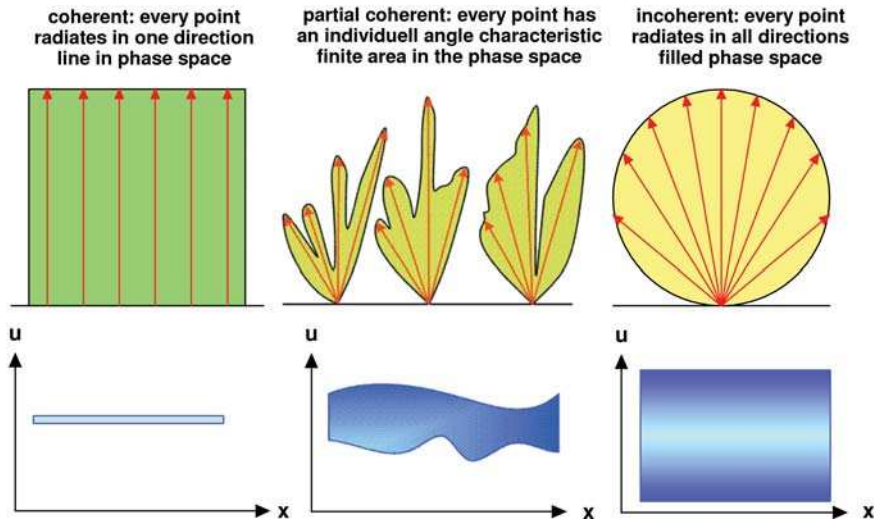


Figure 19-1: Radiation characteristics of light sources with different coherence and the corresponding areas in the phase space with the corresponding density distributions.

19.1.2

Two-beam Interference and Double Slit Diffraction

A monochromatic plane wave incident on an infinitesimal slit produces a cylindrical wave. Thus two cylindrical waves are generated behind a double slit. Hence, the two slits correspond to two sources, which excite two field distributions oscillating in phase with the incident plane wave. The superposition of two waves can be illustrated by the Moiré pattern, e.g., by marking the crests of the waves and superimposing the two patterns at intervals from the two slits (Figure 19-2). Dark Moiré fringes

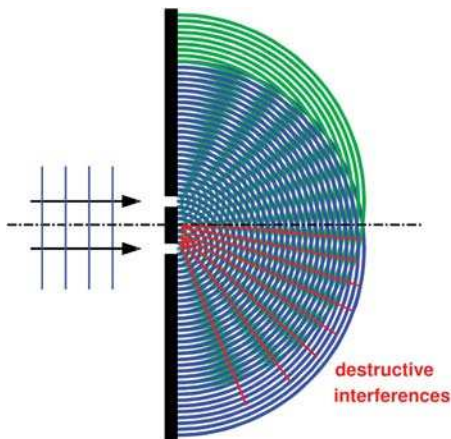


Figure 19-2: Illustrating the interference behind a double slit by a Moiré pattern.

indicate zones where the crest of the first wave coincides with the trough of the other wave and the two wave amplitudes with opposite phases extinguish each other. In contrast, in the light regions the two waves interact in-phase, i.e., the common wave crests are intensified.

Hence the resulting intensity at any point behind a double slit is determined by the superposition of the two field amplitudes U_1 and U_2

$$I(x, z) = |U_1(x, z) + U_2(x, z)|^2. \quad (19-1)$$

Let us assume for simplicity that a monochromatic point light source illuminates the two slits from some distance as shown in figure 19-3. Depending on the position x_s of the light source and the position of observation x_2 , different optical path lengths are obtained for the two slits. In one of the limiting cases, the maximum intensity is obtained for the image point x_2 when the difference in the path lengths is an integer multiple of 2π , i.e., the two wave crests coincide and constructive interference takes place. In the other extreme case one obtains complete extinction when the optical path difference (OPD) s_{OPD} is an odd integer multiple of π , i.e., when wave crests meet wave troughs. The optical path difference depends on the separation z_1 between the light source and the double slit and the distance z_2 between the double slit and the plane of observation, as well as on the slit spacing D , the lateral position of the light source x_s and the lateral position of the detector x_2 . Using eq. (19-1) it is easy to derive the following expression for the resulting intensity:

$$I(x_2) = |A \cdot e^{ik_0 \cdot s_{OPD1}} + A \cdot e^{ik_0 \cdot s_{OPD2}}|^2 \approx 2A^2 + 2A^2 \cos(k_0 \cdot \Delta s_{OPD}), \quad (19-2)$$

$$\Delta s_{OPD} = s_{OPD2} - s_{OPD1}, \quad (19-3)$$

$$s_{OPD1} = a_1 + b_1, \quad (19-4)$$

$$s_{OPD2} = a_2 + b_2. \quad (19-5)$$

Figure 19-3 shows the geometry of the double slit arrangement.

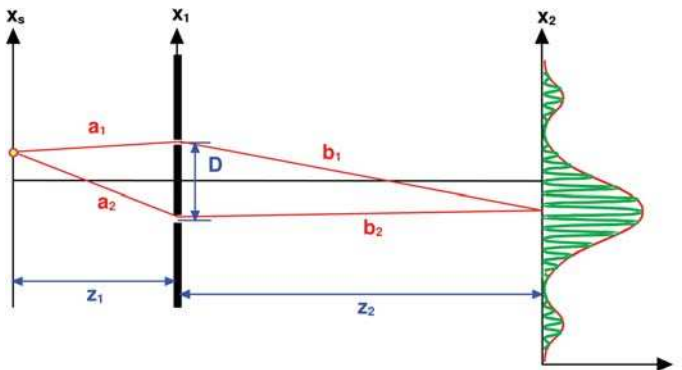


Figure 19-3: Diffraction by a double slit in the case of a monochromatic point light source.

The exact geometrical realization of the above arrangement has a negligible impact on the final result, which can be obtained as a solution of the optical diffraction problem using, e.g., the Kirchhoff integral. In the far-field limit one can apply the Fraunhofer approximation, according to which the diffraction pattern and the intensity distribution are given by the Fourier transform of the transmission function of the double slit. The double slit with a spacing D can be represented by a convolution of two delta-functions with a rect-function for a single slit of width d :

$$T(x_1) = \text{rect}\left(\frac{x_1}{d}\right) \otimes \left[\delta\left(x_1 - \frac{D}{2}\right) + \delta\left(x_1 + \frac{D}{2}\right) \right]. \quad (19-6)$$

After illumination by a plane wave with a propagation angle α relative to the z-axis, the far-field distribution is given by the product of a cosine and a sinc-function:

$$U(x_2) \approx \left[\text{sinc}\left(\frac{d}{\lambda z} \cdot x_2\right) \cdot \cos\left(\pi \frac{D}{\lambda z} x_2\right) \right]. \quad (19-7)$$

The sinc-shaped envelope function degenerates into a constant only for delta-shaped slit apertures, which results in simple cosine-shaped interference fringes as defined in eq. (19-2).

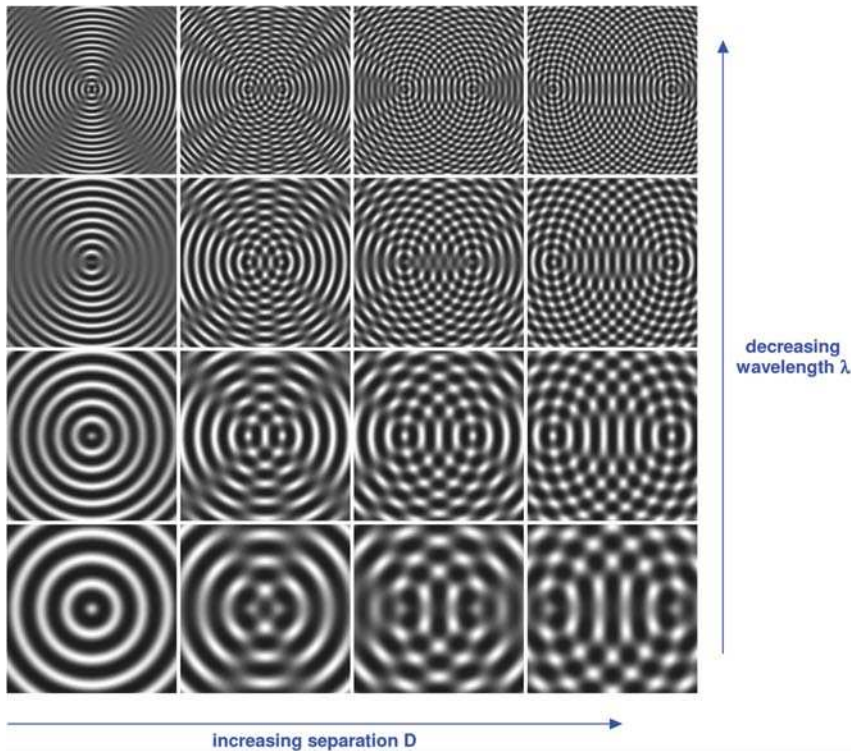


Figure 19-4: Interference patterns resulting from the coherent superposition of two point light sources versus their distance D and wavelength λ .

Figure 19-4 illustrates the interference patterns obtained in a two-pinhole experiment, which is equivalent to the superposition of two coherently emitting light points in a plane containing both of them. It illustrates the interference pattern resulting from the coherent superposition for different separations D and wavelengths λ . The number of interference patterns increases with decreasing wavelength and increasing pinhole distance. This corresponds to the argument in the cos function of the last expression in eq. (19-2), which scales with $\Delta s_{OPD}/\lambda$.

19.1.3

Contributions of Different Points of the Light Source

A light source of finite size can be considered as consisting of many point light sources. Each point light source produces, depending on its position in space, a laterally shifted interference pattern, whereby the lateral position of the far-field distribution is determined by the position of the source point x_s . First the question will be investigated, whether waves, emitted by different light source points, may interfere at all.

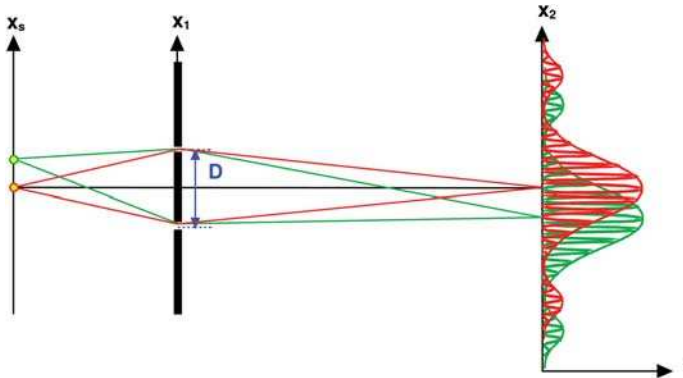


Figure 19-5: Interference between two points of a light source.

For the consideration the interference of two point sources it is assumed that the emitted waves are monochromatic with different frequencies ω_1 and ω_2 :

$$\begin{aligned} U_1 &= A_1 \cos(k_1 r - \omega_1 t + \varphi_1), \\ U_2 &= A_2 \cos(k_2 r - \omega_2 t + \varphi_2). \end{aligned} \tag{19-8}$$

In the special case of equal amplitudes $A_1=A_2=A$ and constant phase difference $\varphi_1 - \varphi_2 = \text{const.}$, one obtains for the field in the position x_2

$$\begin{aligned} U(t) &= A \cdot \cos(\omega_1 t) + A \cdot \cos(\omega_2 t) \\ &= 2A \cdot \cos\left[\frac{(\omega_1 + \omega_2) \cdot t}{2}\right] \cos\left[\frac{(\omega_1 - \omega_2) \cdot t}{2}\right]. \end{aligned} \tag{19-9}$$

It follows from the last equation that the intensity is given by:

$$I(t) = 4A^2 \cdot \cos^2 \left[\frac{(\omega_1 + \omega_2) \cdot t}{2} \right] \cdot \cos^2 \left[\frac{(\omega_1 - \omega_2) \cdot t}{2} \right]. \quad (19-10)$$

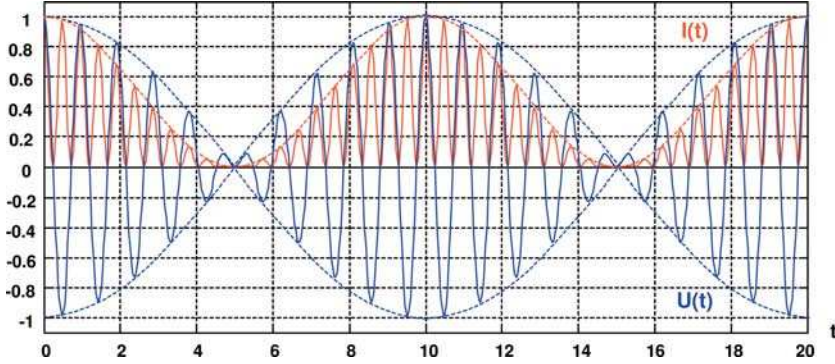


Figure 19-6: Beating as a result of the superposition of two plane waves with different carrier frequencies ($\omega_1 = 2\pi$, $\omega_2 = 2.2\pi$). $U(t)$ is the amplitude and $I(t)$ the intensity of the interference signal.

The resulting interference pattern contains a high-frequency and a low-frequency term. The high-frequency term will be considered first.

19.1.4

The High-frequency Term

Even if the carrier frequencies are equal, there is an oscillating term with frequency $\omega = (\omega_1 + \omega_2)/2$. As a consequence, the interference pattern will, in any case, be “flickering” with an average temporal frequency $1/T = \omega/2\pi$, defined by the propagation velocity c_0 of light divided by the wavelength λ :

$$\frac{\omega}{2\pi} = \frac{1}{T} = \frac{c_0}{\lambda}. \quad (19-11)$$

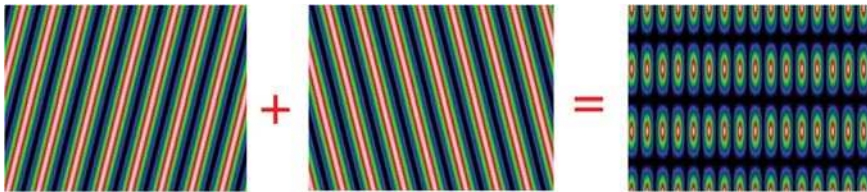


Figure 19-7: Spatial interference pattern of two plane waves: a spatially high-frequency pattern emerges which propagates to the right at $\sim 10^{15}$ Hz.

If, e.g., one considers green light with a wavelength of 500 nm, the corresponding frequency is 0.6 PHz = 600.000 GHz:

$$\lambda = 500 \text{ nm} \rightarrow \frac{1}{T} = \frac{c_0}{\lambda} = 3 \cdot 10^8 \frac{\text{m}}{\text{s}} / 0.5 \cdot 10^{-6} \text{ m} = 0.6 \cdot 10^{15} \text{ Hz} . \quad (19-12)$$

The optical frequencies are in the range $10^{14} - 10^{15}$ Hz and the high-frequency portion of the interference signal with $(\omega_1 + \omega_2)/2$ is of the same order as the average light frequency. Any detection process is associated with time averaging, electrons have eventually to be excited and one has to measure the current or a photographic film has to be blackened. Even modern detectors can operate only up to 10^{10} Hz (10 GHz). Hence it is necessary to carry out an additional temporal averaging in eq. (19-10) over approx. 10^4 oscillations. This time constant of the detector always averages the high-frequency term $(\omega_1 + \omega_2)/2$. This term is actually absent when one chooses the complex representation.

As a comment it has to be reminded, that the frequency is related to the energy of the radiation. The photon energy comes into play, if the photon statistics like the photon noise has to be considered. Such aspects will be disregarded at first.

19.1.5

The Low-frequency Term

In the complex representation, the two interfering waves are written as:

$$U_1 = A_1 e^{i(k_1 r - \omega_1 t + \varphi_1)} , \quad U_2 = A_2 e^{i(k_2 r - \omega_2 t + \varphi_2)} . \quad (19-13)$$

One obtains for the intensity:

$$\begin{aligned} I(r, t) &= |U_1 + U_2|^2 = U_1 U_1^* + U_1 U_2^* + U_2 U_1^* + U_2 U_2^* \\ &= I_1 + I_2 + A_1 A_2 \cdot [e^{i(\Delta k r - \Delta \omega t + \Delta \varphi)} + e^{-i(\Delta k r - \Delta \omega t + \Delta \varphi)}] \\ &= I_1 + I_2 + 2A_1 A_2 \cdot \cos(\Delta k r - \Delta \omega t + \Delta \varphi) . \end{aligned} \quad (19-14)$$

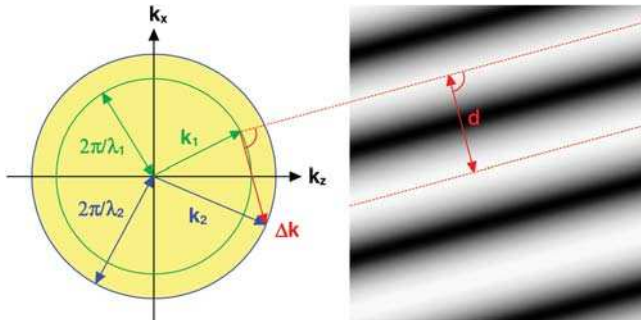


Figure 19-8: Vector construction of interference pattern with two Ewald spheres with different radius $\sim 1/\lambda_1$ and $\sim 1/\lambda_2$; different wavelengths produce a moving interference pattern with phase velocity $\Delta\omega$ and fringe period d .

In the case of light sources with different wavelengths one obtains running interference fringes with a spatial frequency $\Delta k = k_1 - k_2$ which move with a phase velocity of $\Delta\omega$ (see figure 19-8).

In addition, the interference pattern is phase-shifted by $\Delta\varphi$. Assuming at first a constant phase difference of $\Delta k \cdot r + \Delta\varphi = 0$, one obtains after time averaging:

$$\begin{aligned} I(r, t) &= \langle |U_1 + U_2|^2 \rangle = I_1 + I_2 + \frac{2A_1A_2}{T} \int_{-T/2}^{T/2} \cos(\Delta\omega t) dt \\ &= I_1 + I_2 + \frac{4A_1A_2}{T\Delta\omega} \sin\left(\Delta\omega \cdot \frac{T}{2}\right) \\ &= I_1 + I_2 + 2A_1A_2 \cdot \text{sinc}(\Delta\nu_i \cdot T) \end{aligned} \quad (19-15)$$

with $\Delta\omega = 2\pi\nu_i$. The contrast of the interference fringes decreases with the finite integration time of the detector and the frequency separation $\Delta\omega$. Assuming a detector with a temporal resolution of $1/T \sim 10^{10}$ Hz (10 GHz) one can estimate the frequency difference $\Delta\omega$ which still allows the measurement of the moving interference fringes. The first zero of the sinc-function occurs at

$$\Delta\omega = \omega_2 - \omega_1 = \frac{2\pi}{T}. \quad (19-16)$$

The substitution $\omega = 2\pi c/\lambda$ yields, in the limit of small wavelength differences $\Delta\lambda_s$ in the source, approximately

$$\Delta\lambda_s = \frac{\lambda^2}{cT}. \quad (19-17)$$

It follows that, for an average wavelength of 500 nm, the moving interference fringes cannot be observed if the wavelengths of two light sources differ by more than ~ 10 pm. Since the carrier frequencies in the optical range (of the order of 10^{14} s⁻¹) are too high even for the fastest detectors, the averaging of the moving interference fringes in the case of light sources with different wavelengths precludes any interference contrast behind the Young double slit. Only in the special examples like heterodyne interferometry slightly wavelength-shifted light sources and hence moving interference fringes are applied for the measurement, e.g., of distances.

19.1.6

Different Light Source Points with Statistical Phase

In the case of equal wavelengths eq. (19-14) is reduced to the time-independent form:

$$I(r) = I_1 + I_2 + 2A_1A_2 \cos(\Delta k \cdot r + \Delta\varphi). \quad (19-18)$$

Depending on the relative phase of the two waves a standing interference pattern is formed in which the period d of the interference fringes is given by

$$d = \frac{2\pi}{\Delta k} = \frac{\lambda}{2 \cdot \sin\left(\frac{\alpha}{2}\right)} \quad (19-19)$$

where α denotes the relative angle between the wave directions.

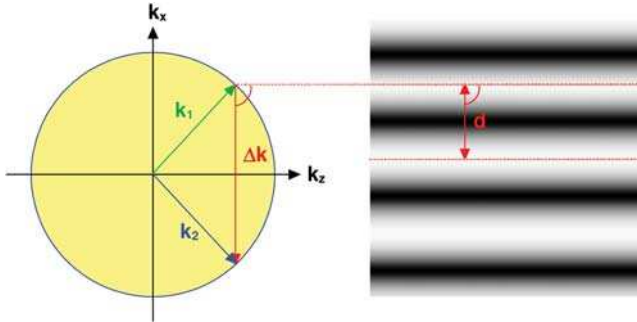


Figure 19-9: Vector construction of an interference pattern with the Ewald sphere for equal wavelengths.

However, even in the case of equal wavelengths waves emitted by different light sources, in general no visible interference fringes are formed if the phase relation $\Delta\varphi$ of different source points are not constant. The resulting intensity is obtained by time averaging over the time-dependent phase difference $\varphi(t)$:

$$I(r, t) = \langle |U_1 + U_2|^2 \rangle_T = I_1 + I_2 + \frac{2A_1A_2}{T} \int_0^T \cos(\Delta\vec{k} \cdot \vec{r} - \Delta\varphi(t)) dt. \quad (19-20)$$

Depending on the phase relation $\Delta\varphi$ between the two point sources, three cases can be distinguished (figure 19-10). The photons emitted by a thermal light source, e.g., an incandescent lamp, have statistical phases since the contribution of the spontaneous emission substantially exceeds that of the stimulated emission. The phase differences between different points of the light source fluctuate stochastically so that the interference term vanishes after time averaging. Thus one can analyse thermal light sources as consisting of separate, uncorrelated point light sources which do not interfere with each other. The resulting interference patterns of the individual light source points are therefore also uncorrelated and can be summed in terms of intensity. This corresponds to the right-hand side case in figure 19-10: two uncorrelated light source points produce no interference fringes and the visibility V vanishes. The case of complete correlation depicted on the left-hand side occurs, e.g., when the interfering waves originate from the same emission process and consequently exhibit a fixed relative phase. The case shown in the middle corresponds to weakly correlated relative phases, corresponding to a finite-sized thermal light source or as might occur in concatenated imaging. Averaging over all wave pairs then results in an averaged phase relation and a correspondingly weak interference contrast.

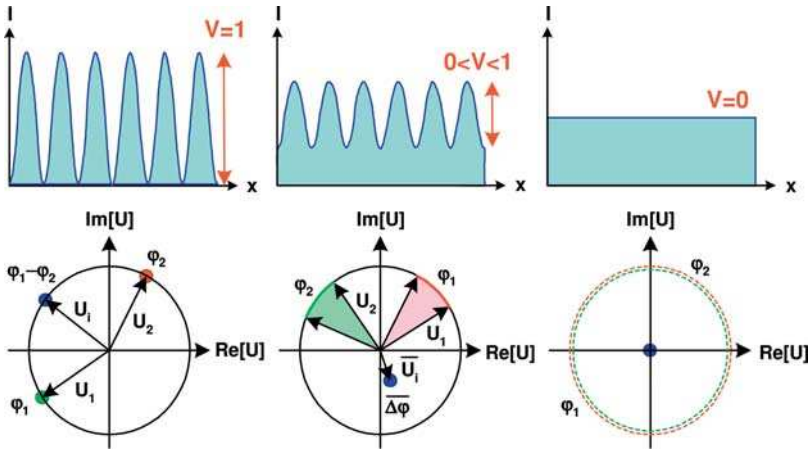


Figure 19-10: Comparison of different phase relationships between two stationary signals.

Thus, interference occurs only with wave trains, which originate from the same light source point and hence oscillate in-phase. Furthermore, interference between field distributions with different temporal frequencies or colour cannot be observed. Interference fringes of different colour occur of course under illumination with a white light source (figure 19-11), and the fringe period is, according to eq. (19-19), a function of the wavelength or colour. These coloured interference fringes emerge, however, always as a result of the interference of two monochromatic waves with subsequent incoherent superposition of the intensities of the different interference fringes corresponding to the different colours.

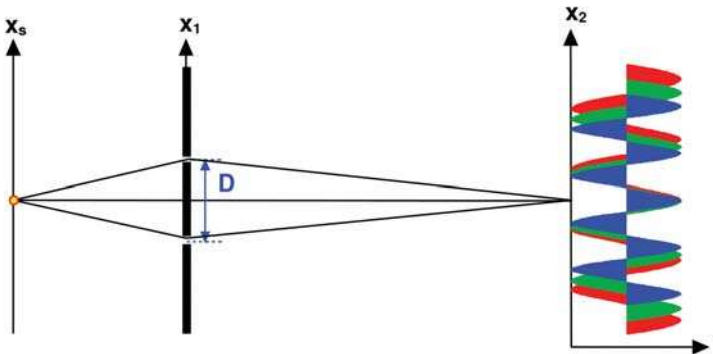


Figure 19-11: Coloured interference fringes by incoherent superposition of many interferences of different colours.

In the case of a finite-sized monochromatic light source, the intensities of the diffraction patterns of the double slit originating from separate source points are superposed in the observation plane. Depending on the separation between two source points this results in extinction of the contrast or in-phase superposition of the interference

fringes. In figure 19-12 two different separations of the source points are depicted. In the upper picture with separation a_1 constructive superposition of the interference fringes takes place, i.e., interference fringes with maximum contrast can be observed. For different separations, as in the case with separation a_2 shown in the lower picture, destructive superposition occurs, i.e., in the worst case the result is total extinction of the interference fringes. It is easy to see that the image contrast or visibility V are proportional to the cosine of the separation a of the light-source points:

$$V = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \sim \cos(B \cdot a) \quad (19-21)$$

where the factor B is determined by the geometry.

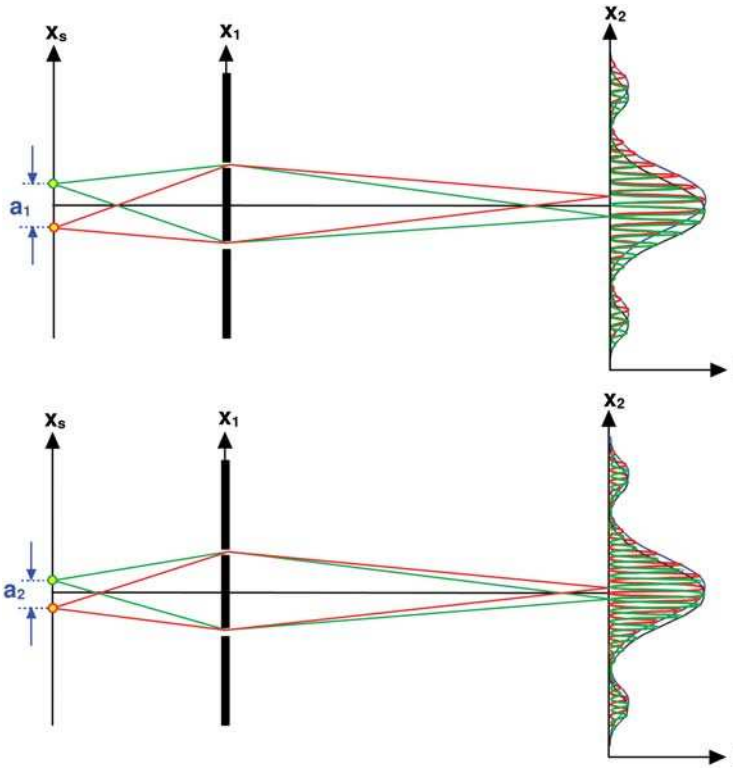


Figure 19-12: Superposition of the intensities in the detector plane for different separations a_1 and a_2 of two source points corresponding to constructive (upper picture) and destructive (lower picture) superposition.

If instead of two light-source points a finite-sized light source with a maximum extension D_S is considered, one obtains for the pattern contrast after integration over all source points

$$V = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \sim \text{sinc}(B \cdot D_S). \quad (19-22)$$

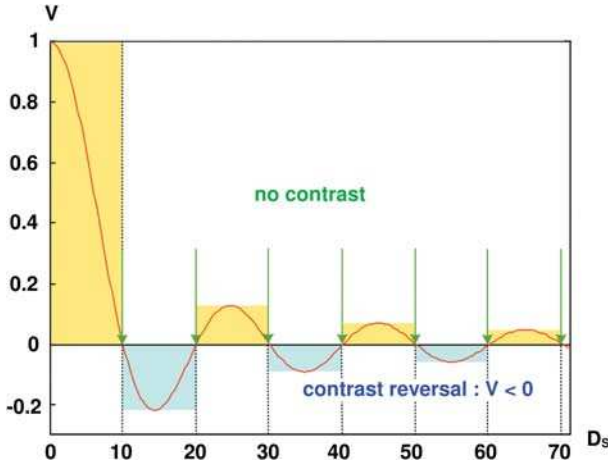


Figure 19-13: Visibility of the interference fringes behind a fixed double slit in dependence on the light-source diameter D_s .

Thus the interference contrast vanishes for increasing light source sizes and one obtains the intensity as an incoherent sum of the displaced diffraction patterns of the separate slits. This visual interpretation will be derived mathematically in the following on the basis of the coherence function. For the time being one can summarize:

- The slit spacing determines the frequency of the interference fringes.
- The slit width determines the extension of the interference region.
- The frequency bandwidth of the light spectrum determines the visibility of the interference fringes; as a rule, interference can be observed in the visible range only for monochromatic light.
- The finite size of the monochromatic light source determines the interference contrast.

19.2

Mathematical Description of Coherence

19.2.1

Coherence Function

For the description of coherence, the coherence function Γ serves as an important mathematical concept in statistical optics. It is defined as the correlation of the complex field amplitudes at different positions and times. The correlation integral is averaged over a time interval T

$$\begin{aligned} \Gamma_{12}(\tau) &= \Gamma(\vec{r}_1, \vec{r}_2, \tau) = \langle U(\vec{r}_1, t + \tau) U^*(\vec{r}_2, t) \rangle_T \\ &= \frac{1}{T} \int_t^{t+T} U(\vec{r}_1, t + \tau) \cdot U^*(\vec{r}_2, t) dt. \end{aligned} \quad (19-23)$$

Since the field in this definition is compared to itself, the coherence Γ is given by the cross-correlation of the field amplitude at different locations r and times t . Figure 19-14 illustrates the formation of the correlation at two different positions x_1 and x_2 within the light.

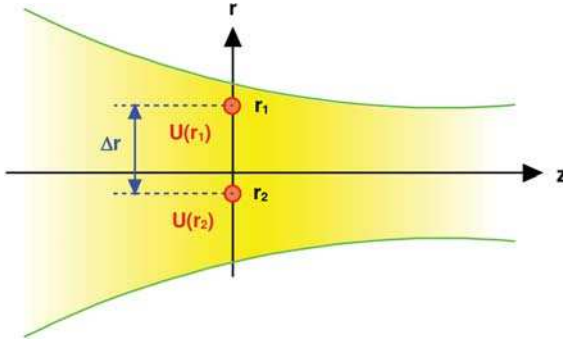


Figure 19-14: Schematic representation illustrating the interpretation of the coherence function as a correlation of the field strengths at two points in the radiation field.

In the general case, the function Γ is hermitian and consequently the permutation of the spatial vectors gives the complex conjugate:

$$\Gamma(\vec{r}_2, \vec{r}_1, \tau) = \Gamma^*(\vec{r}_1, \vec{r}_2, \tau). \quad (19-24)$$

After insertion of the complex amplitude, the first-order mutual coherence function describes the correlation of the two signals U_1 and U_2 in the same position

$$\Gamma(U_n, U_m) = \langle |U_n U_m^*| \rangle_T = \frac{1}{T} \int_0^T A_n A_m \cos [\Delta k_{nm} r - \Delta \omega_{nm} (t + \tau) + \Delta \varphi_{nm}] d\tau \quad (19-25)$$

with

$$\Delta \omega_{nm} = \omega_n - \omega_m, \quad (19-26)$$

$$\Delta k_{nm} = k_n - k_m, \quad (19-27)$$

$$\Delta \varphi_{nm} = \varphi_n - \varphi_m. \quad (19-28)$$

The mutual coherence function corresponds for $n = m$ to the normal intensity and is consequently referred to as mutual intensity. The intensity is obtained from the coherence function by taking coincident positions in space $\vec{r}_1 = \vec{r}_2 = \vec{r}$ and vanishing time difference.

Using eq. (19-25) one can identify the intensity with the autocorrelation function of a signal

$$I = \Gamma(U_m, U_m) = \Gamma(\vec{r}, \vec{r}, 0). \quad (19-29)$$

For multiple superposition of many amplitudes or modes, the coherence function can be written in general as a double sum over all ‘signals’ U_n :

$$I(r, t) = \sum_{n,m} \Gamma(U_n, U_m) = \sum_{n,m} \sqrt{I_n I_m} \gamma(U_n, U_m). \quad (19-30)$$

For mutual incoherent fields or modes of vanishing cross-correlation, eq. (19-30) simplifies to

$$\Gamma(\vec{r}_1, \vec{r}_2) = \sum_m \Gamma_m(\vec{r}_1, \vec{r}_2). \quad (19-31)$$

As a correlation, the coherence function is a mathematical construction and can not be measured directly. An evaluation of the coherence function is possible, since the coherence function specifies the contrast of the interference between two signals U_1 and U_2 . By normalization, the coherence function is transformed to the so-called complex degree of coherence γ with values limited by $0 \leq |\gamma| \leq 1$:

$$\gamma_{12}(\tau) = \gamma(\vec{r}_1, \vec{r}_2, \tau) = \frac{\Gamma_{12}(\tau)}{\sqrt{\Gamma_{11}(0) \cdot \Gamma_{22}(0)}} = \frac{\Gamma(\vec{r}_1, \vec{r}_2, \tau)}{\sqrt{I(\vec{r}_1) \cdot I(\vec{r}_2)}}. \quad (19-33)$$

The correlation between two points in real light fields normally decreases both with increasing time difference τ and spatial separation Δr . A typical dependence of the absolute value of the degree of coherence on these two differences is shown in figure 19-15.

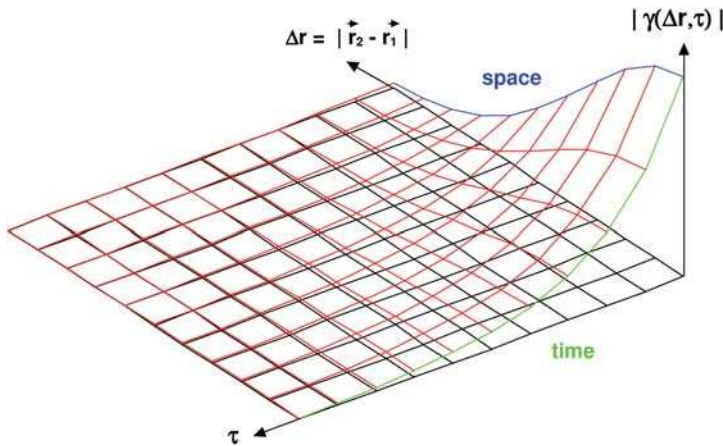


Figure 19-15: Schematic representation of the decreasing degree of coherence with increasing temporal and spatial separation.

Finally, the intensity of the interference pattern of two point light sources is given by

$$\begin{aligned} I(\vec{r}) &= I_1(\vec{r}) + I_2(\vec{r}) + 2\Gamma(\vec{r}_1, \vec{r}_2, 0) \\ &= I_1(\vec{r}) + I_2(\vec{r}) + 2\sqrt{I_1(\vec{r})I_2(\vec{r})}\gamma_{12}(0). \end{aligned} \quad (19-34)$$

The stronger the correlation between the separate light sources, the higher the contrast of the interference pattern. In the limit $\gamma=0$ no phase coupling is present and the two source points are not capable to interfere with each other. The visibility of the interference pattern is defined as

$$V = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} = \frac{2\sqrt{I_1(\vec{r})I_2(\vec{r})}}{I_1(\vec{r}) + I_2(\vec{r})} \cdot \gamma_{12}(0) \quad (19-35)$$

and it corresponds to the known contrast function in terms of intensities.

The temporal Fourier transform of the coherence function is referred to as the cross-spectral density (CSD):

$$S(\vec{r}_1, \vec{r}_2, \omega) = \int_{-\infty}^{\infty} \Gamma(\vec{r}_1, \vec{r}_2, t) \cdot e^{-i\omega t} dt. \quad (19-36)$$

This distribution plays a major role in the consideration of the temporal coherence and will be discussed in more detail in section 19.3.2.

The application of the wave equation in the Helmholtz formulation for the field amplitude U at positions r_1 and r_2 , respectively, leads in general to the two coupled wave equations in both spatial coordinates, applicable not only to the field amplitudes but also to the coherence function

$$\nabla_j^2 \Gamma - \frac{1}{c^2} \cdot \frac{\partial^2 \Gamma}{\partial t^2} = 0, \quad j = 1, 2. \quad (19-37)$$

This equation represents the general transport equation for the coherence function and is a direct consequence of the wave equation. The frequently applied transport integrals of the coherence function, according Cittert-Zernike and Hopkins are time-independent formulations and are frequently given within the limits of Kirchhoff or Fresnel approximation. These purely spatial aspects of the coherence propagation play an essential role in optical systems and will be discussed in section 19-4.

19.2.2

Wigner Distribution Function

The Wigner distribution function (WDF) is directly related to the coherence function and is a powerful and visual description of partially coherent wave fields [19-12]. As a quasi-density function in the phase space the Wigner distribution function W possesses a simple physical meaning. The combined use of coherence function Γ and

Wigner distribution function W is especially useful for the consideration of propagation of coherence properties through optical systems in the case of partially coherent radiation. With the center-of-mass and difference coordinates

$$\vec{r} = \frac{\vec{r}_1 + \vec{r}_2}{2}, \quad \Delta\vec{r} = \vec{r}_1 - \vec{r}_2 \quad (19-38)$$

$$\vec{\nu} = \frac{\vec{\nu}_1 + \vec{\nu}_2}{2}, \quad \Delta\vec{\nu} = \vec{\nu}_1 - \vec{\nu}_2 \quad (19-39)$$

the coherence function may be written as

$$\Gamma(\vec{r}_1, \vec{r}_2) \Rightarrow J(\vec{r}, \Delta\vec{r}). \quad (19-40)$$

The Wigner distribution function $W(x, \nu)$ can now be directly obtained by a Fourier transform of the coherence function J with respect to the difference coordinate Δr [19-13] with transverse spatial frequency coordinate ν :

$$W(\vec{r}, \vec{\nu}) = \int \Gamma\left(\vec{r} + \frac{\Delta\vec{r}}{2}, \vec{r} - \frac{\Delta\vec{r}}{2}\right) \cdot e^{-2\pi i \Delta\vec{r} \cdot \vec{\nu}} d\Delta\vec{r}. \quad (19-41)$$

After inversion of eq. (19-41) it is obtained

$$\Gamma(\vec{r}_1, \vec{r}_2) = \int W\left(\frac{\vec{r}_1 + \vec{r}_2}{2}, \vec{\nu}\right) \cdot e^{2\pi i \vec{\nu} \cdot (\vec{r}_1 - \vec{r}_2)} d\vec{\nu}. \quad (19-42)$$

Frequently a modified formulation of the Wigner distribution function is applied with the transversal spatial frequency ν replaced by the x - and y -components of the optical direction cosine vector \vec{p} , respectively. Accordingly it can be written for eq. (19-41)

$$W'(\vec{r}, \vec{p}) = \int \Gamma\left(\vec{r} + \frac{\Delta\vec{r}}{2}, \vec{r} - \frac{\Delta\vec{r}}{2}\right) \cdot e^{-i k_0 \Delta\vec{r} \cdot \vec{p}} d\Delta\vec{r} \quad (19-43)$$

with the optical direction cosine vector \vec{p} given by the sine of the angle u with respect to the optical axis:

$$\vec{p} = \sin \vec{u} = \lambda \cdot \vec{\nu} = \frac{\lambda}{2\pi} \cdot \vec{k}. \quad (19-44)$$

It is written W' instead of W to indicate the change of the second variable. In the paraxial approximation, the direction cosine p can be approximated by the angle u .

Since Wigner distribution function W' and coherence function are related by a Fourier transformation, both contains the same information, but in a different representation. In general both are four-dimensional functions. It represents a quasi-density function in the phase space: a local angular spectrum at position r and a local averaged phase space density, respectively. The Wigner distribution function thus describes the amplitude of a beam at a position x, y with direction p, q (see

figure 19-16 and figure 19-17). For example, the Wigner distribution function of a light source is limited by the area and the aperture of the source. As the coherence function, the Wigner distribution function W is applicable to radiation with an arbitrary degree of coherence. In addition, it considers wave-optic effects such as interference and diffraction. Since Γ is hermitian it follows that W is always real.

Although the function W' describes the density, its values are not necessarily positive everywhere. Negative values of W' indicate destructive interference effects.

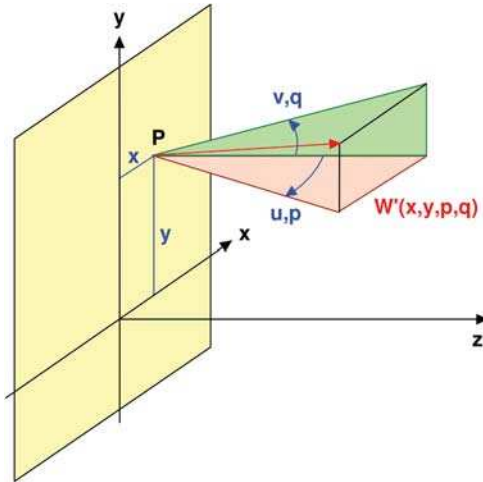


Figure 19-16: Schematic presentation of the Wigner distribution function of a quasi-point light source at point $P(x, y)$ as a local flow of energy in a definite direction (u, v) with the cosines (p, q) .

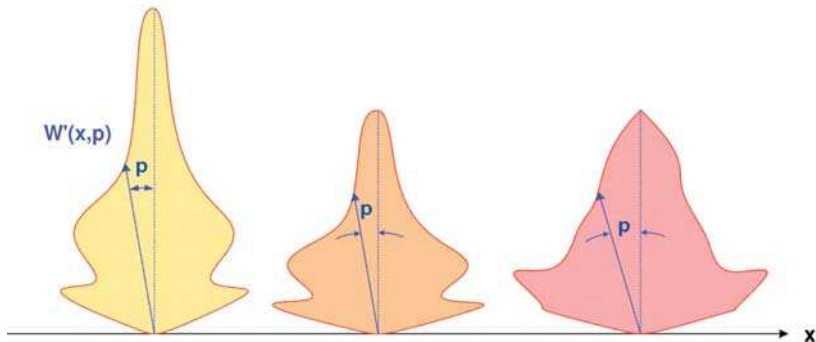


Figure 19-17: Schematic presentation of the Wigner distribution function of a light source as a local distribution of the energy flow in a definite direction in one dimension.

Figure 19-18 shows the geometrical-optical phase space density and the Wigner distribution function in the object and image planes for the imaging of a 10λ -slit with coherent and incoherent illumination. The x -axis is pointing to the right and

the frequency axis upwards. The geometrical-optical phase space densities in the left part of figure 19-18 are limited in the direction of the x -axis by the width of the slit and in the direction of the ν -axis by the illumination aperture with $\nu = \sigma NA / \lambda$. For incoherent illumination with $\sigma = 1$ the frequency spectrum is limited at the end by the aperture $NA = 1$. The Wigner distribution function, however, exhibits more diffraction effects in the object plane due to the diffraction at the edge of the slit opening, which is more distinctive with coherent illumination. The effect of low-pass filtering of the optical system becomes noticeable in a low pass filtering, i.e. in a blocking of the higher propagation angles. At the same time, this leads to fading at the edges of the diffracting structures in the x -direction.

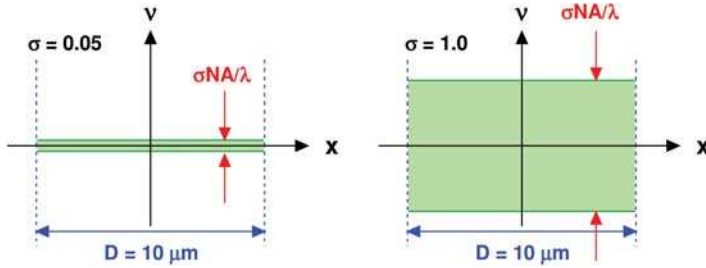


Figure 19-18: Illustration of the coordinates of the Wigner distribution function for imaging of a 10λ -slit with coherent illumination and $\sigma = 0.05$, and with incoherent illumination and $\sigma = 1$.

In figure 19-19 the numerically calculated Wigner distribution functions corresponding to the case of figure 19-18 are shown. The diffraction fringes can be seen clearly. Figure 19-20 shows the Wigner distribution function of an ideal point image.

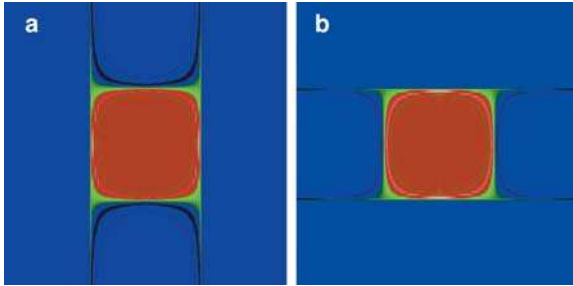


Figure 19-19: Wigner distribution function of a 10λ -slit upon incoherent imaging: a) shows the object and b) the image plane.

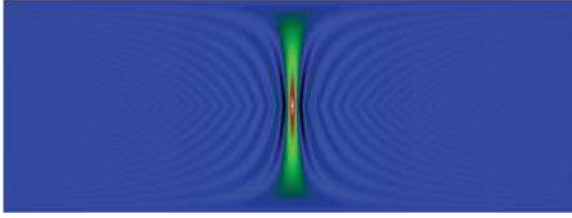


Figure 19-20: Wigner distribution function $W(x,\nu)$ of an ideal point image.

19.2.3

Moments of the Wigner Distribution Function

Measurable physical quantities are obtained by integration of the density function W' . The integral over the spatial coordinate \vec{r} gives the direction spectrum of the intensity

$$I(\vec{p}) = \int W'(\vec{r}, \vec{p}) d\vec{r} \quad (19-45)$$

and, correspondingly, the integration over the direction p yields the spatial distribution of the intensity

$$I(\vec{r}) = \frac{1}{(2\pi)^2} \cdot \int W'(\vec{r}, \vec{p}) d\vec{p} . \quad (19-46)$$

The integral over r and p gives the total power of a field

$$P = \frac{1}{(2\pi)^2} \cdot \iint W'(\vec{r}, \vec{p}) d\vec{r} d\vec{p} = \int I(\vec{r}) d\vec{r} . \quad (19-47)$$

For any arbitrary function f one can define an average value in terms of the radiation field

$$\langle f(\vec{r}) \rangle = \frac{\iint f(\vec{r}) W'(\vec{r}, \vec{p}) d\vec{r} d\vec{p}}{\iint W'(\vec{r}, \vec{p}) d\vec{r} d\vec{p}} . \quad (19-48)$$

The density W' plays here the role of a weighting function. The spatial and angular extents of a beam are given by the second moments for position and angle:

$$\langle x^2 \rangle = \frac{1}{P} \cdot \iint x^2 \cdot W'(x, p) dx dp , \quad (19-49)$$

$$\langle p^2 \rangle = \frac{1}{P} \cdot \iint u^2 \cdot W'(x, p) dx dp , \quad (19-50)$$

with the total power of a beam

$$P = \iint W'(x, p) dx dp . \quad (19-51)$$

The general formulation finds an important application in the definition of higher moments, which describe e.g. the quality of laser beams.

19.2.4

Smoothing of the Wigner Distribution Function and Diffraction Focus

The negative values of the original Wigner distribution function arise from the minimum spatial extend of a wave, i.e. from the non-local character of a wave. Negative values corresponds to interference effects between distant points and arise only in partial coherent or coherent wave fields. These negative values occur as a consequence of the fact that the Wigner distribution function is a bilinear function. The purely geometrical-optical description by rays is based on the use of delta functions and consequently only local effects are included, excluding interaction of light rays at different positions or with different directions, respectively.

In the general case of an arbitrary, coherent or partially coherent radiation, it is always possible to expand the Wigner distribution function in Gaussian functions and represent it as a sum of individual Gaussian modes. The Wigner distribution function of a coherent superposition of two modes is given by

$$W'_{sum}(x, p) = W'_1(x, p) + W'_2(x, p) + W'_{int}(x, p) \quad (19-52)$$

with an interference term W'_{int} . The interference term can be negative, while the Wigner distribution function of a single Gaussian beam is always positive.

As shown above, the conventional Wigner distribution function is a locally averaged quasi-probability in the phase space and can also take negative values [19-14]. Integrating the Wigner distribution function W' over the elementary volume of the phase space with a weight function, results in the averaged and smoothed, purely positive, modified Wigner distribution function Q , which can be directly interpreted as a probability.

$$Q(\vec{r}, \vec{p}) = \iint W'(\vec{r}, \vec{p}) \cdot e^{-\left(\frac{\vec{r}-\vec{r}'}{a}\right)^2 - \left(\frac{\vec{p}-\vec{p}'}{b}\right)^2} d\vec{r}' d\vec{p}' . \quad (19-53)$$

Only the Q -function can be measured, it is a true density and has always positive values. By the Gaussian averaging kernel it takes into account the uncertainty principle, since from quantum theory the exact measurement of conjugated quantities is forbidden. The exact and simultaneous determination of x and p is impossible within the area in phase space specified by a coherent Gaussian basis mode, the averaging kernel. For this reason, certain fine structures of the Wigner distribution function at the resolution limit cannot be obtained by measurements. The convolution leading to the Q -function smoothes these effects in addition to the noise which is intrinsic to the measurement.

In a more general formulation, the convolution with arbitrary smoothing functions G reads

$$Q(x, p) = \iint W'(x, p) \cdot G_x(x - x') \cdot G_p(p - p') dx' dp' . \quad (19-54)$$

In the case of Gaussian functions one has

$$G_x(x - x') = \frac{1}{\sqrt{\Delta x}} \cdot e^{-2\left(\frac{x-x'}{\Delta x}\right)^2} , \quad (19-55)$$

$$G_p(p - p') = \frac{1}{\sqrt{\Delta p}} \cdot e^{-2\left(\frac{p-p'}{\Delta p}\right)^2} . \quad (19-56)$$

In the case of minimal smoothing effect by the averaging kernel, the uncertainty principle requires that the following relationship is satisfied

$$\Delta x \cdot \Delta p = \frac{\lambda}{\pi} . \quad (19-57)$$

The width of the averaging kernel in spatial and angular coordinates may be different as long as eq. (19-57) is satisfied. One can determine the position at a higher precision at the expense of angular resolution and vice versa. The averaging kernel thus takes in general the shape of an elliptical area. A simple example are illustrated schematically in figure 19-21.

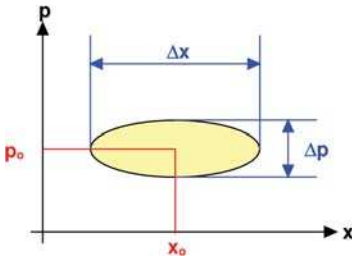


Figure 19-21: Schematic presentation of the integration of the Wigner distribution function over the elementary phase space area.

19.2.5

Wigner Distribution Function of Coherent Fields

In the case of coherent quasi-monochromatic radiation fields, the Wigner distribution function can be directly obtained from the field amplitude, without the coherence function, by

$$W'(x, p) = \int U\left(x + \frac{\Delta x}{2}\right) \cdot U^*\left(x - \frac{\Delta x}{2}\right) \cdot e^{-ik_0 \Delta x \cdot p} d\Delta x . \quad (19-58)$$

Thus the Wigner distribution function is a bilinear function of the complex field amplitude. In the special case of coherent wave fields it is possible to obtain the field amplitude directly by reversal of eq. (19-58) with

$$U(x) = \frac{1}{\lambda \cdot U^*(0)} \cdot \int W' \left(\frac{x}{2}, p \right) \cdot e^{ik_0 p x} dp . \quad (19-59)$$

It is obvious that the complex field cannot be reconstructed from the Wigner distribution function since with unknown $U(0)$ at least the absolute value of the phase remains undefined.

Frequently, coherent waves or beams are characterized at any location x by a single definite direction p . Thus the description of a coherent beam by the two-dimensional Wigner distribution function $W'(x, p)$ can be reduced to a one-dimensional description by amplitude A and phase Φ . With

$$U(x) = A(x) \cdot e^{i\Phi(x)} \quad (19-61)$$

the direction of the wave at position x is given by the derivative of the phase function [19-15]

$$p(x) = \frac{\lambda}{2\pi} \cdot \frac{d\Phi(x)}{dx} \quad (19-62)$$

and with the help of Dirac's delta function $\delta(p)$ one obtains for the Wigner distribution function

$$W'(x, p) = A^2(x) \cdot \delta \left[p - \frac{\lambda}{2\pi} \cdot \frac{d\Phi(x)}{dx} \right] . \quad (19-63)$$

Since the Wigner distribution function is a bilinear function, the superposition of two coherent fields U_1 and U_2 results in an interference term, considering the interference effects between the two fields

$$W'_{\text{sum}}(x, p) = W'_1(x, p) + W'_2(x, p) + W'_{\text{int}}(x, p) \quad (19-64)$$

with

$$\begin{aligned} W'_{\text{int}}(x, p) = & \int U_1 \left(x + \frac{\Delta x}{2} \right) \cdot U_2^* \left(x - \frac{\Delta x}{2} \right) \cdot e^{ik_0 \Delta x \cdot p} d\Delta x \\ & + \int U_2 \left(x + \frac{\Delta x}{2} \right) \cdot U_1^* \left(x - \frac{\Delta x}{2} \right) \cdot e^{ik_0 \Delta x \cdot p} d\Delta x . \end{aligned} \quad (19-65)$$

19.2.6

Ambiguity Function

For a complex coherent field $U(x)$ the so-called ambiguity function can be defined, which depends on the spatial difference $\Delta \vec{r}$ and the angular coordinate

$$\vec{s} = \sin \vec{w} = \lambda \cdot \vec{v} . \quad (19-66)$$

It follows for the ambiguity function

$$A'(\vec{s}, \Delta\vec{r}) = \int U\left(\vec{r} + \frac{\Delta\vec{r}}{2}\right) \cdot U^*\left(\vec{r} - \frac{\Delta\vec{r}}{2}\right) \cdot e^{-i\vec{k}_0 \cdot \vec{r} \cdot \vec{s}} d\vec{r} . \tag{19-67}$$

When the wave field is partially coherent and the field amplitude can be considered to be a statistical quantity, an analogous relation holds in terms of the coherence function

$$A'(\vec{s}, \Delta\vec{r}) = \int J(\vec{r}, \Delta\vec{r}) \cdot e^{-i\vec{k}_0 \cdot \vec{r} \cdot \vec{s}} d\vec{r} . \tag{19-68}$$

Hence, the ambiguity function is given by a one-dimensional Fourier transform of the coherence function with respect to the spatial center coordinate \vec{r} . The intensity can be obtained from the ambiguity function for $\Delta\vec{r} = 0$ according to

$$I(\vec{r}) = \int A'(\vec{s}, 0) \cdot e^{-i\vec{k}_0 \cdot \vec{r} \cdot \vec{s}} d\vec{s} . \tag{19-69}$$

Furthermore, a two-dimensional Fourier transform relates the ambiguity function to the Wigner distribution function

$$A'(\vec{s}, \Delta\vec{r}) = \int W'(\vec{r}, \vec{p}) \cdot e^{i\vec{k}_0 \cdot (\vec{p} \cdot \Delta\vec{r} - \vec{r} \cdot \vec{s})} d\vec{r} d\vec{p} . \tag{19-70}$$

In a similar way to the Wigner distribution function, the ambiguity function can be defined with the spatial frequency $\Delta\nu$ (instead of the angle s) as an argument

$$A(\Delta\vec{\nu}, \Delta\vec{r}) = \int W(\vec{r}, \vec{\nu}) \cdot e^{i2\pi \cdot (\vec{\nu} \cdot \Delta\vec{r} - \vec{r} \cdot \Delta\vec{\nu})} d\vec{r} d\vec{\nu} . \tag{19-71}$$

The ambiguity function for ideal imaging thus corresponds to the Wigner distribution function of a point image in one transverse dimension, rotated by 90° and rescaled. Figure 19-22 illustrates the coordinate transformation according to eq. (19-68) applied to the coherence transfer function $\gamma(\nu_1, \nu_2)$. The Fourier transform along the vertical ν -axis gives the ambiguity function $A(\Delta\nu, \Delta x)$.

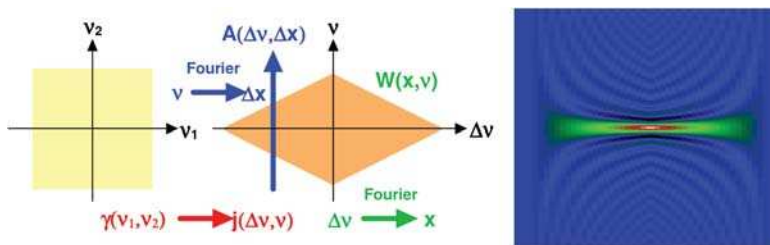


Figure 19-22: Ambiguity function $A(\Delta\nu, \Delta x)$ of an imaging system obtained from the coherence transfer function after coordinate transformation ($\Delta\nu$ -axis pointing to the right, Δx -axis upwards).

19.2.7

The Characterizing Functions in their Context

The relations and the different definitions of the functions describing the coherence are shown for comparison in figure 19-23.

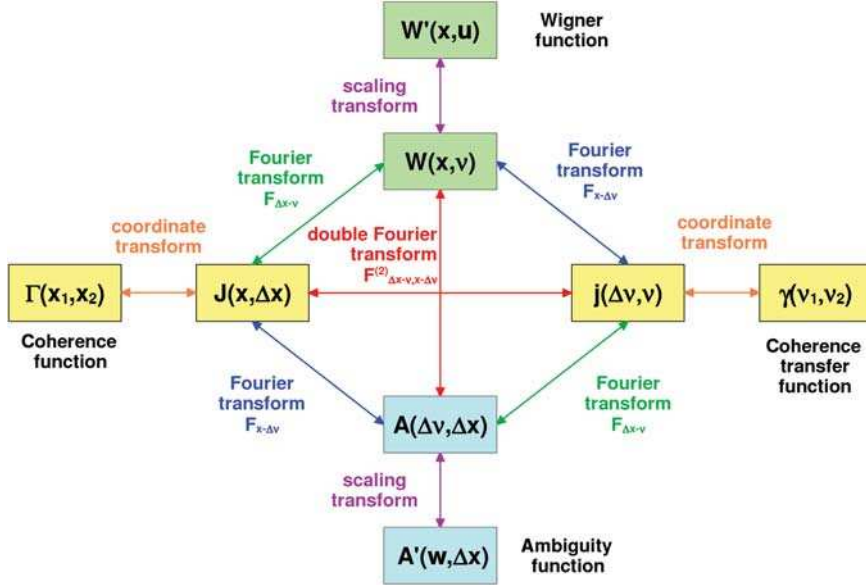


Figure 19-23: Correlations between the functions: Fourier relations, coordinate transformations and scaling transformations.

In one dimension, the following Fourier relations hold between the various functions:

$$W(x, v) = \int J(x, \Delta x) \cdot e^{-2\pi i \Delta x \cdot v} d\Delta x, \tag{19-72}$$

$$A(\Delta v, \Delta x) = \int J(x, \Delta x) \cdot e^{-2\pi i x \cdot \Delta v} dx, \tag{19-73}$$

$$j(\Delta v, v) = \iint J(x, \Delta x) \cdot e^{2\pi i (x \cdot \Delta v - \Delta x \cdot v)} dx d\Delta x, \tag{19-74}$$

$$W(x, v) = \iint A(\Delta v, \Delta x) \cdot e^{2\pi i (x \cdot \Delta v - \Delta x \cdot v)} d\Delta x d\Delta v. \tag{19-75}$$

In addition, the coordinate transform to the separated coordinates x_1, x_2 and v_1, v_2 respectively, connects J with Γ and j with γ . Finally, the scaling from the spatial frequencies $v, \Delta v$ to the angle coordinates p, s delivers the functions W' and A' .

19.3

Temporal Coherence

19.3.1

Superposition of Signals with Different Frequency

As has been shown in section 19.1, the principles of coherence can be understood to a great extent even on the basis of a purely qualitative consideration. The exact mathematical description and, especially, the treatment of coherence in optical imaging, requires a more detailed treatment and a generalization of the mutual coherence function. The superposition of two monochromatic signals of different color

$$U_1(r, t) = A_1 e^{i(k_1 r - \omega_1 t + \varphi_1)} \quad (19-76)$$

and

$$U_2(r, t) = A_2 e^{i(k_2 r - \omega_2 t + \varphi_2)} \quad (19-77)$$

yield, after temporal averaging:

$$\begin{aligned} I(r, t) &= \langle |U_1 + U_2|^2 \rangle_T \\ &= I_1 + I_2 + 2 \frac{1}{T} \int_0^T A_1 A_2 \cos [\Delta k_{12} r - \Delta \omega_{12} (t + \tau) + \Delta \varphi_{12}] d\tau . \end{aligned} \quad (19-78)$$

As an example, the interference of two signals with a fixed phase coupling ($\Delta \varphi = \text{const.}$) is considered. With a integration time T of the detection process, the absolute value of the mutual coherence function is given by a sinc-function with an argument $\Delta \omega \cdot T$ with the frequency difference $\Delta \omega = \omega_1 - \omega_2$:

$$|\gamma(U_1, U_2)| = \frac{1}{T} \int_0^T \cos [\Delta \omega (t + \tau)] d\tau = \text{sinc}(\Delta \omega \cdot T). \quad (19-79)$$

As already established, the contrast of the interference fringes decreases due to the finite integration time of the detector with increasing frequency separation $\Delta \omega$. The frequency difference, for which the interference contrast vanishes as a result of the averaging over one period of the running interference fringes, is given by the first zero of the sinc-function in eq. (19-79).

The coherence function γ corresponds to the correlation of two signals and vanishes for uncorrelated signals, i.e., when, for example, the phases φ exhibit statistical fluctuations as happens in thermal light sources. Interference occurs in the case of uncorrelated light sources only if the two signals passing through the two apertures are emitted by the same source point, as was shown in Young's double slit arrangement.

Two monochromatic amplitudes U_1 and U_2 behind the slit apertures, assumed to be very small, emitted by a single monochromatic source point S_1 , are still fluctuat-

ing but possess a fixed relative phase difference $\Delta\varphi(S_1)$ for small path length differences. Thus the differences of the statistically fluctuating phases are constant for a given wavelength. One obtains:

$$\Gamma(U_1, U_2) = A_1(S_1)A_2(S_1) \cos [\Delta kr - \Delta\varphi(S_1)] . \quad (19-80)$$

A fixed phase relation is obtained likewise for a different source point S_2 , however, the phase difference $\Delta\varphi(S_2)$ is modified as a consequence of the different optical paths. The resulting intensity distribution is determined by summation or integration over all source points. This will be considered in more detail in the section 19.4 about spatial coherence.

19.3.2

Spectral Distribution of a Light Source

The light emission by e.g. an atom occurs in reality only for a finite time Δt . The resulting wave train has a finite length. Since it is not an infinitely extended periodic function it is represented by a finite spectrum according to

$$E(t) = \int A(\omega) \cdot e^{-i\omega t} d\omega . \quad (19-81)$$

$A(\omega)$ above denotes the spectral amplitude density. A wave train of finite length should not, in principle, be monochromatic. For a hard rectangular truncation of a sine-wave one has, e.g.,

$$A(\omega) = \frac{\sin(\omega \cdot \Delta t)}{\omega \cdot \Delta t} . \quad (19-82)$$

The width of this spectral distribution (the separation of the zeros) is given by

$$\Delta\omega = \frac{2\pi}{\Delta t} . \quad (19-83)$$

The shorter the wave train the broader the corresponding spectral width. The decomposition of a finite wave train into its spectral components is schematically illustrated in figure 19-24.

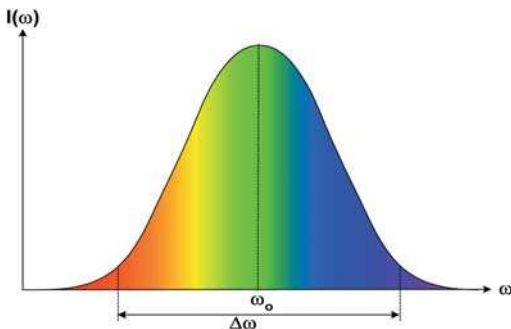


Figure 19-24: Chromatic representation of the spectrum of a light source.

19.3.3

Bandwidth-limited Signals

Bandwidth-limited signals are often found, e.g., in the spectra of sources dominated by spectral lines like gas discharge sources, or by restricting the detected wavelength range by a band-gap filter given by the transmission properties of optical materials or the sensitivity of the detector.

As an example, the interference of two bandwidth-limited signals is considered, i.e., interference of signals with finite spectral extent. The averaging as a result of the measurement is expressed by the function $M(t)$, e.g., $M(t) = 1/T \cdot \text{rect}(t/T)$. The geometrical factor $\Delta k \cdot r - \Delta \varphi$ can be represented in general by a time difference τ . One then obtains for the dependence of the coherence function on the time difference between the two signals,

$$\Gamma_{nm}(\tau) = \langle |U_n U_m^*| \rangle_T = \int U_n(t + \tau) U_m^*(t) M(t) dt . \quad (19-84)$$

Substitution of the spectrum and reordering yields

$$\begin{aligned} \Gamma_{nm}(\tau) &= \iiint u_n(\omega') e^{-i\omega'(t+\tau)} d\omega' \cdot u_m^*(\omega) e^{+i\omega t} d\omega \cdot M(t) dt \\ &= \iint u_n(\omega') u_m^*(\omega) e^{-i\omega\tau} \cdot \int M(t) e^{i(\omega-\omega')t} dt d\omega d\omega' . \end{aligned} \quad (19-85)$$

Since the integration time is in general much longer than the oscillation period $1/\nu_t$ the time integral can be approximated by a delta-function and one obtains

$$\Gamma_{nm}(\tau) = \int u_n(\omega) \cdot u_m^*(\omega) \cdot e^{-i\omega\tau} d\omega . \quad (19-86)$$

The product of the two spectra in eq. (19-86) is referred to as the mutual spectral density:

$$S_{nm}(\omega) = u_n(\omega) \cdot u_m^*(\omega) . \quad (19-87)$$

Hence, the mutual coherence function Γ_{nm} is given by the Fourier transform of the mutual spectral density S_{nm} , also known as the cross-spectral density, which is in general a function of the spatial coordinates

$$S(\vec{r}_1, \vec{r}_2, \omega) = \int_{-\infty}^{\infty} \Gamma(\vec{r}_1, \vec{r}_2, \tau) \cdot e^{-i\omega\tau} d\tau . \quad (19-88)$$

For identical signals and coinciding positions $r_1 = r_2$, the cross spectral density is given by the auto-correlation of the complex signal and is also called the power spectral density (PSD)

$$S(\omega) = S(\vec{r}, \vec{r}, \omega) = S(\vec{r}, \omega) = S_{nn}(\omega) = |u_n(\omega)|^2 . \quad (19-89)$$

The intensity of a multi-spectral field can be computed as an integral over the power spectral density $S(\omega)$ in the form

$$I = \int_0^{\infty} S(\omega) d\omega . \tag{19-90}$$

Real light sources can be described in an approximate manner by analytical expressions for the spectral function. Several such model functions are included in table 19-1.

Table 19-1: Spectral distribution functions of practical importance.

Distribution shape	Spectral function	Spectral width
Gaussian Gas lamp, Doppler effect	$S(\omega) = \tau_c \cdot e^{-\tau_c^2 \omega^2}$	$\Delta\omega_{\text{fwhm}} = \frac{\ln 2}{\tau_c}$
Top hat	$S(\omega) = \begin{cases} 1 & \omega \leq \omega_0 - \frac{\Delta\omega}{2} \\ 0 & \text{otherwise} \end{cases}$	$\Delta\omega_{\text{fwhm}} = \frac{1}{\tau_c}$
Lorentz Collisional line-broadening in gas light sources	$S(\omega) = \frac{\Delta\omega}{2\pi} \cdot \frac{1}{(\omega - \omega_0)^2 + (\frac{\Delta\omega}{2})^2}$	$\Delta\omega_{\text{fwhm}} = \frac{1}{\pi \tau_c}$
Exponential	$S(\omega) = \tau_c \cdot e^{-\tau_c(\omega - \omega_0)}$	$\Delta\omega = \frac{1}{\tau_c}$

With $n = m$ from eq. (19-86) one obtains the Wiener-Khinchin theorem according to which the autocorrelation of a signal is given by the Fourier transform of the power spectral density S_{nn}

$$\Gamma_{nn}(\tau) = \int |u_n(\omega)|^2 e^{-i\omega\tau} d\omega = \hat{F}\{S_{nn}(\omega)\} . \tag{19-91}$$

In a simplified form this can be written as

$$\Gamma(\tau) = \int_{-\infty}^{\infty} S(\omega) \cdot e^{-i\omega\tau} d\omega . \tag{19-92}$$

The inverse Fourier transform gives the power spectrum as the Fourier transform of the temporal coherence function

$$S(\omega) = \int_{-\infty}^{\infty} \Gamma(\tau) \cdot e^{i\omega\tau} d\tau . \tag{19-93}$$

19.3.4

Axial Coherence Length

As a consequence of the Wiener-Khinchin theorem, a spectrally broad light source exhibits a very short coherence time, and vice versa. Since the temporal and spectral widths of two functions related by a Fourier transformation are inversely proportional, the coherence time and the spectral line-width of a light source are inversely proportional

$$\tau_c = \frac{2\pi}{\Delta\omega} . \quad (19-94)$$

The interpretation of this equation is shown in figure 19-25. As a result of the statistical emission processes, independent wave trains of length l are emitted (figure 19-25, right hand side). The length l of a wave train is referred to as the coherence length of a light source. Spectral width $\Delta\omega$ and temporal length l of a wave train are given by eq. (19-94) in equivalence to the uncertainty principle.

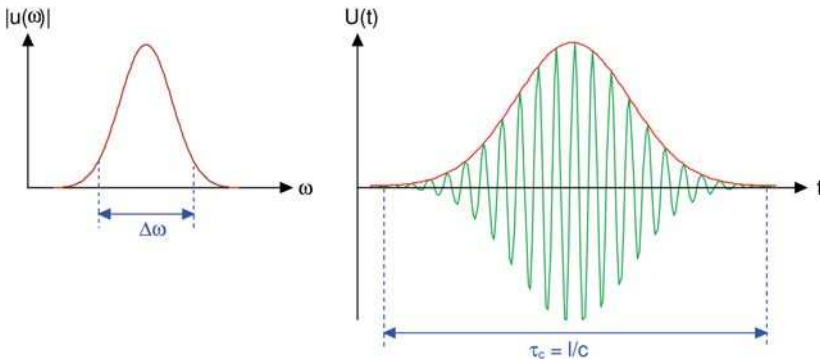


Figure 19-25: Schematic illustration of the spectrum of a bandwidth-limited signal (left) and a coherent wave train (right).

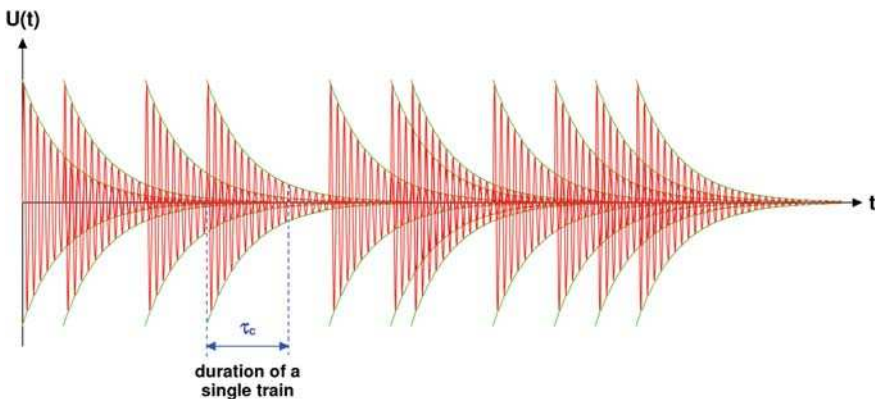


Figure 19-26: Schematic illustration of the emission of single wave trains with duration τ_c .

Assuming that the initial phases of the individually emitted wave trains fluctuate statistically and that each single wave exhibits a decay time of τ_c , it follows that τ_c is precisely equal to the coherence time of the radiation. This scenario is illustrated in figure 19-26.

The application of Parseval's theorem to the Fourier pair of the Wiener-Khinchin theorem yields the relationship between the power spectral density and the coherence time

$$\tau_c = \int_{-\infty}^{\infty} |I(\omega)|^2 d\omega . \quad (19-95)$$

An alternative definition of the coherence time is based on the average of the square of the coherence function

$$\tau_c = \int_{-\infty}^{\infty} |\gamma(\tau)|^2 d\tau . \quad (19-96)$$

The axial coherence length is obtained simply as a product of the coherence time and the wave propagation speed c

$$l_c = c \cdot \tau_c . \quad (19-97)$$

The coherence length of a non-monochromatic wave may be illustrated at the example of two waves of different wavelengths λ_1 and λ_2 . Starting with coinciding phases of the two plane waves, the same phase is obtained at a distance l_c proportional to the frequency difference $\Delta\omega$

$$l_c = c \cdot \tau_c = \frac{2\pi \cdot c}{\Delta\omega} . \quad (19-98)$$

The distance l_c in eq. (19-98) corresponds to the longitudinal coherence length of (19-97). The superposition of two waves results in beating, the maxima of which are separated by l_c . This situation is illustrated in figure 19-27. The interference of the two waves is a harmonic wave of frequency $\Delta\omega$. The detected contrast vanishes for multiples of the integration time τ_c , corresponding to multiples of the coherence length l_c .

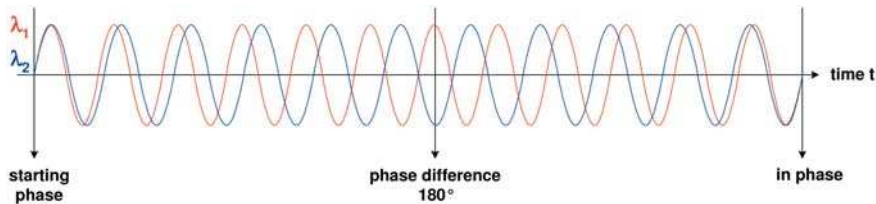


Figure 19-27: Temporal coherence between different wavelengths.

Table 19-2: Typical bandwidths, coherence times, and axial coherence lengths of some common light sources.

Light source	τ_c	l_c
Incandescent lamp	8.3 fs	2.5 μm
Hg-lamp	2 ps	600 μm
Hg-high-pressure lamp, 546 nm line	67 fs	20 μm
Hg-low-pressure lamp, 546 nm line	200 ps	6 cm
Kr-isotope lamp, 606 nm line	2.3 ns	70 cm
LED	67 fs	20 μm
Multimode HeNe-laser, cavity length $L=1$ m	0.67 ns	20 cm
Single frequency HeNe-laser	1 μs	300 m

To summarize, in order to interfere with itself, the time difference τ should not exceed the wave train length divided by the propagation speed c , i.e.:

$$\tau < \frac{1}{\Delta\omega} = \frac{l}{c} \quad (19-99)$$

with the coherence length determined according eq. (19-98) by

$$l \approx \frac{2\pi \cdot c}{\Delta\omega} \approx \frac{\lambda^2}{\Delta\lambda} \quad (19-100)$$

with the frequency difference approximated by

$$\Delta\omega = \frac{2\pi \cdot c}{\lambda^2} \cdot (\lambda_2 - \lambda_1) . \quad (19-101)$$

Example: an excimer laser at 193 nm possesses a bandwidth of 0.35 pm. Hence, its coherence length is given by

$$l \approx \frac{193 \text{ nm}^2}{0.35 \text{ pm}} = \frac{3.7 \cdot 10^{-14} \text{ m}^2}{0.35 \cdot 10^{-12} \text{ m}} = 10.6 \text{ cm} . \quad (19-102)$$

The coherence length l , especially in the case of monochromatic imaging, is substantially longer than the path length differences which normally occur in the imaging systems. In the ideal case the latter possess equal optical lengths for all light paths (conservative or orthotomic systems). In order to avoid disturbing interference effects it is therefore advantageous deliberately to introduce path length differences for the different light paths in the illumination systems, which minimizes the capability for temporal interference.

19.3.5
Thermal Light Sources

The coherence function of narrow-band thermal light sources as spectral or discharge lamps is dominated by spectral broadening due to collisions. The resulting field distribution of N independent light sources is given by

$$U(t) = \sum_i U_i \cdot e^{-i\omega_0 t + \Phi_i} \tag{19-103}$$

The coherence function for a single emission process can be specified as

$$\begin{aligned} \langle U_i(t) U_i^*(t + \tau) \rangle_T &= U_i^2 e^{-i\omega_0 \tau} \cdot \langle e^{i[\Phi(t) - \Phi(t+\tau)]} \rangle_T = U_i^2 e^{-i\omega_0 \tau} \int_{t-\tau}^{\infty} p(t) dt \\ &= U_i^2 e^{-i\left(\omega_0 \tau + \frac{\tau}{T_i}\right)} \end{aligned} \tag{19-104}$$

with the probability

$$p(t) = \frac{1}{T_i} e^{-\frac{t}{T_i}} \tag{19-105}$$

where T_i corresponds to the average time-of-flight of a free atom in the distribution of the source.

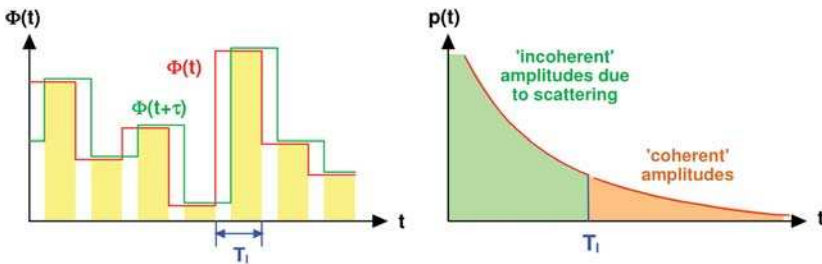


Figure 19-28: Statistical distribution of the wave trains occurring as a result of emission processes interrupted by collisions.

Figure 19-28 shows schematically the phase distribution $\Phi(t)$ versus time; the light emitted between two collisions has the same phase while collisions cause abrupt phase changes. The probability that no collisions have taken place and two signals can still interfere, decreases with the time t and this is expressed in the exponential decay of $p(t)$.

The power spectrum $S(\omega)$ is given by the Fourier transform of the coherence function and its shape corresponds to the so-called Lorentz spectrum:

$$u(\omega) \sim \frac{\rho}{(\omega_0 - \omega)^2 + \rho^2} \tag{19-106}$$

with the “damping constant” $\rho = 1/T_0$.

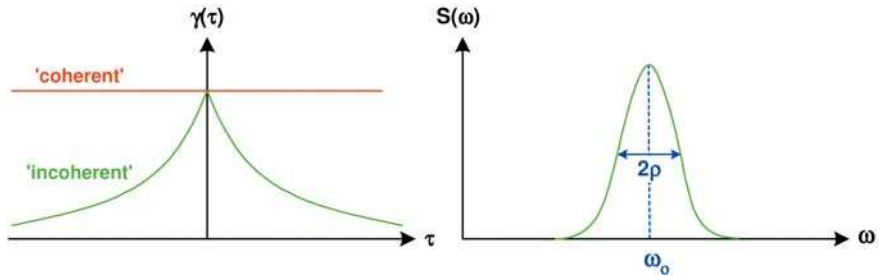


Figure 19-29: Coherence function $\gamma(\tau)$ and power spectrum $S(\omega)$ with a Lorentz shape.

An ion in a plasma experiences about 1000 collisions in the process of emission. Typical average times-of-flight for free ions amount to $\approx 10^{-11}$ s which results in a spectral bandwidth ρ of about 10^{11} /s corresponding to a coherence length of ≈ 3 mm. In contrast, the “lifetime” of an emission process in a laser with a cavity length of 30 cm typically amounts to $T_0 \sim 10^{-8}$ s which gives a coherence length of about 3 m.

It should be mentioned that an additional spectral broadening occurs as a result of the Doppler shift of the power spectrum, which can be taken into account by the Maxwell distribution of the ion velocities in the plasma. The Doppler broadening leads to a power spectral density of Gaussian shape.

19.3.6

Temporal Coherence in the Michelson Interferometer

In the Michelson interferometer a beam splitter divides the amplitude into two parts to form signal and reference. The experimental setup is sketched in figure 19-30.

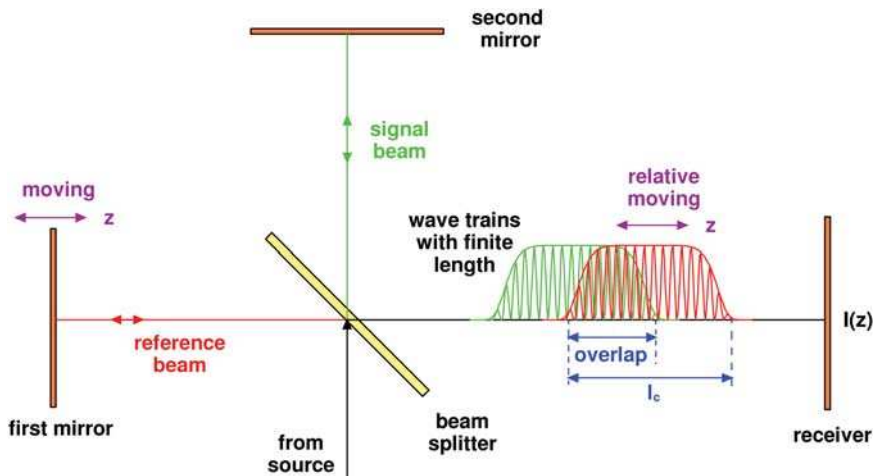


Figure 19-30: Arrangement of a Michelson interferometer.

By variation of the path length in either the reference or signal path of the Michelson interferometer, the temporal coherence function can be determined. Different lengths of the interferometer arms result in temporal shifting of the wave trains along the propagation direction. With increasing path length difference, this leads in general to a contrast loss. Vanishing contrast is obtained for path length differences larger than the coherence length l_c . The spectral distribution of the light source can be obtained by so-called Fourier spectroscopy. Writing for the detected intensity as a function of path length difference z

$$I(z) = \sum_m I_m + 2 \sum_{n < m} \sqrt{I_n \cdot I_m} \cdot \cos[\varphi_n(z) - \varphi_m(z)], \quad (19-107)$$

where the phase difference is proportional to the path difference

$$\Delta\varphi = \varphi_n - \varphi_m = 2 \Delta k \cdot z = \frac{4\pi \cdot \Delta\lambda \cdot z}{\lambda^2}, \quad (19-108)$$

the power spectral density of the signal is obtained by Fourier-transformation of $I(z)$. The behavior of the signal $I(z)$ as a function of the path difference z is shown in figure 19-31.

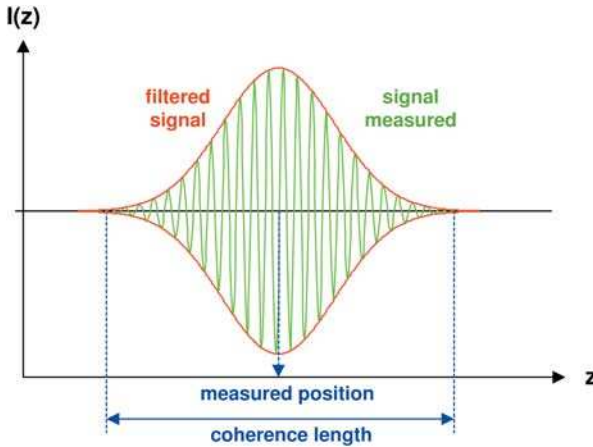


Figure 19-31: Signal from a Michelson interferometer.

19.4 Spatial Coherence

19.4.1 Introduction

Uncorrelated light from a source S can interfere only through signals emitted by the same source points. In this case, the differences of the statistically fluctuating phases, for given path length differences within the coherence length of the light source, are also constant (cf. figure 19-32).

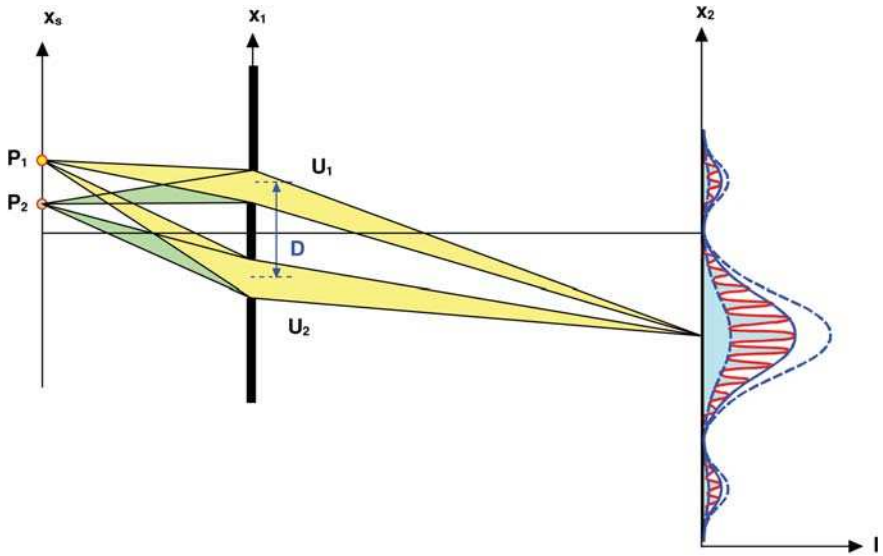


Figure 19-32: Two correlated signals U_1 and U_2 from an extended light source can be produced, e.g., by a double slit. An interference pattern with visibility V is seen on a distant screen.

Although, for a monochromatic source point P_1 , the two amplitudes U_1 and U'_1 behind the slit apertures are fluctuating, they have a fixed relative phase difference $\Delta\varphi(S_1)$ determined by the path length difference. One obtains:

$$\Gamma(U_1, U'_1) = A_1(S_1)A'_1(S_1) \cos [\Delta kr - \Delta\varphi(S_1)]. \quad (19-109)$$

For another source point P_2 , again a fixed phase relation is obtained. However, the modified path length difference leads to a different phase difference $\Delta\varphi_2$. The mutual coherence function of the signals coming from different source points P_1 and P_2 of a thermal light source, vanishes:

$$\Gamma(U_1, U'_2) = 0. \quad (19-110)$$

This consideration can be extended to the more general case of the superposition of many signals. In the general case, the intensity may be written as a double sum or integral over all light source points

$$I(r, t) = \sum_{n,m} \sqrt{I_n I_m} |\gamma(U_n, U_m)| \cos(\arg \gamma(U_n, U_m)). \quad (19-111)$$

An extended thermal light source can be considered as consisting of many mutually incoherent emitting light source points, whereby all mixed terms in eq. (19-111) vanish. The visibility V or the maximum possible contrast of the interference fringes is given in general by

$$V = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} = \frac{2 \sum_{n < m} \sqrt{I_n I_m} |\gamma(U_n, U_m)|}{\sum_m I_m} \quad (19-112)$$

For a finite-sized light source, the spatial distribution of the coherence function indicates the limits of the spatial region where the phases of the waves are still correlated. Figure 19-33 illustrates a source of finite extension and two points in a detection plane. Even if the source elements $d\sigma$ are strictly uncorrelated, the resulting complex amplitudes in image points P_1 and P_2 may be correlated. This is represented by coherence function γ with a lateral diameter L_c . A simplified example of the coherence function is schematically shown in figure 19-34.

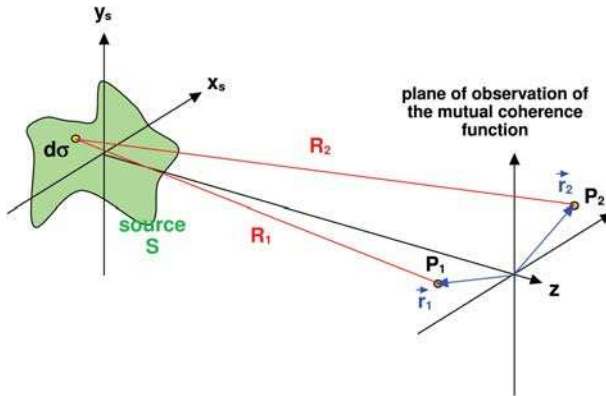


Figure 19-33: The mutual coherence function of two points P_1 and P_2 is determined by the source size.

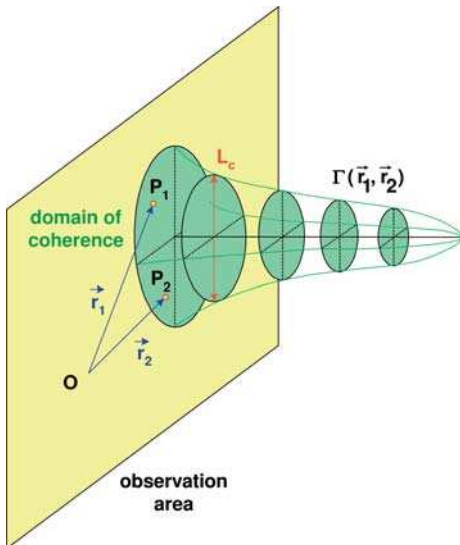


Figure 19-34: Spatially coherent domain within the illuminated surface.

19.4.2

Propagation of the Coherence Function

The coherence properties of a wave field in general change with the light propagation. The spatial degree of coherence increases with the distance z . For a given source size, the effective source area decreases with increasing distance between source and the detection points. From another point of view, illustrated in figure 19-35, the size of the common area in the source plane, which contributes to both of the detection points P_1 and P_2 , increases while the source size in the source plain remains constant. If the source is smaller compared to the common area, the mutual coherence at points P_1 and P_2 will be high and vice versa.

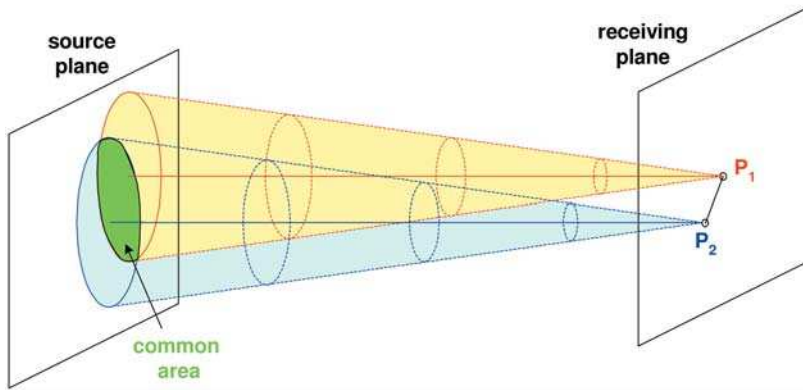


Figure 19-35: Change in the coherence state of light emitted by a finite source with the distance.

The mutual coherence at points P_1 and P_2 is thus determined by the visible angle of the source, i.e. the effective source size, which decreases with increasing distance. For a rough estimation, the following approximation can be made. At some distance an incoherent light source with dimension $2a$ appears to be a partially coherent source. If the distance exceeds the value

$$z = \frac{a^2}{\lambda} \quad (19-113)$$

the source turns out to be coherent. This corresponds to a Fresnel number of the associated free-space propagation of smaller than 1. Therefore phase differences larger than $\lambda/2$ are not possible from different points of the light source and destructive interference cannot occur. Even a large light source can be considered to be coherent for large distances as long as the visible angle satisfies the relation

$$\theta \leq \frac{\lambda}{a} . \quad (19-114)$$

In a general consideration, the propagation of the coherence function follows from the linearity of the Maxwell equations and the Helmholtz equation for the complex field amplitude may be applied to the coherence function as well:

$$\nabla_j^2 \Gamma - \frac{1}{c^2} \cdot \frac{\partial^2 \Gamma}{\partial t^2} = 0, \quad j = 1, 2. \quad (19-115)$$

From a general formulation of the solution of Helmholtz transport equation with help of Greens function in three dimensions as propagator,

$$U_i(\vec{r}') = \iiint G(\vec{r} - \vec{r}') \cdot U_0(\vec{r}) d\vec{r}, \quad (19-116)$$

the generalized transport equation for the coherence function can be formulated and the Hopkins integral is obtained:

$$\Gamma_i(\vec{r}'_1, \vec{r}'_2) = \iiint \iiint G_o(\vec{r}_1, \vec{r}_2) \cdot G(\vec{r}_1 - \vec{r}'_1) \cdot G(\vec{r}_2 - \vec{r}'_2) d\vec{r}_1 d\vec{r}_2. \quad (19-117)$$

The intensity in the detection plane is given by the diagonal elements of the coherence function with $\vec{r}'_1 = \vec{r}'_2 = \vec{r}'$

$$I_i(\vec{r}') = \iiint \iiint G_o(\vec{r}_1, \vec{r}_2) \cdot G(\vec{r}_1 - \vec{r}') \cdot G(\vec{r}_2 - \vec{r}') d\vec{r}_1 d\vec{r}_2. \quad (19-118)$$

It should be noted, that in the absence of stationary conditions or when the coherence length of the light source is in the order of the path length differences in the optical system, the time should also be taken into account as well in Hopkins integral. Both the time dependence of the source field $U(\vec{r}, t)$ and the consideration of propagation time in the transferring optical system then have to be considered. For the complex amplitude, one obtains the representation

$$U_i(\vec{r}', t') = \iiint \iiint G(\vec{r} - \vec{r}', t - t') \cdot U_0(\vec{r}, t) d\vec{r} dt. \quad (19-119)$$

In the paraxial approximation and after insertion of the paraxial free-space transfer function for Green's function in the transverse coordinates only

$$G_{\text{paraxial}}(\vec{r}, \vec{r}') = e^{-\frac{ik}{2z}(\vec{r}-\vec{r}')^2} \quad (19-120)$$

the paraxial approximation for free-space propagation of the complex amplitude is obtained

$$U(\vec{r}', z) = \frac{-ik}{2\pi z} \iint U(\vec{r}, 0) \cdot e^{-\frac{ik}{2z}(\vec{r}-\vec{r}')^2} d^2\vec{r}, \quad (19-121)$$

where with plane source and detection planes the integration is reduced to two dimensions. From this the paraxial transport equation for the coherence function is obtained

$$\Gamma(\vec{r}'_1, \vec{r}'_2, z) = \left(\frac{k}{2\pi z} \right)^2 \cdot \iiint \iiint \Gamma(\vec{r}_1, \vec{r}_2, 0) \cdot e^{-\frac{ik}{2z}(\vec{r}_2 - \vec{r}'_2)^2 + \frac{ik}{2z}(\vec{r}_1 - \vec{r}'_1)^2} d^2\vec{r}_1 d^2\vec{r}_2. \quad (19-122)$$

The computation of the coherence function requires a double integration over the start planes. Accordingly, the calculation of the coherence propagation is in general a rather complex task.

19.4.3

Van Cittert-Zernike Theorem

In the case of incoherent light sources the coherence function of a light source in general can be simplified by

$$\Gamma_0(\vec{r}_1, \vec{r}_2) = I_0(\vec{r}_1) \cdot \delta(\vec{r}_1 - \vec{r}_2). \quad (19-123)$$

Insertion in the general transport equation (19-117) the Van Cittert-Zernike theorem for incoherent light sources is obtained:

$$\Gamma_i(\vec{r}_1', \vec{r}_2') = \iint I_0(\vec{r}) \cdot G(\vec{r} - \vec{r}_1') \cdot G(\vec{r} - \vec{r}_2') d^2\vec{r}. \quad (19-124)$$

The van Cittert-Zernike theorem considers the propagation of the coherence function of an incoherent light source through an arbitrary optical system and represents a special case of the general transport equation. In the special case of a free-space propagation over a distance z , and assuming the paraxial approximation, one obtains

$$\Gamma_i(\vec{r}_1', \vec{r}_2', z) = \left(\frac{k}{2\pi z} \right)^2 \cdot e^{+i\frac{k}{2z}(\vec{r}_1'^2 - \vec{r}_2'^2)} \cdot \iint I_0(\vec{r}, z) \cdot e^{-i\frac{k}{2z}\vec{r} \cdot (\vec{r}_1' - \vec{r}_2')} d^2\vec{r}. \quad (19-125)$$

According to this equation the coherence function at infinite distance to an incoherent light source is given by the Fourier transform of the intensity distribution of the source with respect to the difference of the coordinates.

19.4.4

The Coherence Function of a Circular Source

Circular light sources are of special importance in classical optical systems with rotational symmetry. In practice, the consideration of the image formation in the case of partially coherent illumination includes the pupil as a secondary light source which very often is of circular shape. Within the paraxial regime, the coherence function of a circular and homogeneously emitting effective source with radius a and distance z is given by the Fourier transform of the circular area

$$\Gamma(\vec{r}_1, \vec{r}_2) = C \cdot \frac{J_1\left(2\pi \cdot a \cdot \frac{|\vec{r}_2 - \vec{r}_1|}{\lambda \cdot z}\right)}{|\vec{r}_2 - \vec{r}_1|}. \quad (19-126)$$

The geometry of this arrangement is schematically shown in figure 19-36.

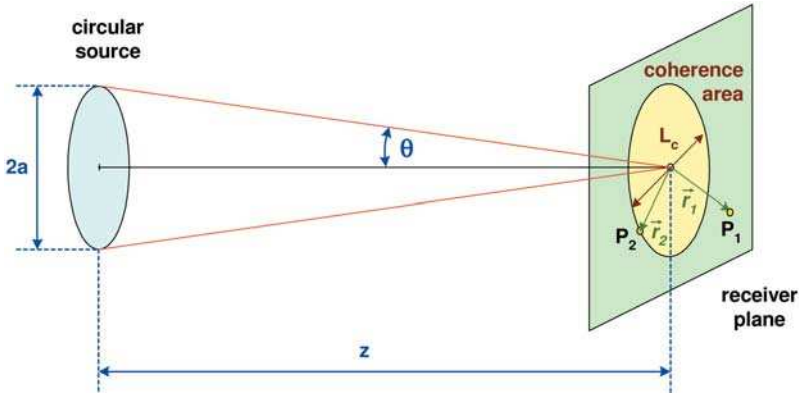


Figure 19-36: Evolution of the coherence state of light emitted by a finite circular source with the distance z .

The diameter of the coherence region is approximately given by

$$L_c = |\vec{r}_2 - \vec{r}_1| = 1.22 \frac{\lambda \cdot z}{R} . \quad (19-127)$$

For the contrast between a point at radius r and the centre at $r = 0$ one obtains

$$V(r) = \frac{\lambda \cdot z}{\pi \cdot ar} \cdot J_1 \left(\frac{2\pi ar}{\lambda z} \right) . \quad (19-128)$$

The visibility or contrast according to eq. (19-128) is illustrated in figure 19-37. At certain radii, the zeros of the Bessel function lead to vanishing contrast. Thus, an interference experiment with a circular light source of radius a one obtains the vanishing interference contrast for detection point distances r corresponding to the first zero of the Bessel function

$$\frac{2\pi a}{\lambda} \cdot \frac{r}{z} = \frac{2\pi a}{\lambda} \cdot NA = 3.83 \quad (19-129)$$

from which for the visible angle of the source follows:

$$\sin \theta_{\text{null}} = 0.61 \cdot \frac{\lambda}{a} . \quad (19-130)$$

This approach is applied e.g. in Michelson's stellar interferometer for the analysis of star diameters.

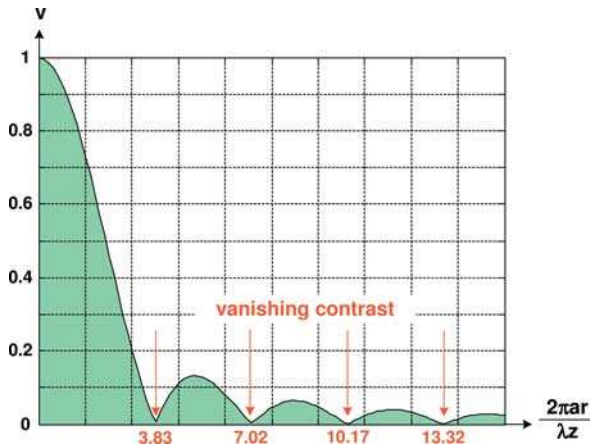


Figure 19-37: Contrast of the radiation from a circular light source as a function of the distance z .

According to a convenient definition, the region of coherence is given by the angle of the light cone, at which the Bessel function drops by 12% to 0.88 of its maximum value. This corresponds to arguments of the Bessel function equal to 1. One obtains for the angle

$$\sin \theta_{\text{coh}} = 0.16 \cdot \frac{\lambda}{a}. \quad (19-131)$$

As a consequence, the field at a distance z from the light source is considered to be coherent over a coherence area having a diameter

$$L_c = 2z \sin \theta_{\text{coh}} = 0.32 \cdot \frac{\lambda \cdot z}{a}. \quad (19-132)$$

For incoherent imaging, where the pupil of the imaging system is completely filled by the light source, the coherence function is equivalent to the point-spread function of the optical system. If a smaller light source is applied, the coherence radius increases in comparison with the point-spread function. In general, the smaller the source, the longer the wavelength or the longer the distance to the source, the larger the coherence range.

A further example with practical importance is the illumination of the earth by the sun. The sun is seen from the earth at an angle of $\sim 32'$, corresponding to ~ 9.3 mrad. Hence one obtains for $a/z \sim 0.005$. The coherence range of monochromatic sunlight on the earth thus is given by

$$L_{c,\text{Sun}} \approx 0.12 \text{ mm} \quad \text{for } \lambda = 500 \text{ nm}. \quad (19-133)$$

Hence, even using a color filter, it is difficult to observe interference effects with sunlight. One needs a double slit with a spacing of less than $50 \mu\text{m}$ or a grating with a period of less than $50 \mu\text{m}$.

19.4.5

Coherence Function behind a Double Slit

In Young's interference experiment, the light from a quasi-monochromatic light source at a distance z_1 is diffracted at two pinholes or a double slit with spacing D . This new source consisting of two points, forms an interference pattern in a certain region on a screen at a distance z_2 behind the apertures. Figure 19-38 shows the geometry of this arrangement.

In a simplified approach of complete coherence the primary light source is considered as a point. The intensity distribution, with x as a coordinate on the screen, is given by

$$I(x) = 4I_0 \cdot \cos^2 \frac{\pi D x}{\lambda \cdot z_2}. \quad (19-134)$$

For the period of the interference fringes it follows

$$\Delta x = \frac{\lambda \cdot z_2}{D}. \quad (19-135)$$

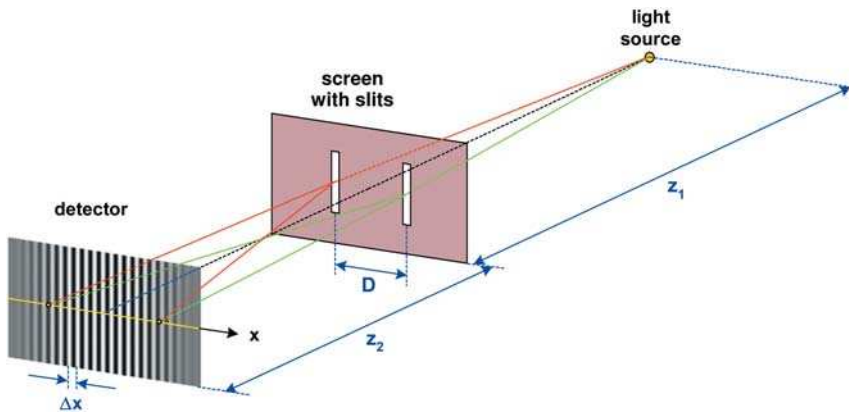


Figure 19-38: Young's double slit experiment with interference fringes.

The general case of a finite-sized light source with lateral dimension a is illustrated in figure 19-39. The finite extent of the incoherently emitting light source in general decreases the contrast of the interference fringes, depending on the geometry of the set-up.

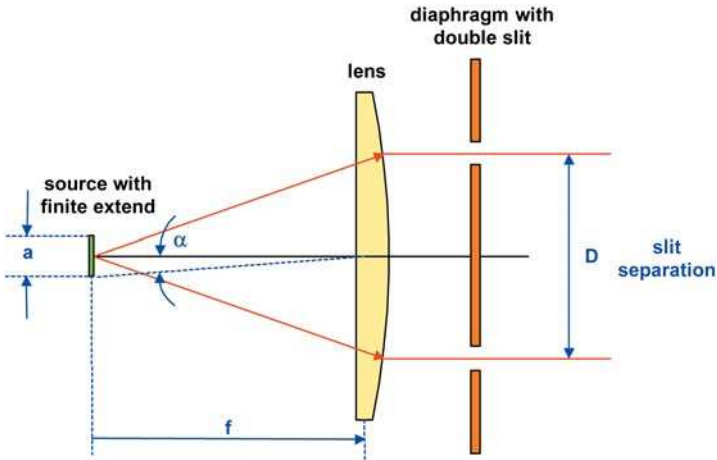


Figure 19-39: Double slit experiment with an extended light source.

Figure 19-40 shows a typical intensity distribution obtained by Young's double slit experiment. The contrast reduction when increasing the width D of the slits is illustrated in the cross-section on the right side of figure 19-40.

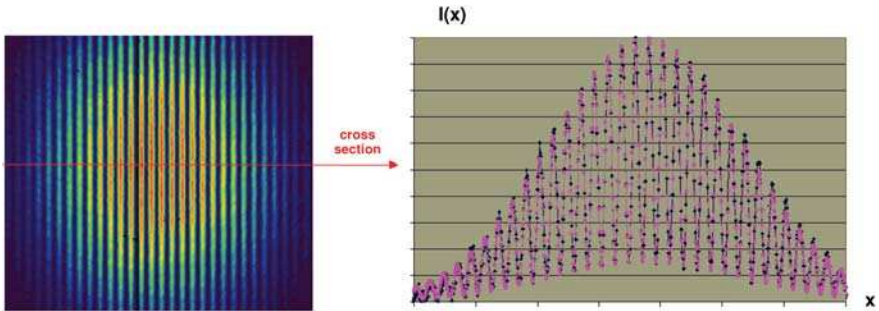


Figure 19-40: Interferogram and cross-section obtained in Young's experiment with a double slit using a partially coherent light source.

Figure 19-41 compares the coherence function $\Gamma(x_1, x_2)$ in the image plane behind a double slit for a small and a large light source. The main diagonals of the coherence function correspond to the intensity $I(x_1) = \Gamma(x_1, x_1)$, which, in the case of a large light source, degenerates into a constant function (figure 19-41b). In contrast, interference fringes are visible in the case of a small light source. The opposite diagonal describes the coherence function $\Gamma(x_2 - x_1)$.

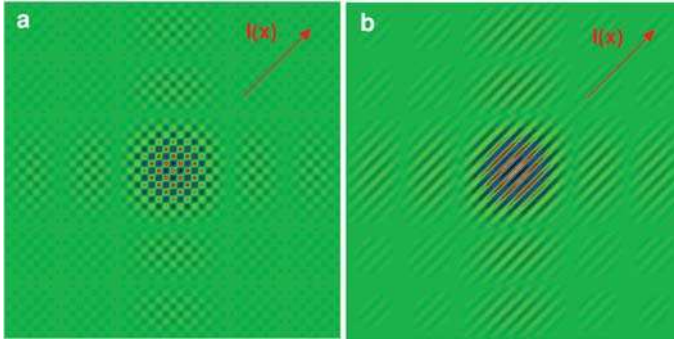


Figure 19-41: Coherence functions $I(x_1, x_2)$ behind a double slit for different sizes a of the incoherent light source: a) small light source corresponding to coherent illumination; b) incoherent source.

The typical behaviour of the contrast function in dependence on the double slit separation D is illustrated in figure 16-42 at the example of several measurements with an excimer laser emitting at a wavelength of $\lambda = 193$ nm as source. With increasing slit separation D the visibility of the interference pattern is decreasing. The difference in the propagation time for the two-beam paths increases with increasing double slit distance and the two slits, as secondary sources, become more and more incoherent.

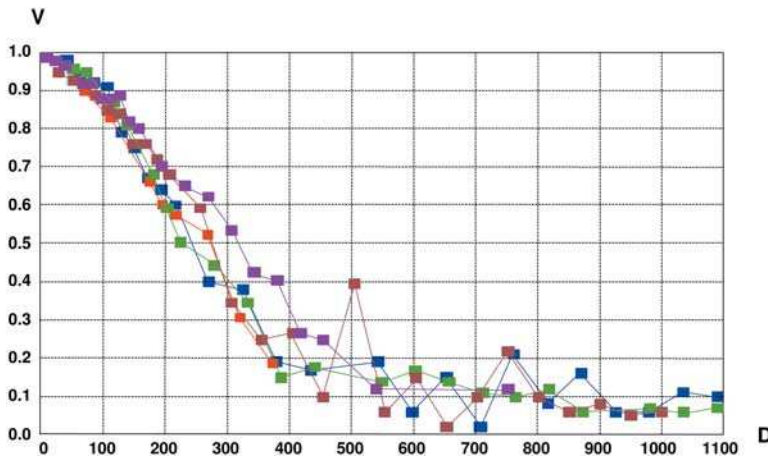


Figure 19-42: Examples of the contrast function in Young's double slit experiment with a partially coherent excimer laser. The slit separation D is given in μm .

Figure 19-43 shows the intensity distribution of a double slit experiment with a finite size light source and varying partial coherence. The coherence is quantified by the parameter σ , the effective size of the source is given by σa . The distance of the double slit is four times the Airy diameter, with slit widths equal to the Airy diameter. The decreasing coherence results in reduced contrast, and for $\sigma = 0.4$ almost no interference is visible.

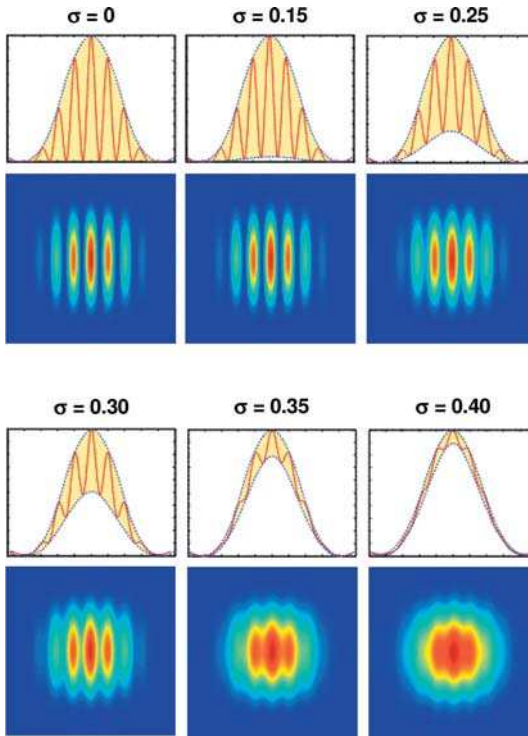


Figure 19-43: Contrast function in Young's double slit experiment with partially coherent illumination.

19.4.6

Propagation of the Wigner Distribution Function

The changes in the coherence properties of wave field due to propagation can be described by the Wigner distribution function. Particularly for paraxial systems, the treatment of propagation with the help of the Wigner distribution function is considerably simpler and can be numerically accomplished at a higher speed compared to propagation of the coherence function.

With the Greens function G as the coherent transfer function of the field U , the coherent propagation of light can be written in the most general form

$$U(\vec{r}_2) = \iint G(\vec{r}_2, \vec{r}_1) \cdot U(\vec{r}_1) d\vec{r}_1 . \tag{19-136}$$

A general propagator applied to the Wigner distribution function, however, must be four-dimensional, and for the propagation of the Wigner distribution function it may be written

$$W'(\vec{r}_2, \vec{p}_2) = \iiint K(\vec{r}_2, \vec{p}_2, \vec{r}_1, \vec{p}_1) \cdot W(\vec{r}_1, \vec{p}_1) d\vec{r}_1 d\vec{p}_1 \tag{19-137}$$

with the transfer kernel

$$K(\vec{r}_2, \vec{p}_2, \vec{r}_1, \vec{p}_1) = \frac{1}{4\pi^2} \iint G\left(\vec{r}_2 + \frac{\vec{r}_2'}{2}, \vec{r}_1 + \frac{\vec{r}_1'}{2}\right) G^*\left(\vec{r}_2 - \frac{\vec{r}_2'}{2}, \vec{r}_1 - \frac{\vec{r}_1'}{2}\right) e^{-ik_0(\vec{r}_2'\vec{p}_2 - \vec{r}_1'\vec{p}_1)} d\vec{r}_2' d\vec{r}_1' \quad (19-138)$$

The propagation of the Wigner distribution function in free space is given in accordance with the geometrical interpretation in section 19.2.2 by

$$x' = x + z \cdot \tan u = x + z \cdot \frac{v}{\sqrt{n^2/\lambda^2 - v^2}} \quad (19-139)$$

$$v' = v, \quad u' = u. \quad (19-140)$$

In the phase space this corresponds to a shearing in direction of the x -axis as shown in figure 19-44. In the paraxial approximation, the shear according eq. (19-139) may be simplified by a linear shear. The propagation is then given by a simple linear scale transformation of the position x .

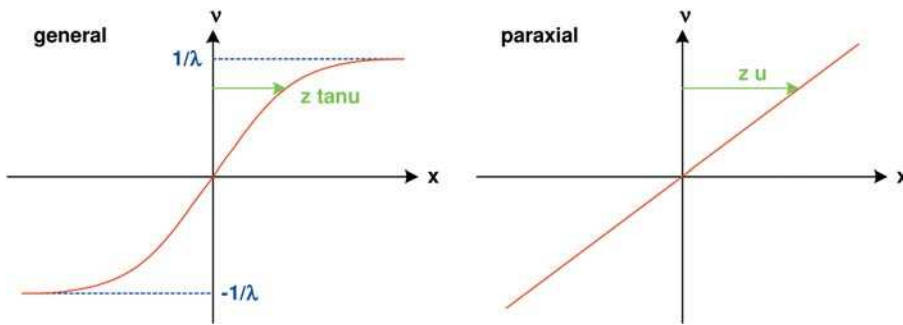


Figure 19-44: Free-space propagation in the phase space is characterized by angular deformation in the general and the paraxial case.

The propagation of the Wigner distribution function is illustrated in figure 19-45 at the example of a point light source. The spatial intensity distribution $I(x)$ is obtained by integration over v . The shearing of the function $W(x, v)$, however, obviously leads to broadening of the intensity $I(x)$ with increasing distance z .

In imaging optical systems, the propagation kernel for the Wigner distribution function takes over a more complex form. In one dimension it may be written as

$$K(x_1, p_1, x_2, p_2) = \frac{k}{2\pi|B|} \iint e^{-ikV(x_2+x_2'/2, x_1+x_1/2) + ikV^*(x_2-x_2'/2, x_1-x_1/2) - ikp_2x_2' + ikp_1x_1'} dx_2' dx_1' \quad (19-141)$$

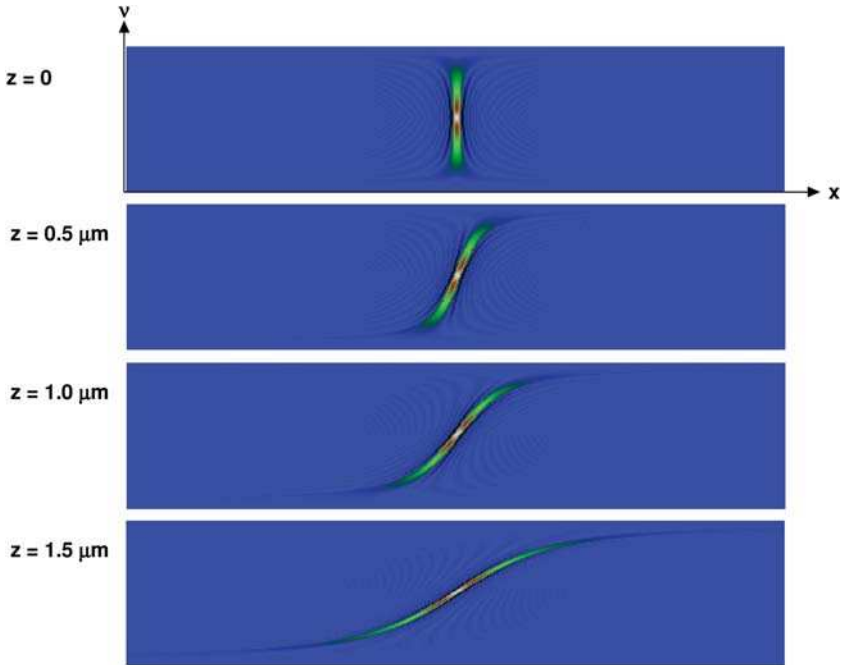


Figure 19-45: Wigner distribution functions of a point image (5 nm slit) by free-space propagation from $z=0 \mu\text{m}$ to $1.5 \mu\text{m}$.

with the point eikonal $V(x, x')$ (see paragraph 20). In practical cases, this formulation is simplified to a convolution of the incoming Wigner distribution function with a filtering function, which characterizes the optical system. The Wigner distribution function in the image plane may thus be written as

$$W'_{\text{out}}(x, p) = \iint W'_{\text{in}}(x, p') \cdot W'_{\text{filt}}(x - x', p - p') dp' dx' . \quad (19-142)$$

In ideal or paraxial systems, the point eikonal is represented by the elements of the ABCD matrix. The transport of the Wigner distribution function can be written as

$$W'(\vec{r}') = W'(\mathbf{M}^{-1} \cdot \vec{r}) \quad (19-143)$$

with the spatial and the angle coordinates written as vector \vec{r} . In two dimensions, this equation takes the well-known simple form with the angle u instead of p for the paraxial approximation

$$W'_{\text{out}}(x, u) = W'_{\text{in}}(D \cdot x - B \cdot u, -C \cdot x + A \cdot u) . \quad (19-144)$$

Within the scope of this approximation, the Wigner distribution function has a constant value along a paraxial ray and, for numerical evaluation of propagation, only an interpolation between the starting and the receiving coordinate grids has to be taken into account, as illustrated in figure 19-46.

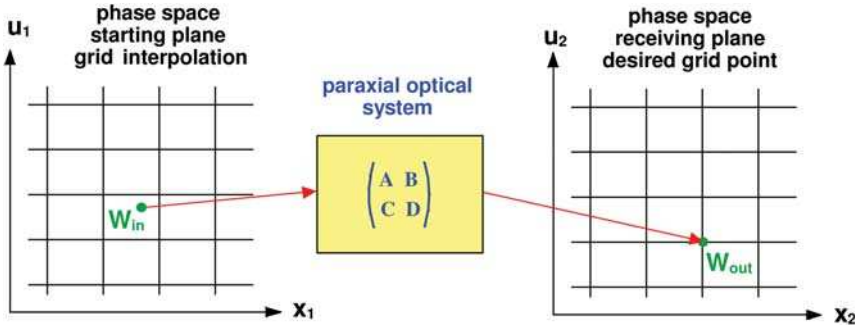


Figure 19-46: Propagation of the Wigner distribution function in a paraxial optical system.

A second type of propagation problem is important for the evaluation in practical examples. If an optical element can be approximated by an infinitely thin object, it has no effect on the dependence from the coordinate x and the propagation of the Wigner distribution function through the element reduces to a change of the optical direction cosines p as a function of the coordinate x . If the object, for example, is pure imaginary, as the case for wave aberrations,

$$T(x) = e^{i\phi(x)} \tag{19-145}$$

and the variation of the phase with the position x is smooth, the Wigner distribution function of the object is given by

$$W'_{\text{object}}(x, p) = \int T\left(x + \frac{x'}{2}\right) \cdot T^*\left(x - \frac{x'}{2}\right) \cdot e^{-ikx'p} dx' \tag{19-146}$$

and the transport can be calculated by a convolution with respect to the angle coordinate only

$$W'_{\text{out}}(x, p) = \int W'_{\text{in}}(x, p') \cdot W'_{\text{object}}(x, p - p') dp' . \tag{19-147}$$

This special case is frequently applied to incorporate e.g. wave aberrations in the propagation through optical systems.

19.5 Gaussian Schell Beams

19.5.1 Definition of Gaussian Schell Beams

There exists a simple type of partially coherent beam which can be analytically described in close analogy to the Gaussian beams in the coherent case [19-16], [19-17], [19-18], [19-19], [19-20]. These so-called Gaussian Schell beams rest upon

simple statistical assumptions concerning the coherent basis functions of Γ . Their transport properties are fully equivalent with those of the Gaussian beams with the consequence that their propagation can be described analytically, see, for example, eq. (19-155). In practice, these ideal solutions are useful as a first approximation for the description of partially coherent radiation and will be considered in more detail in the following.

Gaussian Schell beams are partially coherent Gaussian beams with a Gaussian shaped coherence function. In the beam waist, where at $z=0$ the phase front is plane, the coherence function is real and is given by the expression

$$\Gamma(r_1, r_2) = e^{-\frac{r_1^2 + r_2^2}{w_0^2}} \cdot e^{-\frac{(r_1 - r_2)^2}{2L_c^2}}. \quad (19-148)$$

The transverse coherence length L_c in the above formula is an additional parameter responsible for the description of the coherence. In the limiting case $L_c \rightarrow \infty$ one obtains the conventional coherent Gaussian beams. With decreasing L_c

- the emitting area radiates more incoherent radiation,
- the beam divergence increases for a constant emitter area,
- the beam depth of focus decreases,
- the value of the beam parameter product becomes larger and the beam quality degrades.

These dependences are shown in figs 19-47 to 19-49. Besides L_c the following equivalent quantities are commonly used:

1. Normalized transverse coherence length

$$\alpha = L_c / w_0. \quad (19-149)$$

2. Auxiliary parameter ε according to

$$\varepsilon = 1 + \frac{1}{\alpha^2} = 1 + \left(\frac{w_0}{L_c}\right)^2 > 1. \quad (19-150)$$

3. Degree of coherence β

$$\beta = \frac{1}{\sqrt{1 + (w_0/L_c)^2}} = \frac{1}{\sqrt{1 + \alpha^{-2}}} = \frac{1}{\sqrt{\varepsilon}} < 1. \quad (19-151)$$

As opposed to the classical Gaussian beams, three parameters are allocated for the description of the Gaussian Schell beams. The waist radius w_0 and the far-field divergence θ_0 are independent parameters for a given wavelength λ . In particular, the depth of focus for the Gaussian Schell beams is shorter in comparison to the coherent Gaussian beams

$$z_0' = \frac{z_0}{\sqrt{\varepsilon}} = \beta \cdot z_0 < z_0. \quad (19-152)$$

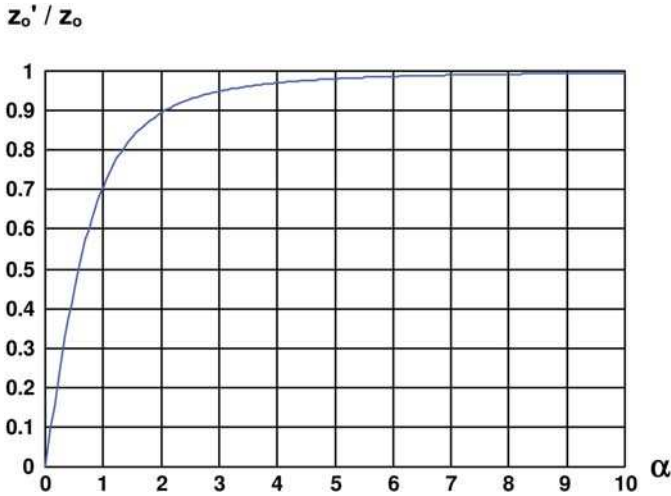


Figure 19-47: Depth of focus as a function of the normalized transverse coherence length α for Gaussian Schell beams.

The parameter z'_0 decreases for reduced coherence (figure 19-47).

At a constant far-field divergence θ_0 the waist radius of Gaussian Schell beams increases when L_c decreases, i.e., when the degree of coherence β becomes smaller. Figure 19-48 demonstrates this behavior for three different values of the parameter β .

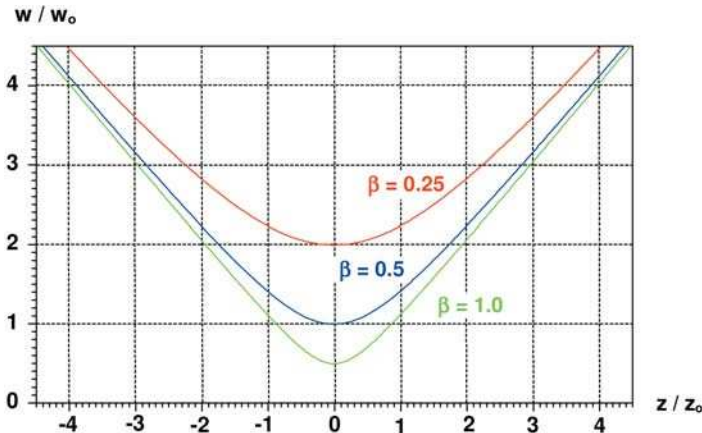


Figure 19-48: Beam radius of Gaussian Schell beams for constant divergence angle and different degrees of coherence in the waist region.

When, on the other hand, the waist radius w_0 is kept constant, the far-field divergence angle grows with decreasing coherence. Figure 19-49 shows some examples of this case. Practically, the meaning of this relationship is that, in order to achieve the same focal size, the light from a more incoherent source should be focussed with a larger numerical aperture.

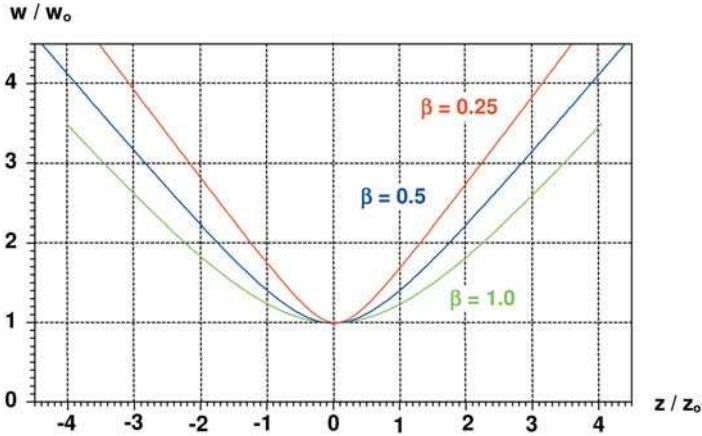


Figure 19-49: Beam radius of Gaussian Schell model beams for constant waist radius and different degrees of coherence in the waist region.

This dependence of the far-field divergence angle from the coherence can be quantitatively described by the expression

$$\theta_o' = \theta_o \cdot \sqrt{\varepsilon} . \quad (19-153)$$

The coherence length L_c in the waist can be calculated for given λ , w_o and θ'_o from the following formula

$$L_c = \frac{w_o}{\sqrt{\left(\frac{\pi w_o \theta'_o}{\lambda}\right)^2 - 1}} . \quad (19-154)$$

Gaussian Schell beams exhibit a value of the degree of coherence β which is between the geometrical-optical incoherent case with $\beta=0$ and the ideal coherent Gaussian beams with $\beta=1$. This can also be seen in the image formation behavior, which is shown in figure 19-50 for a single lens with focal length f . The image formation or transformation equation for Gaussian Schell beams reads

$$\frac{1}{z_T'} - \frac{1}{z_T} \cdot \frac{1}{\left[1 + \frac{z_o'^2}{z_T \cdot (f + z_T)}\right]} = \frac{1}{f} . \quad (19-155)$$

and deviates, similar to the Gaussian beams, from the geometrical-optical models. In an analogous way, the position of the beam waist at the image-side can be calculated analytically.

It should be taken into account that the normalized degree of coherence in the transport of Gaussian Schell beams does not change with the coordinate z

$$\beta(z) = \text{const} . \tag{19-156}$$

It should also be noticed, that Gaussian Schell model beams are by definition paraxial and the relation (19-156) is valid only in this approximation.

Recently, the simple partially coherent Gaussian Schell beams described here were generalized using various beam models. These include twisted beams [19-21], [19-22], [19-23], flatter or more complex intensity profiles [19-24] and vector properties [19-25].

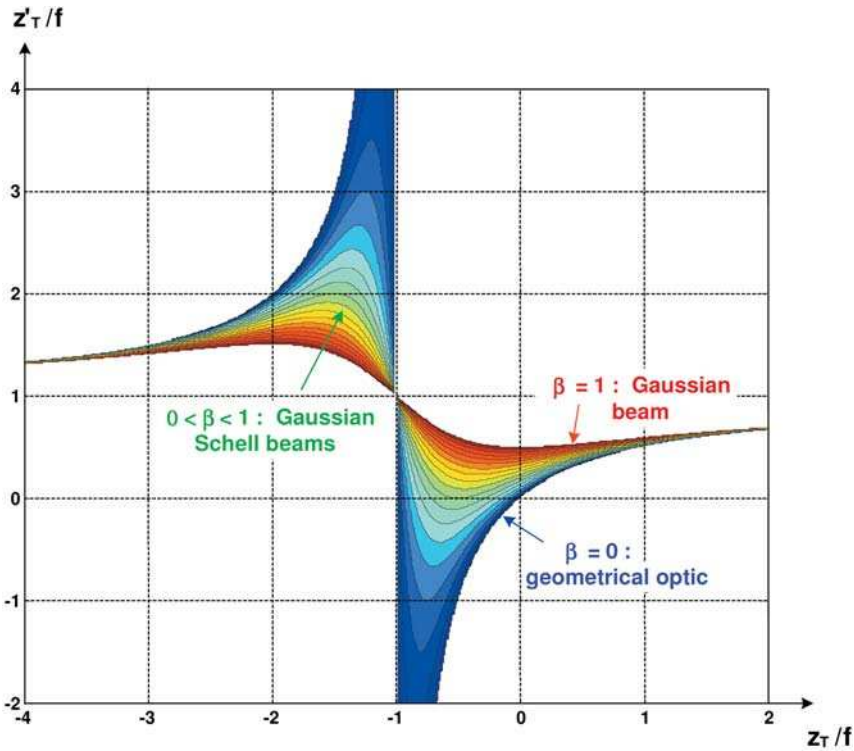


Figure 19-50: Illustrating the image formation properties of Gaussian Schell beams with object-side and image-side waist positions z_T and z'_T in dependence on the coherence parameter β .

19.5.2

Coherence and Wigner Functions of Gaussian Schell Beams

The coherence function of Gaussian Schell beams has the following analytical form for arbitrary planes z

$$\Gamma(\vec{r}_1, \vec{r}_2, z) = I_0 \cdot \left[\frac{w_0}{w(z)} \right]^2 e^{-\frac{\vec{r}_1^2 + \vec{r}_2^2}{w^2(z)}} e^{-\frac{(\vec{r}_1 - \vec{r}_2)^2}{2L_c^2(z)}} e^{-\frac{i\pi(\vec{r}_1^2 - \vec{r}_2^2)}{\lambda R(z)}}. \quad (19-157)$$

The coherence function reduces to eq. (19-148) at the beam waist where the curvature term vanishes and L_c takes the value at the waist plane

$$\Gamma(\vec{r}_1, \vec{r}_2, 0) = I_0 \cdot e^{-\frac{\vec{r}_1^2 + \vec{r}_2^2}{w_0^2}} e^{-\frac{(\vec{r}_1 - \vec{r}_2)^2}{2L_c^2}}. \quad (19-158)$$

The Wigner function for Gaussian Schell beams reads, in paraxial approximation

$$\begin{aligned} W(\vec{r}, \vec{u}) &= \frac{I_0}{2\pi} \cdot \frac{w^2(z) \cdot L_c^2(z)}{w^2(z) + L_c^2(z)} \cdot e^{-\frac{2\vec{r}^2}{w^2(z)}} \cdot e^{-\frac{2\pi^2}{\lambda^2} \frac{w^2(z) \cdot L_c^2(z)}{w^2(z) + L_c^2(z)}} \left(\vec{u} - \frac{\vec{r}}{R} \right)^2 \\ &= \frac{I_0}{2\pi} \cdot \frac{w^2(z) \cdot L_c^2(z)}{w^2(z) + L_c^2(z)} \cdot e^{-\frac{2\vec{r}^2}{w^2(z)}} \cdot e^{-\frac{2}{\theta^2} \left(\vec{u} - \frac{\vec{r}}{R} \right)^2}. \end{aligned} \quad (19-159)$$

The intensity distribution of a Gaussian Schell beam follows a Gaussian function in both the transverse dimension as $I(x)$ and as a function of the angle $I(u)$. Figure 19-51 illustrates these relationships for the Wigner function $W'(x, u)$ by the intensity profiles in a general x - u phase-space representation. A Gaussian Schell beam is represented in the phase space by a set of ellipses in which each ellipse corresponds to a W' contour. According to this, the full width of the angular distribution is found only in the beam center on the axis while the angular spreading in the peripheral regions is substantially smaller.

The above diffraction-related angular distribution overlaps at any arbitrary position outside the beam waist with a geometrical-optical term resulting from the parabolic phase contribution. The latter corresponds to a paraxial radius of curvature R and acts in the phase space as a tilting of the contour lines in accordance with the straight line $u = x/R$. The stronger the focussing of the Gaussian Schell beams, the larger the relative geometrical-optical angular contribution. Then, the phase space ellipses of the Wigner function become increasingly elongated.

It is known from Fourier optics that the spatial distribution occurring in the focus of a lens is proportional to the angular spectrum. This operation corresponds in phase space to a 90° rotation. In accordance with this the Wigner function in the focus of a paraxial lens has the same form as in front of the lens being only rotated about an angle of 90° .

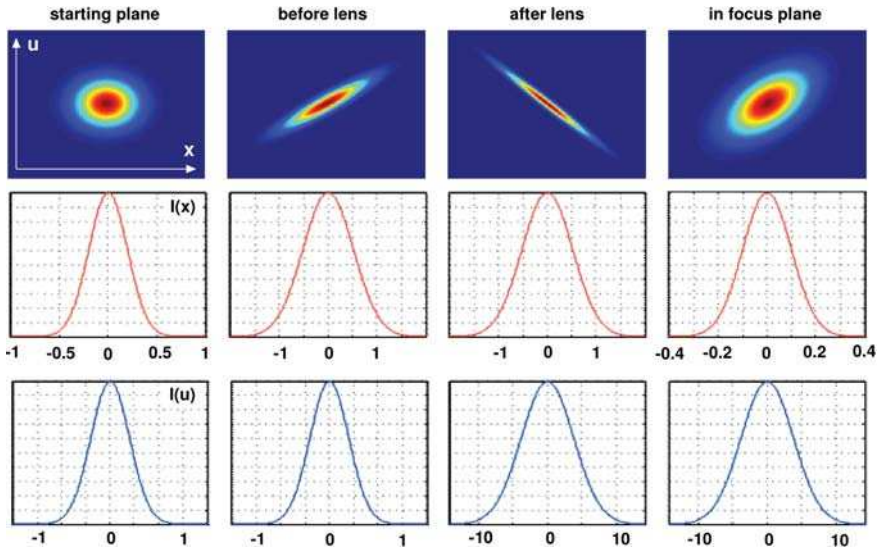


Figure 19-51: Transformation of a Gaussian Schell beam by an ideal lens. The bottom row shows the angular distribution and the middle row shows the spatial distribution. The axes are scaled in arbitrary units, the scale steps in the angular extent at the lens and in the spatial extent for the focussing can be clearly seen.

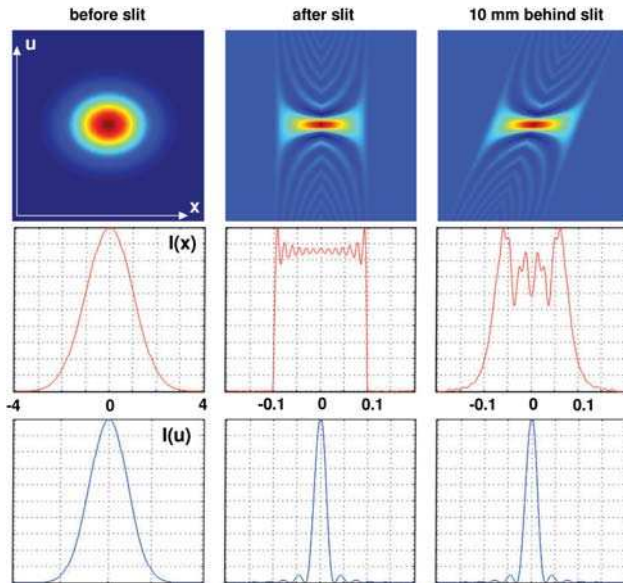


Figure 19-52: Transformation of a Gaussian Schell beam by a slit aperture which truncates its profile. The bottom row shows the angular distribution and the medium row shows the spatial distribution.

If a Gaussian Schell beam is truncated by an aperture, the corresponding Wigner function exhibits characteristic diffraction patterns. Figure 19-52 shows an example. One can see the effect of the edge diffraction and the comparatively unaffected inner region of the beam. The aperture interacts mainly with the outer region of the beam. At the same time the interference capability and therefore the occurrence of the characteristic diffraction ripples depends on the degree of coherence. The more incoherent the beam, the less pronounced are the diffraction effects.

According to their definition, however, Gaussian Schell beams are paraxial beams that are not transversally truncated. In the example presented, the incoming beam is a Gaussian Schell beam but the action of the lens changes its character and the beam is no longer of the Gaussian Schell type behind the aperture. Nevertheless it is still possible to describe the beam using the Wigner function.

19.5.3

Basis Mode Expansion of Partial Coherent Fields

The coherence function obeys the system of differential equations (19-115). According to the theorem of Mercer in the spectral theory of partial differential equations, the eigenfunctions of the equations can be used to express an arbitrary solution in the form of an expansion. In the special case of the coherence function, these coherent basic mode functions are superimposed incoherently and constitute the partial coherent field [19-26]. The expansion of the coherence function then takes the form

$$\Gamma(\vec{r}_1, \vec{r}_2) = \sum_{n=0}^{\infty} \lambda_n \varphi_n(\vec{r}_1) \varphi_n^*(\vec{r}_2) = \sum_{n=0}^{\infty} \lambda_n \Gamma_n(\vec{r}_1, \vec{r}_2) \quad (19-160)$$

with the mode functions $\varphi_n(\vec{r})$, the eigenvalues λ_n and the mode index n . The corresponding eigenvalue equation with the coherence function as kernel reads

$$\int \Gamma(\vec{r}_1, \vec{r}_2) \varphi_n(\vec{r}_1) d\vec{r}_1 = \lambda_n \varphi_n(\vec{r}_2). \quad (19-161)$$

The intensity of the field distribution can be obtained from eq. (19-160) in the usual manner and is given by

$$I(\vec{r}) = \sum_{n=0}^{\infty} \lambda_n |\varphi_n(\vec{r})|^2 = \sum_{n=0}^{\infty} \lambda_n I_n(\vec{r}). \quad (19-162)$$

The last equation shows that the eigenvalues determine the power distribution of the superposition of the independent coherent basic modes. It should be noted, that an expansion must not necessarily use the eigenfunctions. It can also take arbitrary non-orthogonal complete systems of basic functions. For example, it is very useful to choose localized beamlet functions with a gaussian form. This delivers the decomposition of the field according to the theory of Gabor [19-27].

In a very similar manner, the Wigner distribution function can be expressed by the corresponding basic modes in the expansion

$$W'(x, p) = \sum_{n=0}^{\infty} \lambda_n \psi_n(x, p) \tag{19-163}$$

where the functions $\psi_n(x, p)$ are calculated by the basic functions $\varphi_n(\vec{r})$ according to

$$\psi_n(x, p) = \int \varphi_n\left(x + \frac{\Delta x}{2}\right) \cdot \varphi_n^*\left(x - \frac{\Delta x}{2}\right) \cdot e^{-ik_0 \Delta x \cdot p} d\Delta x . \tag{19-164}$$

In the simple case of, for example, Gauss Schell model beams, the representation of the field in the form of a basic mode expansion can be performed analytically in several ways. One possibility uses the classical Gauss-Hermite basis functions [19-28], another way is to use elementary gaussian modes in special geometrical configurations such as displacement, tilt or twist axes [19-29].

Figure 19-53 shows the approximation of a super gaussian profile into single gaussian beamlets according to eq. (19-162). This one-dimensional picture is generalized to rotational symmetry in figure 19-54. The axis of the elementary beamlets are located on a cone.

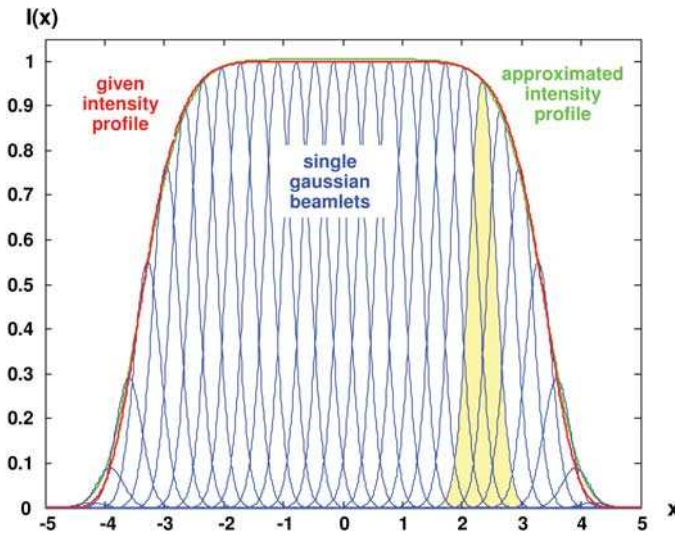


Figure 19-53: Expansion of an intensity profile into single non-orthogonal gaussian beamlets.

In particular, the generalization of beams with intrinsic twist properties can be illustrated very easily in this framework. Figure 19-55 shows the geometry for two single beamlets. The axis of these modes are skew to one another. This causes a coupling of the x- and y-section planes, the resulting field contains an internal twist, the beam shows a kind of rotation during the propagation. The individual beamlets are given by the field strength

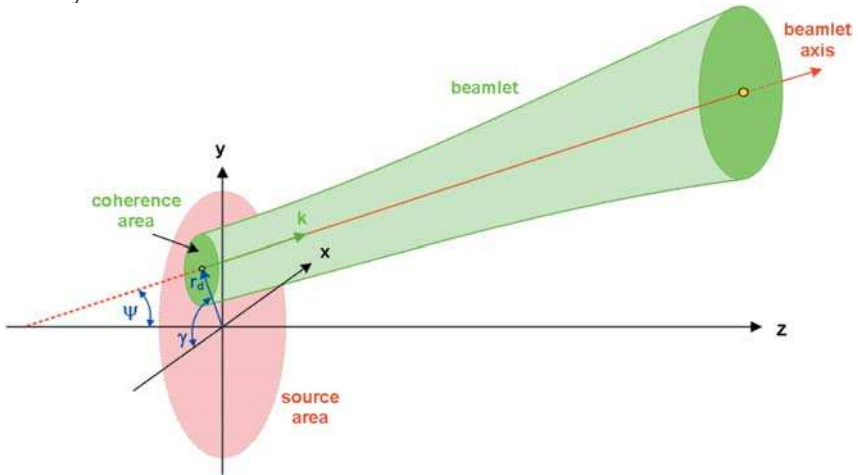


Figure 19-54: Expansion of an intensity profile into single non-orthogonal gaussian beamlets for rotational symmetrical geometry.

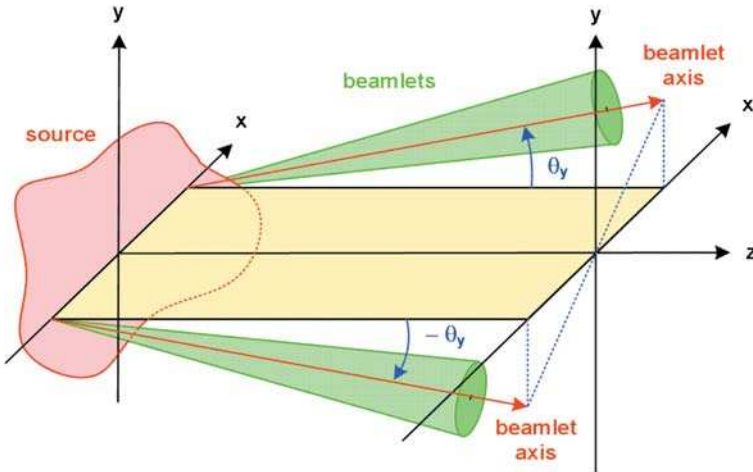


Figure 19-55: Expansion of an intensity profile into single non-orthogonal gaussian beamlets, general geometry with twist of the beam axis.

$$U(x, y, x_0, y_0) = U_0 \cdot e^{-\frac{(x-x_0)^2}{w_{0x}^2} - \frac{(y-y_0)^2}{w_{0y}^2}} \cdot e^{2\pi i \cdot (\alpha \cdot x_0 y - \beta \cdot y_0 x)} \quad (19-165)$$

with two parameters α and β describing the twist. The coherence function is given by the equation

$$\Gamma(x_1, y_1, x_2, y_2) = \iint P(x_0, y_0) \cdot U(x_1, y_1, x_0, y_0) \cdot U^*(x_2, y_2, x_0, y_0) dx_0 dy_0 \quad (19-166)$$

with a weighting function $P(x_0, y_0)$, which is determined by the intensity profile of the source field.

19.6 Statistical Optics and Speckle

19.6.1

Photon Statistics

The predominant part of the light from thermal sources is emitted spontaneously, i.e., almost all photons are statistically independent of each other. Each photon can be considered as an independent degree of freedom of the electromagnetic field and the photon number determines the noise of an image by its statistical character. This can be easily observed with modern digital cameras in the darkness.

In contrast, the photons emitted from a laser by stimulated emission are coupled and cannot be distinguished from each other. These identical photons are described by the spatial modes of the laser. Thus the degree of freedom is limited to a great extent by the number of modes. Many lasers emit in the fundamental mode only, others like the excimer lasers emit simultaneously in more than 1000 modes. The number of modes, which dominates the statistical effects in the intensity formation, is thus typically much smaller than the number of photons.

The different effects will be sketched in the following using the image formation as an example. As already mentioned in the beginning it is the carrier frequency which determines the photon energy:

$$Q = \hbar \cdot \omega = \frac{h}{T} = \frac{hc}{\lambda} = \frac{6.626 \cdot 10^{-34} \text{ Js} \cdot 2.99792 \cdot 10^8 \frac{\text{m}}{\text{s}}}{\lambda} = \frac{19.865 \cdot 10^{-17} \text{ J nm}}{\lambda} \quad (19-167)$$

where $h = 6.626 \cdot 10^{-34}$ Js denotes Planck's constant and $c = 2.998 \cdot 10^8$ m/s is the speed of light. The same expression in eV reads

$$Q[\text{eV}] = \frac{1986.5 \cdot 10^{-19} \text{ 1 eV}}{1.60219 \cdot 10^{-19}} \frac{1}{\lambda} \text{ nm} = \frac{1239.9 \text{ eV}}{\lambda} \text{ nm} . \quad (19-168)$$

Thus the quantum energy of light with a wavelength of 12.4 nm amounts to 100 eV. Using eq. (19-167) one obtains for the number of photons N per energy interval

$$\frac{N}{Q} = \lambda \cdot \frac{1 \text{ J}}{c \cdot h} \frac{1}{\text{J}} = 5.034 \cdot 10^{15} \frac{\lambda}{\text{nm}} \frac{1}{\text{J}} . \quad (19-169)$$

The parameter used in radiometry instead of power is called the radiant flux $\Phi_e = \partial Q / \partial t$, measured again in Watts [W]. Since a radiant flux of 1 Watt corresponds to 1 Joule per second, one can calculate the number of photons of a source per second from the radiant flux:

$$\frac{N}{\Phi} = \frac{\lambda}{c \cdot h} \left[\frac{1}{\text{W}} \right] = 5.034 \cdot 10^{15} \frac{\lambda}{\text{nm}} \frac{1}{\text{W}} . \quad (19-170)$$

A source emitting 1000 photons per second in the visible has therefore a flux of 2 femto-Watts (1 fW = 10^{-15} W). A 30 mW HeNe-laser emits about 10^{17} photons/s. Since the photon energy decreases with the wavelength, the number of photons, for a constant radiant flux, increases with the wavelength. As a consequence of the statistical fluctuations of the photon number (Poisson statistics and shot noise, respectively) any power measurement is affected by a definite signal-to-noise ratio (SNR) given by

$$\text{SNR} = 10 \log_{10} \left(\frac{\sigma_s^2}{\sigma_n^2} \right) \sim \frac{\sqrt{N}}{N} \quad (19-171)$$

where σ_s^2 denotes the average square of the signal amplitude and σ_n^2 is the variance of the noise. Table 19-3 shows, for comparison, the necessary photon number and the resulting noise per resolvable image point for a wavelength of 1000 nm, 100 nm and 10 nm, assuming a detector or a film requiring illumination intensities of 1 mW/cm^2 whereas the resolution is, of course, wavelength dependent ($\text{NA} = 0.6$). Figure 19-56 shows the double logarithmic dependence of the noise on the wavelength for different detector sensitivities and apertures of the imaging optics.

Table 19-3: Photon number and noise per image point for a radiation flux density of $10 \mu\text{W/cm}^2$ and different wavelengths.

	1000 nm	100 nm	10 nm
Frequency $\frac{1}{T} = \frac{c_0}{\lambda}$	$3 \cdot 10^{14}$ Hz	$3 \cdot 10^{15}$ Hz	$3 \cdot 10^{16}$ Hz
Energy $Q = h\nu$	1.24 eV	12.40 eV	123.99 eV
Photons/Watt $\frac{N}{\Phi}$	$5.03 \cdot 10^{15}/\text{mW}$	$5.03 \cdot 10^{14}/\text{mW}$	$5.03 \cdot 10^{13}/\text{mW}$
Photons/area at 1 mW/mm^2 (in N/mm^2)	$5.03 \cdot 10^{13}$	$5.03 \cdot 10^{12}$	$5.03 \cdot 10^{11}$
Point image radius for $\text{NA} = 0.6$	$1.02 \mu\text{m}$	$0.10 \mu\text{m}$	$0.01 \mu\text{m}$
Number of photons per resolvable image point	$1.6 \cdot 10^8$	$1.6 \cdot 10^5$	$1.6 \cdot 10^2$
Statistical noise \sqrt{N}/N	0.008 %	0.25 %	7.81 %

When using lasers, one should take into account that the stimulated emission results in a number of in-phase photons per laser mode and the number of statistically independent photons is consequently substantially lower. The signal-to-noise ratio SNR is then given by the number of independent modes M with $\text{SNR} = \sqrt{M}/M$. The statistical effects, which occur to a much larger extent when using lasers in comparison with thermal light sources, are usually denoted as speckle [19-30], [19-31], [19-32].

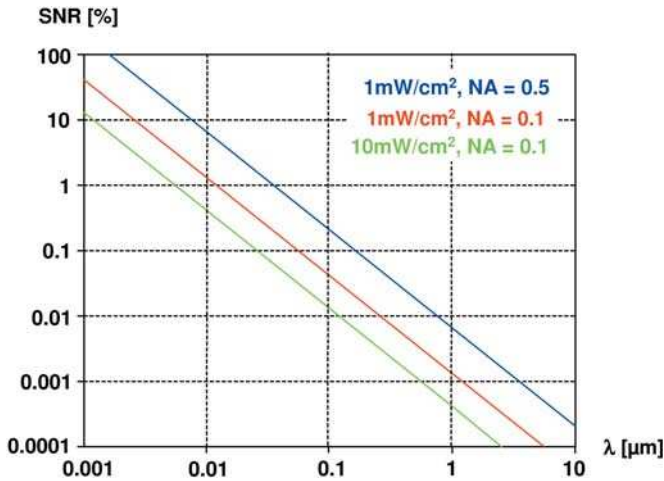


Figure 19-56: SNR per image point for different apertures and intensities.

19.6.2

The Speckle Effect

When highly coherent laser light is incident on a rough surface, a grainy area instead of a uniform light spot is observed. This is the so-called speckle pattern. The light of a laser beam scattered on the wall forms a complex standing wave field. When the roughness is of the same order or larger than the wavelength λ , the interference from different points of the surface possesses a statistical phase. The grainy pattern can be seen sharply by focussing on an arbitrary point in space.

Since in this type of consideration the exact optical path differences depend on the observation point, the exact shape of the speckle pattern is a function of the observation position and vice versa, and the grainy structure seems to move and change when the observer is moving with respect to the scattering surface. This form of the effect is called objective speckle.

Considering an illuminated area of size D and observation at a distance z , one obtains from the outermost points of the illuminated field, a fringe width of

$$d = \frac{\lambda \cdot z}{D} = \frac{\lambda}{2 \sin u} \approx \frac{1}{2} \cdot D_{\text{airy}}. \quad (19-172)$$

This lateral parameter is the smallest grain diameter that can appear in the speckle pattern. This is the typical grain size of the so-called objective granulation, which occurs without additional optics. Depending, however, on the exact form of the surface segment in the aperture, for a given observation point, the grain size can exceed the Airy diameter. The diffraction pattern represents a lower limit for the speckle structural size.

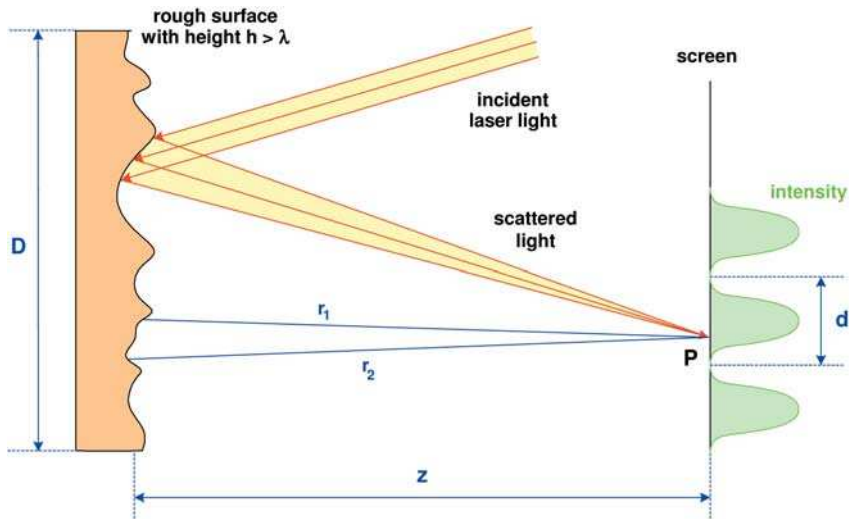


Figure 19-57: Occurrence of speckle upon reflection of coherent radiation at a rough surface.

Since, according to eq. (19-172) the size of typical speckle regions is proportional to the separation z , the speckle image typically depends on the numerical aperture and the distance of the image plane, respectively. This is shown in figure 19-58, using several examples.

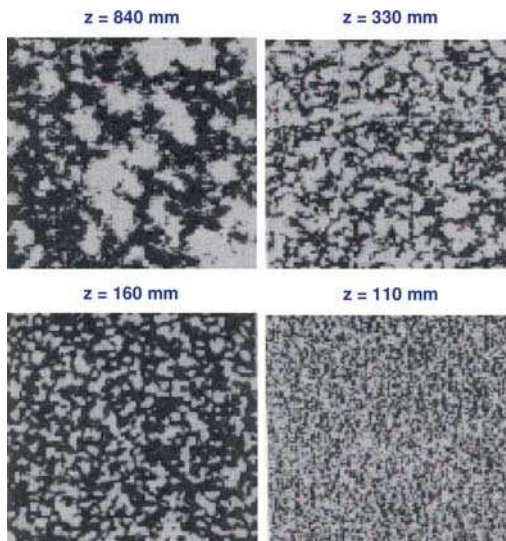


Figure 19-58: Speckle images corresponding to different distances z .

19.6.3

Speckle Parameters and Surface Structure

Let the function $h(x,y)$ describe the surface contour of a scattering area which is illuminated by coherent light with amplitude $U_{in}(x,y)$ at an angle of incidence θ and let $C_h(\Delta x, \Delta y)$ be the normalized autocorrelation function of the surface. In the most general case the relationship between the surface properties and the reflected wave is very complicated because the following effects have to be considered:

1. Fresnel reflection at different local angles.
2. Multiple scattering.
3. Vignetting.
4. Extremely small radii of curvature of the surface, which are comparable with the wavelength.

Assuming an average reflectivity of r one obtains as an approximation for the reflected wave amplitude

$$U_r(x, y) = r \cdot U_{in}(x, y) \cdot e^{ik \cdot (1 + \cos \theta) \cdot h(x, y)} . \quad (19-173)$$

Denoting by σ_h^2 the variance of the surface roughness and assuming a Gaussian dependence of the surface function, one obtains for the complex coherence factor of the reflected field

$$\gamma(\Delta x, \Delta y) = \langle e^{ik(1 + \cos \theta) \cdot [h(x_2, y_2) - h(x_1, y_1)]} \rangle = e^{-k^2(1 + \cos \theta)^2 \cdot \sigma_h^2 \cdot [1 - C_h(\Delta x, \Delta y)]} . \quad (19-174)$$

Assuming further an isotropic Gaussian distribution of the surface correlation function with a coherence length r_c

$$C_h(r) = e^{-\left(\frac{r}{r_c}\right)^2} \quad (19-175)$$

where r is the linear distance, it is possible to obtain curves for the dependence of the coherence factor on the coordinate difference for different variances of the surface, as shown in figure 19-59.

The number of correlation domains which contribute to the intensity when considering the imaging of a scattering surface with a pupil area A_{pup} and a corresponding surface roughness, amounts by use of (19-172) to

$$N = \frac{\lambda^2 \cdot z^2}{A_{pup} \pi \cdot r_c^2} . \quad (19-176)$$

The average contrast in the speckle field

$$\langle V \rangle = \frac{\sigma_I}{\langle I \rangle} \quad (19-177)$$

has a dependence on the roughness variance for different N as shown by the curves in figure 19-60. It should be noticed here, that this modified definition of the contrast contains statistical quantities and therefore does not correspond to the classical normalized definition of V . A contrast of $V = 1$ is approached only for sufficiently large peak-to-valley heights. The contrast slowly increases also with the number N of the contributing correlation cells. For small variances the contrast exhibits a quadratic dependence.

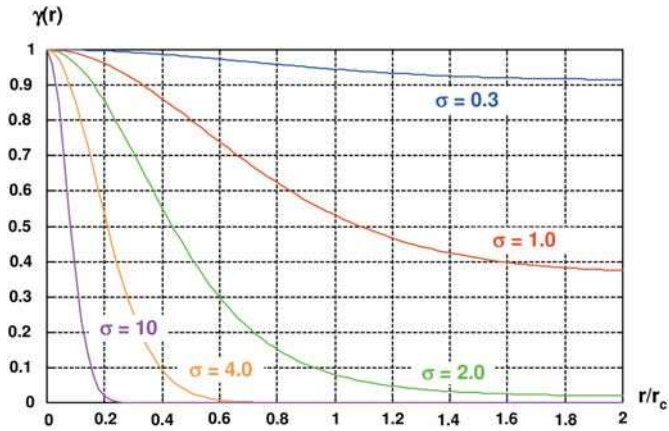


Figure 19-59: Coherence factor as a function of the normalized distance r/r_c and the roughness σ .

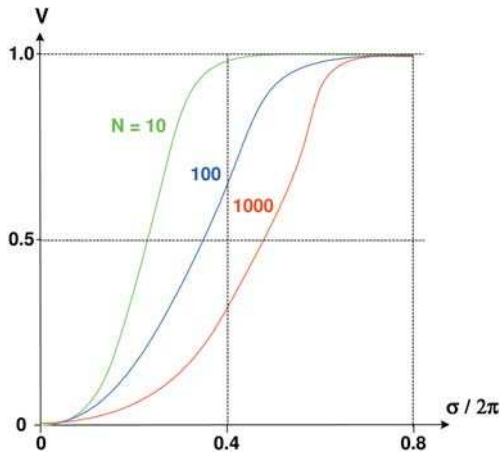


Figure 19-60: Contrast of a speckle structure as a function of the roughness and the number of correlation cells.

19.6.4

Computation of Speckle Effects

Consider an optical system in which an incident coherent field distribution can take many paths with different optical lengths to a single point in the plane of observation. The field distribution results from the interference of all contributing fields. In general the path lengths will depend on time – for instance if a rotating diffuser is considered. The instantaneous field is described by (cf. figure 19-61)

$$U(\vec{r}, t) = \sum_i U_i(\vec{r}) \cdot e^{-i[\omega_0 t + \Phi_i(t)]} = A(\vec{r}, t) \cdot e^{-i[\omega_0 t + \varphi(\vec{r}, t)]} . \tag{19-178}$$

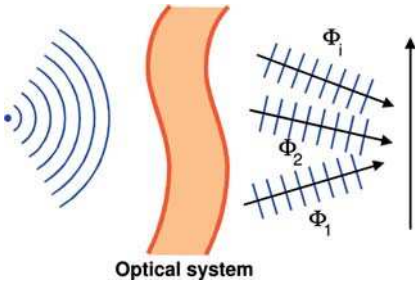


Figure 19-61 Interference of mutually coherent waves of arbitrary phase.

Assuming time independence, i.e. a fixed coupling of all phases, a superposition of constant amplitudes $A(t)$ with equal phases takes place and the resulting phase φ is constant.

If all interfering waves U_i are, on the other hand, of statistical origin, i.e., all amplitudes A_i and relative phases Φ_i are different, the total amplitude $A(r,t)$ and the resulting phase $\varphi(r,t)$ are, in general, position and time dependent. For random signals the amplitude $A(r,t)$ is given by the random walk of the single amplitudes (cf. figure 19-62, right picture).

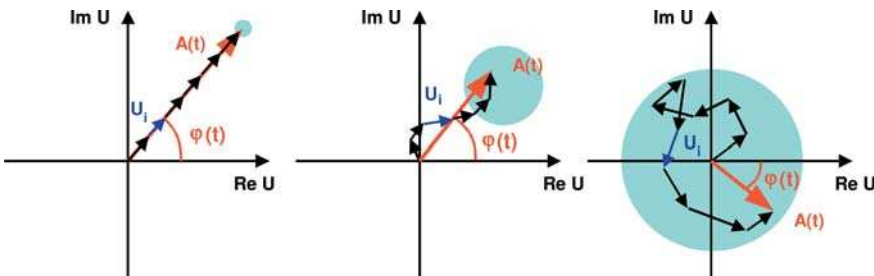


Figure 19-62: N independent field amplitudes are superimposed with correlation decreasing from left to right. The average amplitude is given by the center of the blue circle.

Finally, the correlation of the field distribution $U(t)$ at the position x is given by

$$\langle U(t) U^*(t + \tau) \rangle = \left\langle \sum_i \sum_j U_i U_j^* \cdot e^{i[-\omega_0 \tau + \phi_i(t) - \phi_j(t + \tau)]} \right\rangle. \quad (19-179)$$

For M field distributions of identical phase and amplitude, eq. (19-179) can be rewritten in the form (cf. figure 19-62, left picture)

$$\Gamma_{\text{coh}}(\tau) = M^2 \langle U_i(t) U_i^*(t + \tau) \rangle = M^2 U_i^2 \quad (19-180)$$

which describes the coherent case. In the case of correlated signals the intensity grows quadratically with the number of interfering field distributions. In the other extreme case of uncorrelated field distributions one obtains (cf. figure 19-62, right picture)

$$\Gamma_{\text{incoh}}(\tau) = M \langle U_i(t) U_i^*(t + \tau) \rangle. \quad (19-181)$$

Hence, in the case of uncorrelated signals, time averaging is equivalent to ensemble averaging and the intensity increases linearly with the integration time. That is why in the case of incoherent imaging it makes no difference whether the image is recorded at once, by multiple illumination or, e.g., by scanning [19-33], [19-34].

Figure 19-63 shows typical speckle patterns resulting from the superposition of M independent modes. The contrast (visibility V) reaches a maximum with $V = 1$ for $M = 1$ because the ‘random walk’ can lead to complete mutual extinction of the amplitudes.

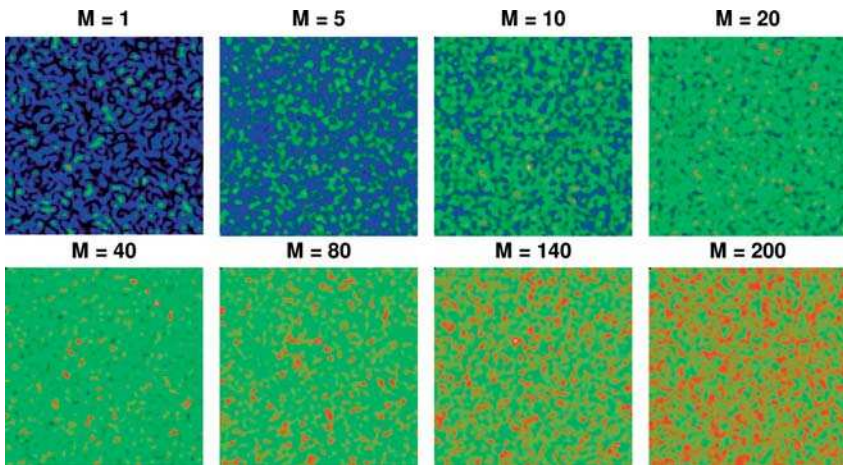


Figure 19-63: Typical speckle patterns resulting from the superposition of M independent modes.

In the case of partially coherent radiation fields, produced by the superposition of statistically independent field distributions and modes, respectively, the overlapping speckle patterns are, depending on the number M of the modes, statistically independent. The RMS-value of the speckle pattern decreases with \sqrt{M}/M . However, one should keep in mind that this does not hold for the contrast (figure 19-64).

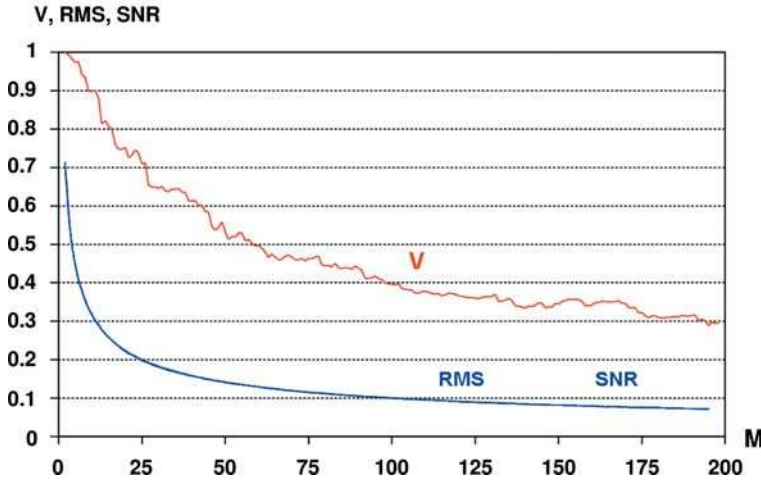


Figure 19-64: RMS and visibility V of a typical speckle pattern resulting from the superposition of M mutually incoherent modes.

In most practical cases, the observed speckle pattern is a superposition of several separate patterns. The overlap of the various beams normally takes place from the same scattering screen. The speckle size remains unchanged, but the statistics and the probability for the occurrence of a definite intensity are different.

In the incoherent case when the correlation of two overlapping distributions is completely absent, the probability distribution of the intensity is given by

$$w(I) = \frac{4I}{I_0^2} \cdot e^{-2\frac{I}{I_0}}. \quad (19-182)$$

Hence, the probability for dark areas with $I = 0$ is vanishing and a concentration with averaged brightness takes place. This trend proceeds for the superposition of an increasing number of patterns and a homogeneous illumination occurs in the limiting case.

In the general case of superimposing M ideal uncorrelated speckle images, the probability distribution for the intensity is given by

$$w_M(I) = \frac{M^M}{(M-1)!} \cdot \frac{I^{M-1}}{I_0^M} \cdot e^{-\frac{M \cdot I}{I_0}}. \quad (19-183)$$

These probability distributions are plotted in figure 19-65 for several values of the number M of the speckle images. The probability tends to have a peak at the value I_0 with increasing number of overlapping single images. Hence, one obtains a homogeneous illumination with constant intensity in the limiting case. This is the underlying principle of speckle reduction.

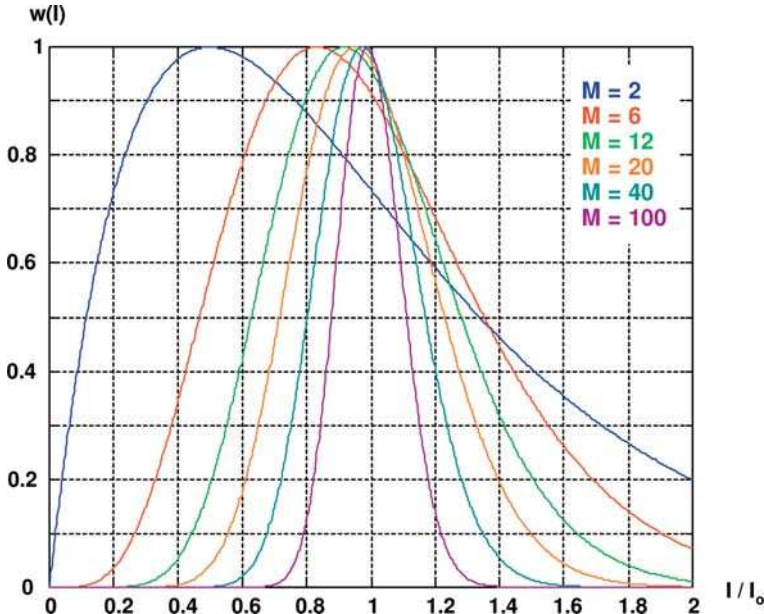


Figure 19-65: Probability distribution of the intensity for incoherent superposition of N speckle images.

For the computation of the intensity one can use the approach described in section 19-4, whereas an additional integration over all modes has to be taken into account. The source distribution is represented as a sum of mutually incoherent modes:

$$I(\vec{r}_s) = \sum_n I_n(\vec{r}_s) = \sum_n |U_n(\vec{r}_s)|^2 . \tag{19-184}$$

For this reason the coherence function is also expressed as a sum over all incoherent modes:

$$\Gamma(\vec{r}_1, \vec{r}_2) = \sum_n \Gamma_n(\vec{r}_1, \vec{r}_2) \tag{19-185}$$

where the coherence functions of the separate modes are defined by

$$\Gamma_n(\vec{r}_1, \vec{r}_2) = \int_{S'} \int_S G(\vec{r}_1 - \vec{r}_s) G^*(\vec{r}_2 - \vec{r}'_s) \cdot U_n(\vec{r}_s) \cdot U_n^*(\vec{r}'_s) d\vec{r}_s d\vec{r}'_s . \tag{19-186}$$

19.6.5

Speckle Reduction

Since speckle is a coherence phenomenon and the coherence, as described above, possesses a spatial-temporal structure, different possibilities for the suppression of speckle effects exist. There are several different approaches possible:

1. Illumination with temporally partially coherent light, which is equivalent to lowering the temporal coherence.
2. Illumination with spatially partially coherent light, which is equivalent to lowering the spatial coherence.
3. Time averaging by a moving aperture.

The optical path lengths in an ideal imaging system are the same for all rays so that disturbing interference effects in the case of coherent illumination have no impact. A real imaging systems under coherent illumination, however, suffers from disturbing interference effects caused in particular by:

- Parasitic interferences by scattering particles (dust, scratches, contaminations) as well as double reflections and, in general, ghost light.
- Air turbulence, particularly in astronomy.

By diffraction or other types of wave-splitting, the wave is decomposed into several mutually coherent waves. These overlap in the observation plane, generating an intensity pattern, which, depending on the origin of the splitting, can have a comparable or higher contrast than the desired image contrast. This speckle pattern is time-independent or stationary for steady-state systems, which indicates the way in which one can avoid speckle effects. Speckle patterns, as interference effects, can be easily affected by temporally changing elements producing different speckle patterns which cancel each other after time averaging [19-35], [19-36], [19-37].

The only possibility of reducing the temporal coherence is, in principle, connected with the manipulation of the radiation spectral bandwidth. Since the coherence length

$$l_c = \frac{\lambda^2}{\Delta\lambda} \quad (19-187)$$

is inversely proportional to the spectral bandwidth, in order to suppress coherence effects, the bandwidth should be increased.

In order to lower the spatial coherence different setups are possible:

1. Insert a moving scattering plate (diffusing plate, ground glass) into the beam path. This plate produces a statistical phase-mixing across the beam cross-section as is shown in figure 19-66.

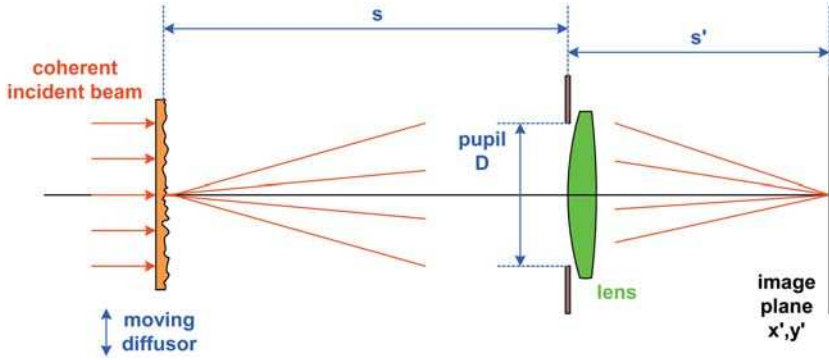


Figure 19-66: Speckle elimination with the help of a moving scattering plate.

Let the surface relief of the scattering plate be given in the normal plane by the height function $h(x, y)$ with vanishing average and a RMS-value of

$$\langle h^2(x, y) \rangle = \delta^2 . \quad (19-188)$$

If the lateral roughness scale is given by σ , then the autocorrelation function of the scattering plate reads

$$C_{\text{Diff}} = \langle h(x, y)h(x + \Delta x, y + \Delta y) \rangle = \delta^2 \cdot e^{-\frac{\Delta x^2 + \Delta y^2}{\sigma^2}} . \quad (19-189)$$

A displacement of the diffusing plate produces a new field distribution through the imaging optics. The cross-correlation between the light amplitudes with and without displacement of the scattering plate indicates the displacement for which a new uncorrelated granulation is obtained, that can be statistically overlapped with the old one. Let D be the pupil diameter of the optical system and s the distance between the pupil and the scattering plate. Assuming that the cone of the scattered light fills the pupil, one has

$$\frac{\sigma}{\delta \cdot k} \ll \frac{\lambda \cdot s}{D} \quad (19-190)$$

and one obtains for the correlation of the field amplitudes in the image plane x', y' for a displacement $\Delta x', \Delta y'$

$$\langle U(x', y')U^*(x' + \Delta x', y' + \Delta y') \rangle = \langle U(x', y')U^*(x', y') \rangle \cdot \text{sinc} \frac{\Delta x' \cdot D}{\lambda \cdot s'} \cdot \text{sinc} \frac{\Delta y' \cdot D}{\lambda \cdot s'} . \quad (19-191)$$

Thus, for a displacement of the scattering plate given by

$$\Delta x' = \frac{\lambda \cdot s}{D} \quad (19-192)$$

one observes a sufficiently strong decorrelation in the image plane. The speckle suppression with a moving scattering plate is, in principle, carried out as ensemble averaging on the time axis. However, a prerequisite for this is to have detector integration times, which are sufficiently large in comparison with the characteristic time constant of the motion.

2. An improved modification of the method is based on two sequential scattering plates of which only one is moving. The double scattering results in considerable improvement of the phase modulation and hence strongly improves the speckle suppression. This principle is shown in figure 19-67.

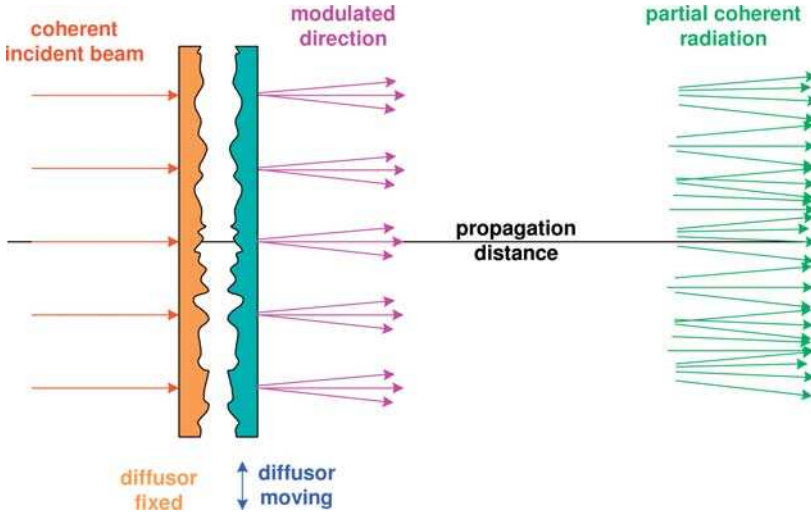


Figure 19-67: Speckle elimination by two diffusers, one is moved.

One obtains the following expression for the autocorrelation in the image plane

$$\begin{aligned} & \langle U(x', y') U^*(x' + \Delta x', y' + \Delta y') \rangle \\ &= \langle U(x', y') U^*(x', y') \rangle \cdot e^{-\frac{k^2 \cdot (\Delta x'^2 + \Delta y'^2)}{(\sigma_1/\delta_1)^2 + (\sigma_2/\delta_2)^2}} \operatorname{sinc} \frac{\Delta x' \cdot D}{\lambda \cdot s'} \cdot \operatorname{sinc} \frac{\Delta y' \cdot D}{\lambda \cdot s'} . \end{aligned} \quad (19-193)$$

The minimum required displacement of the moving scattering plate is given, for equal parameters of the two diffusers, by

$$\Delta x' = \frac{\lambda}{\sqrt{2\pi}} \cdot \frac{\sigma}{\delta} . \quad (19-194)$$

This expression is independent, in contrast to the arrangement with a single diffuser, of the geometrical imaging parameters, and is a function only of the diffuser characteristics. It can be verified numerically that this approach requires a substantially smaller displacement, which is more convenient for practical realizations.

19.7

Array Homogenizer

19.7.1

Setup of the System

In this section, as a concrete application of coherence theory, the intensity homogenization of a partial coherent beam with a lenslet array component and a condenser lens will be demonstrated. Figure 19-68 illustrates the geometry of the array component. The array consists of N lenses. A single lenslet has a focal length f_{arr} and a diameter

$$D_{\text{sub}} = \frac{D_{\text{arr}}}{N} . \quad (19-195)$$

The array is modeled as a thin element and can be described in this approximation as a thin mask. A ray reaches the array at the height x_{ray} in the subaperture No. j . If the center of this lens is x_{cent} with

$$x_{\text{cent}} = D_{\text{sub}} \cdot \left(j - \frac{N+1}{2} \right) \quad (19-196)$$

the change in direction of the ray is in paraxial approximation

$$\Delta u(x) = -\frac{x_{\text{ray}} - x_{\text{cent}}}{f_{\text{arr}}} \quad (19-197)$$

and the new ray direction is

$$u_{\text{out}}(x) = u_{\text{in}}(x) + \Delta u(x) \quad (19-198)$$

A single lenslet for one sub-aperture is described in paraxial approximation by the transmission function

$$T_{\text{sub}}(x) = e^{-\frac{i(x_{\text{ray}} - x_{\text{cent}})^2}{2\lambda f_{\text{arr}}}} . \quad (19-199)$$

The whole array therefore can be modeled by the transmission function [19-38], [19-39]

$$T_{\text{arr}}(x) = \left[\text{rect}\left(\frac{x}{D_{\text{sub}}}\right) \cdot e^{\frac{ix^2}{2\lambda f_{\text{arr}}}} \otimes \text{comb}\left(\frac{x}{D_{\text{sub}}}\right) \right] \cdot \text{rect}\left(\frac{x}{D_{\text{arr}}}\right) . \quad (19-200)$$

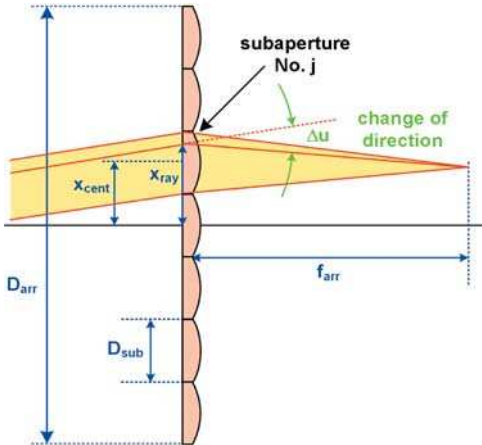


Figure 19-68: Ray trace at an array component.

The complete system for the homogenization of an intensity profile consists of the array and a condenser lens with focal length f_{con} . Figure 19-69 shows the principle of the setup. The array splits the incoming beams into the subaperture beamlets. Every beamlet is propagated into the receiver plane. All the light of the subapertures is superimposed there. Within the paraxial approximation of the propagation of the beamlets according to Fresnel approximation, the size of the illuminated area is given by

$$D_{ill} = D_{sub} \cdot \frac{f_{con}}{f_{arr}} \tag{19-201}$$

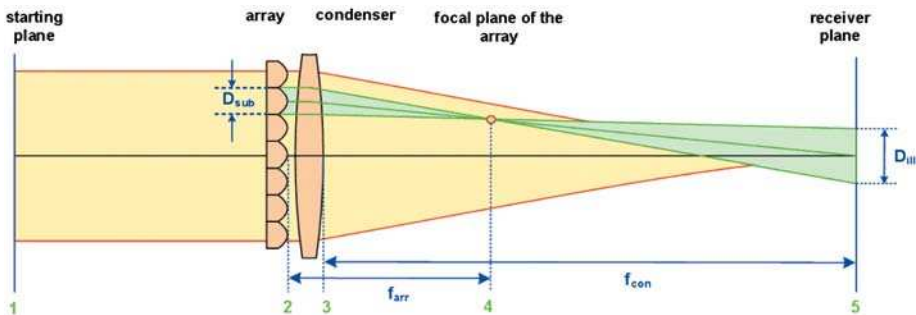


Figure 19-69: Array illuminator system to homogenize the intensity distribution in one selected plane. The numbers indicate the planes for which the calculation results are shown in figure 19-76.

Every subaperture contains a different part of the incoming intensity profile. The superposition of all these contributions gives a good homogenization, if a large

number of subapertures is illuminated. This principle works well in the ideal incoherent case, where the intensities are added in the receiving plane.

One important parameter for the design and optimization of the diffraction array homogenizer is the Fresnel number of a single subaperture, which is defined by (cf. eq. 18-94)

$$N_F^{(\text{sub})} = \frac{\pi D_{\text{sub}}^2}{\lambda \cdot f_{\text{arr}}} \quad (19-202)$$

The illumination field is located in the Fresnel range of the subaperture. Therefore, if the light inside one subaperture can be regarded as coherent, the intensity distribution of each subaperture shows typical Fresnel structure with diffraction ripple. The number of the ripple roughly corresponds to the Fresnel number of the propagation. The larger the Fresnel number, the better the homogeneity of the field and the easier it is to make the gradient of the illumination field boundary steep. But if the size of the desired illumination field is given, a large Fresnel number causes a small number of subapertures, so the superposition and the generation of the mean do not work well.

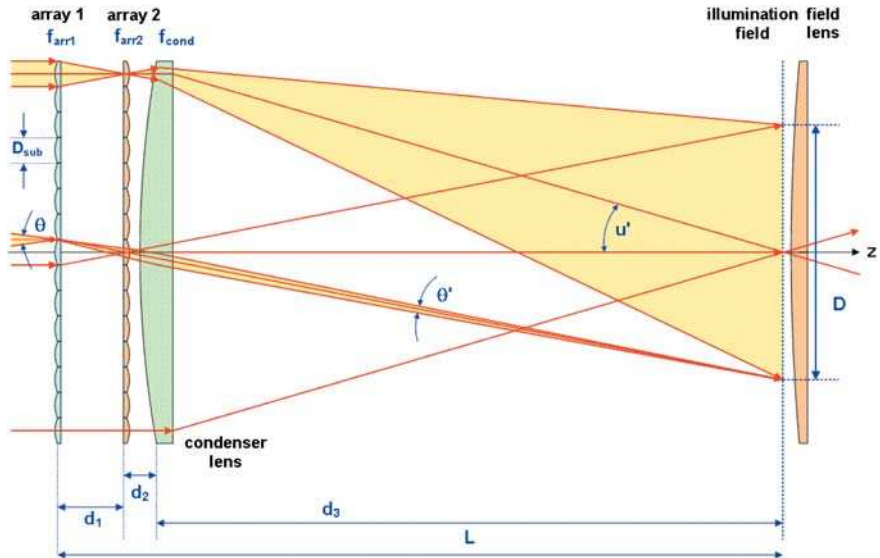


Figure 19-70: Full setup of a two-staged illuminator with two arrays.

So, a more sophisticated setup of a homogenizing array illuminator is shown in figure 19-70. The concept of this two-staged system shows a significant better performance. The first array focusses the incoming collimated light beam onto the second array, which acts as a field lens. The second array therefore images the first array into the plane of homogeneous illumination. One advantage of this more complicated design is a nearly perfect decoupling of the output illumination field from the properties of the incoming light beam. Furthermore, the finite divergence angle θ of the beam does not cause the field edges to be less sharp in the receiving plane.

19.7.2

Pupil Filling

In the plane of the homogenized illumination field, the distribution of the light is not homogeneous with respect to the angle of propagation. This corresponds to a structured illumination of the pupil. The maximum aperture angle reads

$$\sin u'_m = \frac{(N - 1) \cdot D_{arr}}{2f_{cond}} \tag{19-203}$$

In the center of the light field, the angular distribution of the illumination light concentrates around the lines, which connects the point with the center points of the lenslets of the array. Around this selected direction, the light is distributed according to the angular divergence θ of the incoming beam. Depending on this value in comparison to the aperture angle of the individual subapertures $u'/(N-1)$, the angular distribution forms a structured illuminated array and separates in the extreme case into single illumination areas in the pupil. Figure 19-71 shows a section through the angle u' for these two different cases, figure 19-72 illustrates the pupil illumination for the case of isolated areas in two dimensions.

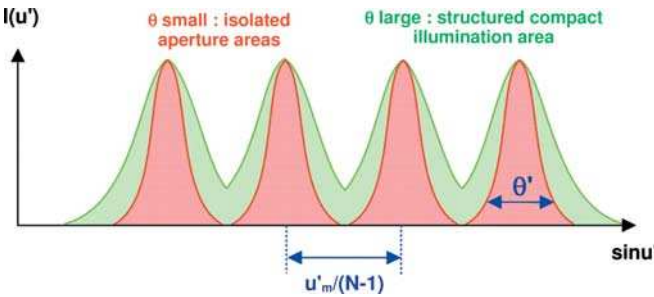


Figure 19-71: Composition of the pupil illumination.

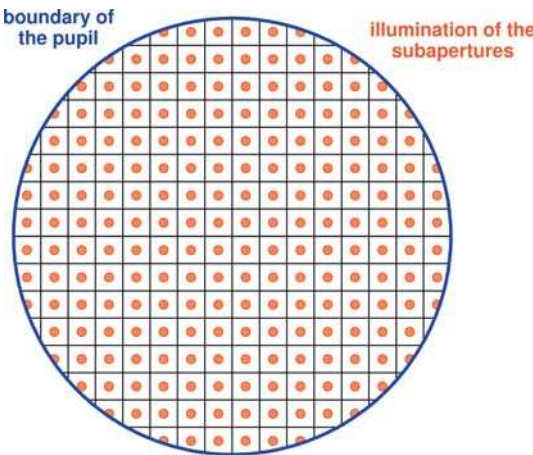


Figure 19-72: Structured pupil illumination for the array homogenizer in two dimensions.

Since every subaperture focusses its individual beam, this decomposition of the pupil illumination can be regarded as a new secondary source of point sources in the corresponding distance of the illumination field. The concept for this idea is shown in figure 19-73.

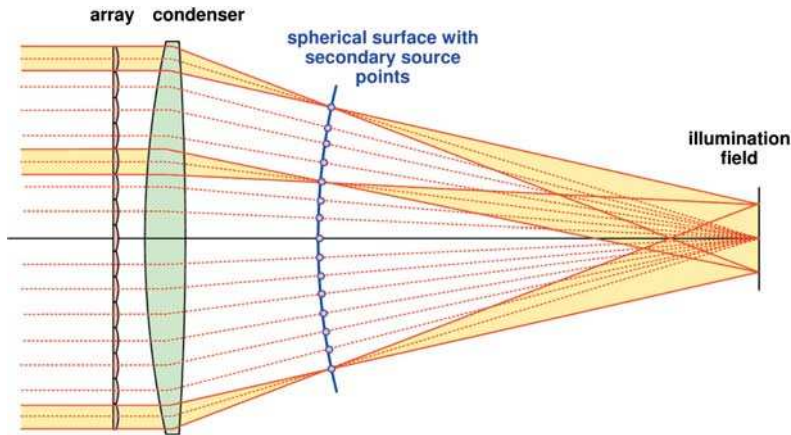


Figure 19-73: Model of the secondary point sources.

The structured illumination of the pupil in the angular domain causes the effect that the homogeneity of the illumination is destroyed immediately, if the receiver is defocussed against the ideal illumination plane.

19.7.3

Coherence Effects

According to the model of the secondary point sources, the occurrence of diffraction and interference effects in the simple diffraction array illuminator can be understood easily [19-40], [19-41]. If the incoming beam has a non-zero transverse coherence length, the different point sources can interfere with a finite contrast and the generation of the resulting illumination corresponds to grating diffraction according to the picture in figure 19-73. Figure 19-74 defines the relevant terms for the coherence consideration of the array setup.

With the help of the Hopkins formula, the degree of coherence can be written with these parameters in the form

$$\gamma(x) = \left| \frac{\sin\left(\frac{\pi \cdot N \cdot x \cdot X'}{\lambda \cdot d}\right)}{N \cdot \sin\left(\frac{\pi \cdot x \cdot X'}{\lambda \cdot d}\right)} \cdot \frac{\sin\left(\frac{\pi \cdot x \cdot d'}{\lambda \cdot d}\right)}{\frac{\pi \cdot x \cdot d'}{\lambda \cdot d}} \right|. \quad (19-204)$$

This function is shown in figure 19-75. It resembles the grating function. The first term is the fast oscillating one and results from the subaperture interference. The second factor is the envelope function and comes from the subaperture profile itself.

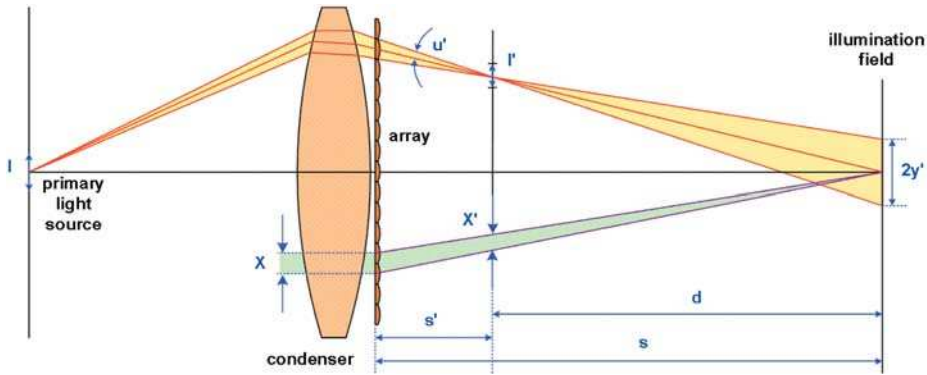


Figure 19-74: Coherence model for the simple array illuminator.

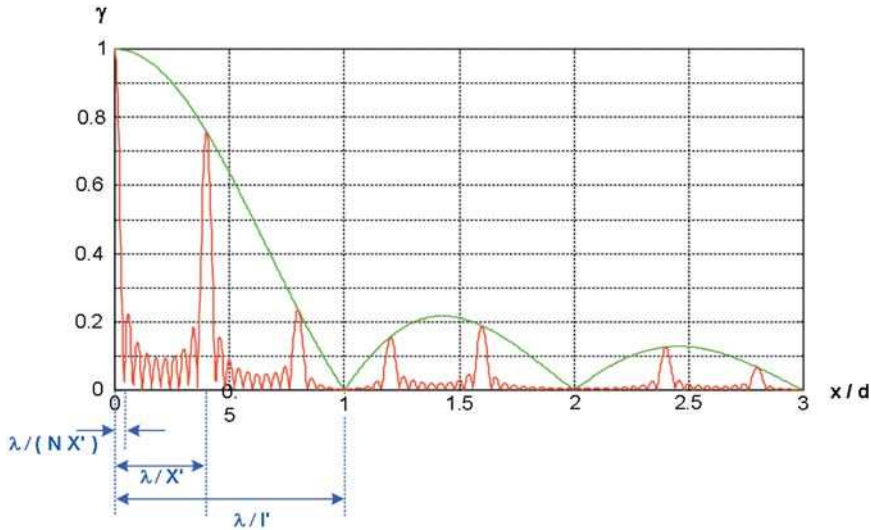


Figure 19-75: Coherence factor of the simple array illuminator.

19.7.4

Example Calculation

If the incoming beam has a certain degree of coherence, a residual non-uniformity or speckle effect remains in the receiving plane. This effect depends on the relation between the lateral coherence length L_c of the incoming beam to the size of the beamlets D_{sub} . Figure 19-76 shows a calculation of the above setup with the plot of the Wigner distribution function and the intensity distribution as a function of the spatial and the angle coordinate in the selected planes 1 to 5 as indicated in figure 19-69. As an incoming beam, a Gauss Schell model beam is used. In the second

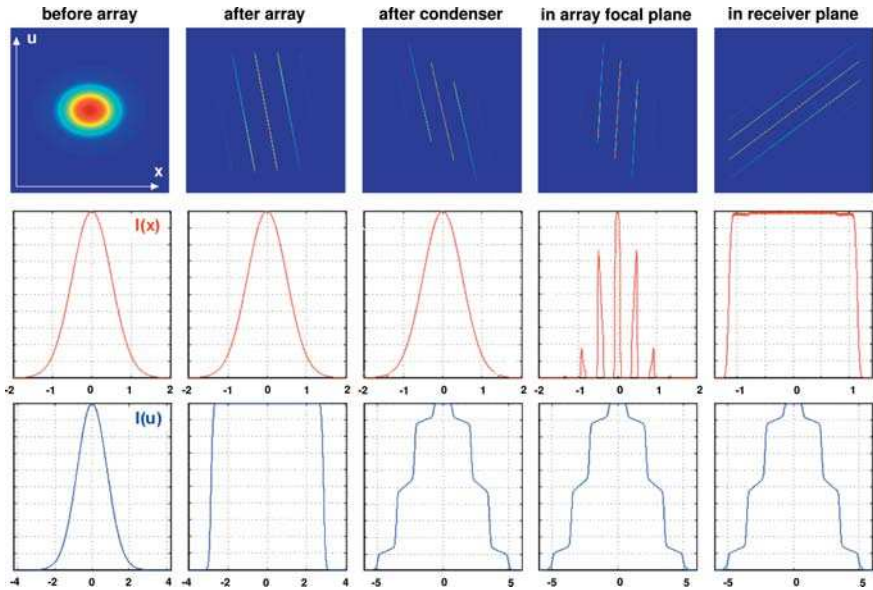


Figure 19-76: Wigner distribution function, in the selected planes of figure 19-69, for an array illuminator system with partially coherent illumination. The second row shows the intensity as a function of the spatial coordinate $I(x)$, the

last row shows the angular distribution of the intensity $I(u)$. It should be noted that the coordinate scales in the first picture of the row is stretched by a factor of 10 in comparison with the other four cases.

column of the series of plots it can be seen that the array has divided the beam into beamlets. The fourth column shows the focussing of the beams in each subaperture, and in the fifth column, the nearly perfect homogenized intensity profile in the spatial domain can be recognized in the second row. It should be noticed that the angle distribution is not homogeneous, as can be seen in the right picture of the last row.

In the right picture of the second row, the residual speckle as a result of the coherence of the incoming beam shows the limits of the setup. The pictures in the bottom row, up to the third, indicate the structured illumination of the pupil as an angle distribution.

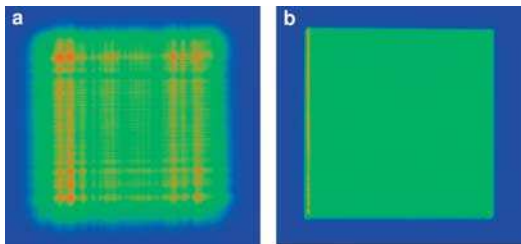


Figure 19-77: Measured intensity profile of an excimer laser, homogenized by: a) a simple array; and b) a double array setup.

As a practical example, figure 19-77 shows the measured beam profile of an excimer laser in the left part, which is homogenized by a single array illuminator. The residual interference effects can be seen very clearly. If, on the other hand, the same beam is homogenized with a full system with two arrays, according to the arrangement of figure 19-70, the observed profile shows a much better homogeneity as can be seen in the right side of figure 19-77. Here only negligible residual diffraction effects appear in the illumination field.

19.8 Miscellaneous

19.8.1

General Coherence Length

A general averaged transverse coherence length can be defined over the cross-section of a partially coherent beam as a normalized second moment of the square of the coherence function [19-42]

$$L_c = \sqrt{\frac{\iint |\Gamma(\vec{r}_1, \vec{r}_2)|^2 \cdot |\vec{r}_1 - \vec{r}_2|^2 d\vec{r}_1 d\vec{r}_2}{\iint |\Gamma(\vec{r}_1, \vec{r}_2)|^2 d\vec{r}_1 d\vec{r}_2}}. \quad (19-205)$$

This coherence length can be interpreted as an average lateral size of the coherence region within the beam cross-section. The fact that this moment is defined, not by the contrast and hence by the absolute value of Γ , but instead by its square, is related to the characteristic properties of the correlation function. The averaging requires integration over both coordinates. Each integral itself transforms in the limiting case of equal coordinates, into an integral over the intensity, giving the power as a physically meaningful quantity.

Substitution of the Wigner function for the coherence function results in an alternative formula for the computation of the general coherence length

$$L_c = \sqrt{\frac{\iint \left(\frac{\partial W(\vec{r}, \vec{p})}{\partial \vec{p}}\right)^2 d\vec{r} d\vec{p}}{\iint W^2(\vec{r}, \vec{p}) d\vec{r} d\vec{p}}} = \sqrt{\frac{\iint \left(\hat{\vec{p}} \cdot \nabla_{\vec{u}} W(\vec{r}, \vec{p})\right)^2 d\vec{r} d\vec{p}}{\iint W^2(\vec{r}, \vec{p}) d\vec{r} d\vec{p}}}. \quad (19-206)$$

In the one-dimensional case, the definition of the transverse coherence length should be supplemented with an additional scaling factor of 2

$$L_c = \sqrt{\frac{2 \cdot \iint |\Gamma(x_1, x_2)|^2 \cdot |x_1 - x_2|^2 dx_1 dx_2}{\iint |\Gamma(x_1, x_2)|^2 dx_1 dx_2}}. \quad (19-207)$$

For a Gaussian Schell beam with a coherence function in the waist given by

$$\Gamma(x_1, x_2) = I_0 \cdot e^{-\frac{x_1^2+x_2^2}{w^2}} e^{-\frac{(x_1-x_2)^2}{2L^2}} \tag{19-208}$$

one obtains the expression

$$L_c = \frac{L}{\sqrt{1 + \left(\frac{L}{w}\right)^2}} \tag{19-209}$$

In the limiting case of ideal coherence the last expression gives

$$L_c(L \rightarrow \infty) = w \tag{19-210}$$

For strongly incoherent beams, the asymptotically obtained coherence length corresponds to that of the conventional Gaussian Schell beam definition.

$$L_c(L \rightarrow 0) = L \cdot \left(1 - \frac{L^2}{2w^2}\right) \tag{19-211}$$

The lateral coherence length can be measured with the help of Young’s double pin-hole experiment. As a characteristic parameter, one records in this case the contrast V of the fringe pattern in dependence on the decreasing hole separation. This quantity equals the absolute value of the normalized coherence function

$$V(x_2 - x_1) = |\Gamma(x_2, x_1)| \tag{19-212}$$

Typically, the half-width a of the decaying contrast curve is taken as a lateral measure of the coherence but a definite relationship with the above defined quantity L_c can be obtained only for given forms of the decaying coherence curve.

1. For a contrast with Gaussian decay shape one has

$$a_{\text{fwhm}} = \sqrt{\ln 2} \cdot L_c = 0.833 \cdot L_c \tag{19-213}$$

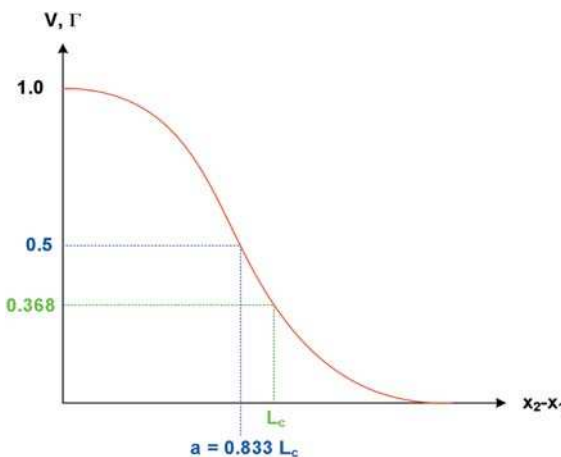


Figure 19-78: Contrast curve with Gaussian shape.

2. For a top-hat coherence function it follows that

$$a_{fwhm} = \sqrt{\frac{3}{2}} \cdot L_c = 1.225 \cdot L_c . \tag{19-214}$$

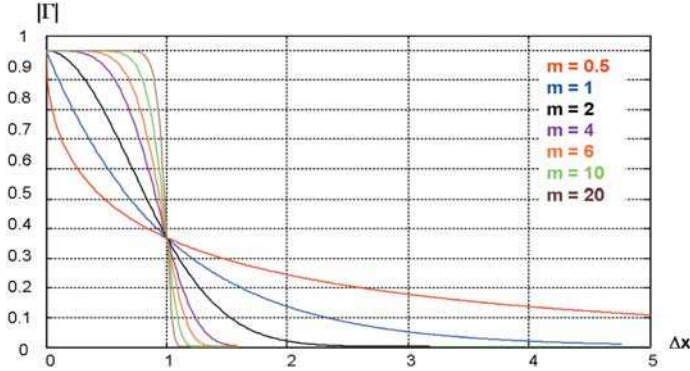


Figure 19-79: Contrast curves with super-gaussian shapes.

In the general case of coherence functions with super-gaussian profile and shapes such as those shown in figure 19-79 for an exponent $m=0.5 \dots 20$,

$$\Gamma(\Delta x) = e^{-\Delta x^m} , \tag{19-215}$$

it is possible to define the transverse coherence length in several different ways:

1. Second moment with Γ

$$L_{c1} = \sqrt{\frac{2 \cdot \iint |\Gamma(x_1, x_2)| \cdot |x_1 - x_2|^2 dx_1 dx_2}{\iint |\Gamma(x_1, x_2)| dx_1 dx_2}} . \tag{19-216}$$

2. Second moment with Γ^2

$$L_{c2} = \sqrt{\frac{2 \cdot \iint |\Gamma(x_1, x_2)|^2 \cdot |x_1 - x_2|^2 dx_1 dx_2}{\iint |\Gamma(x_1, x_2)|^2 dx_1 dx_2}} . \tag{19-217}$$

3. Threshold value of Γ at 50 %

$$|\Gamma(x_2, x_1)|_{L_{c3}=x_2-x_1} = 0.5 . \tag{19-218}$$

4. Threshold value of Γ at $1/e = 0.3678$

$$|\Gamma(x_2, x_1)|_{L_{c4}=x_2-x_1} = 1/e = 0.3678 . \tag{19-219}$$

Depending on the exponent m , one obtains for the lateral coherence length L_c , the curves shown in figure 19-80. The corresponding distributions of the coherence function are scaled by their $1/e$ -width. Realistic values of $m > 1.3$ (the contrast decay converges significantly to zero) result in differences in the coherence lengths as large as a factor of 5, depending on the curve shape of $\Gamma(\Delta x)$.

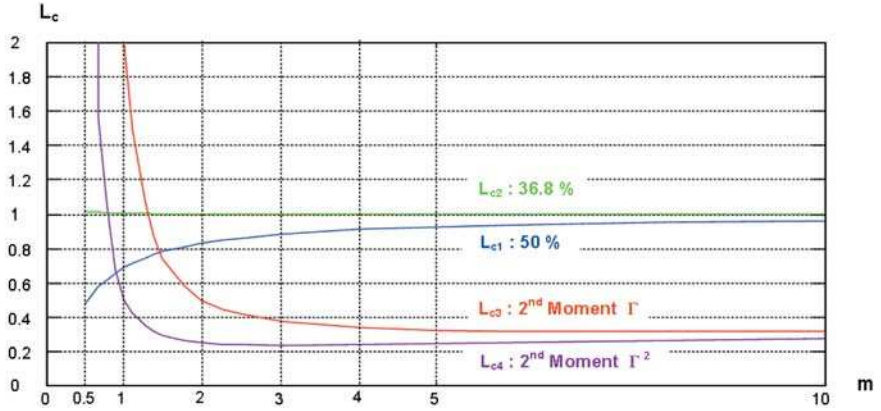


Figure 19-80: Several definitions of the transverse coherence length for contrast decays with super-gaussian shape as shown in figure 19-79.

19.8.2

General Degree of Coherence

A definition of the degree of coherence with global validity and an averaging property across the beam cross-section can be formulated with the help of the coherence or the Wigner function in the form

$$\gamma^2 = \frac{\iint |\Gamma(x_1, x_2)|^2 dx_1 dx_2}{\left(\int \Gamma(x, x) dx\right)^2}, \tag{19-220}$$

$$\gamma^2 = \lambda \cdot \frac{\iint W'^2(x, p) dx dp}{\left(\int \int W'(x, p) dx dp\right)^2}. \tag{19-221}$$

19.8.3

Coherence and Polarization

The polarization, or the composition, of the directional components of the electromagnetic field, remains unchanged for free-space propagation. Hence, a separate coherence function can be defined for each vector component and the consideration can be limited to the scalar case.

Passage through optical systems, however, can result in cross-talk between the vector components, e.g., an x -component at the input of an optical system can be transformed into a y -component at the output. Hence, an x -component of the light source can actually interfere with a y -component. Because of that mutual correlation functions are defined for the different polarization components. The resulting polarization coherence function matrix has the form

$$\underline{\Gamma}(\vec{r}_1, \vec{r}_2) = \begin{pmatrix} \Gamma_{xx} & \Gamma_{xy} & \Gamma_{xz} \\ \Gamma_{yx} & \Gamma_{yy} & \Gamma_{yz} \\ \Gamma_{zx} & \Gamma_{zy} & \Gamma_{zz} \end{pmatrix} \quad (19-222)$$

where the mutual coherence functions of the light source are given by

$$\Gamma_{ij}(\vec{r}_s, \vec{r}'_s) = \iint_{S'} \iint_S E_{i,s}(\vec{r}_s) \cdot E_{j,s}^*(\vec{r}'_s) d\vec{r}_s d\vec{r}'_s. \quad (19-223)$$

The polarization coherence function Γ_{xy} characterizes the capability of the different vector components to interfere. Through the correlation of two polarization directions the polarization coherence function describes the polarization degree of a given field distribution. It has to be noted, that the coherence matrix according to eq. (19-222) is over-specified because it stems from the superposition of two-dimensional Jones vectors (see chapter 28 for more details). For vanishing coherence function the light is unpolarized, otherwise it can be partially polarized or polarized:

$$\begin{array}{ll} \Gamma_{xy} = 1 & \text{polarized,} \\ 0 < \Gamma_{xy} < 1 & \text{partially polarized,} \\ \Gamma_{xy} = 0 & \text{unpolarized.} \end{array} \quad (19-224)$$

Propagation through an optical system can be described in an abstract form by a propagator or a Green function, where the Green function has a matrix form. One has for the field vector in the observation plane

$$\vec{E}(\vec{r}) = \int \underline{G}(\vec{r} - \vec{r}_s) \cdot \vec{E}_s(\vec{r}_s) d\vec{r}_s. \quad (19-225)$$

The cross-talk between different vector components is accounted for in the non-diagonal components of the Green function matrix. Thus the correlation integral can again be determined for each vector component in the image plane. The coherence function itself is also represented by a vector with x,y,z -components:

$$\Gamma_{ij}^k(\vec{r}, \vec{r}') = \int_{S'} \int_S \int_S G_{ki}(\vec{r} - \vec{r}_s) E_{i,s}(\vec{r}_s) \cdot G_{kj}^*(\vec{r}' - \vec{r}'_s) E_{j,s}^*(\vec{r}'_s) d\vec{r}_s d\vec{r}'_s. \quad (19-226)$$

The mutual intensity in the image plane is given now by a triple sum over all diagonals of the polarization coherence functions

$$I(\vec{r}) = \sum_{i,j,k} \Gamma_{ij}^k(\vec{r}, \vec{r}). \quad (19-227)$$

In each case, two arbitrary vector components i and j , in the input plane, can contribute to a vector component, e.g., to the x -component, in the image plane.

The last chapter of this volume will be devoted to further consideration of the polarization coherence function.

19.9

Literature

- 19-1** F. Zernike, *Physica V*, 785 (1938), The concept of degree of coherence and its application to optical problems.
- 19-2** E. L. O'Neill, *Introduction to Statistical Optics* (Dover, New York, 1991).
- 19-3** M. J. Beran and G. B. Parrent, *Theory of Partial Coherence* (Prentice Hall, New York, 1964).
- 19-4** A. S. Marathay, *Elements of Optical Coherence Theory* (Wiley, New York, 1982).
- 19-5** L. Mandel and E. Wolf, *Optical Coherence and Quantum Optics* (Cambridge University Press, Cambridge, 1995).
- 19-6** J. W. Goodman, *Statistical Optics* (Wiley Interscience, New York, 1985).
- 19-7** G. O. Reynold, J. B. DeVelis, G. B. Parrent and B. J. Thompson, *Physical Optics Notebook* (SPIE Press, Bellingham, 1989).
- 19-8** U. Vokinger, *Propagation, Modification and Analysis of Partially Coherent Light Fields* (UFO Atelier für Gestaltung & Verlag, Konstanz, 2000).
- 19-9** J. Shamir, *Optical Systems and Processes* (SPIE Press, Bellingham, 1999).
- 19-10** M. Born and E. Wolf, *Principles of Optics* (Cambridge University Press, Cambridge, 1999).
- 19-11** B. E. Saleh and M. C. Teich, *Fundamentals of Photonics* (J. Wiley & Sons, Hoboken, 1991).
- 19-12** W. Mecklenbräuker and F. Hlawatsch, *The Wigner Distribution* (Elsevier, Amsterdam, 1997).
- 19-13** K.-H. Brenner, *Phasenraumdarstellungen in Optik und Signalverarbeitung*, Dissertation, Erlangen 1983.
- 19-14** D. Dragoman, *Optik* **111**, 179 (2000), The origin of negative values of the Wigner distribution function.
- 19-15** D. Dragoman, *Appl. Opt.* **35**, 161 (1996), Wigner distribution function applied to third-order aberrations.
- 19-16** J. Turunen and A.T. Friberg, *Optics and Laser Techn.* **18**, 259 (1986), Matrix representation of Gaussian Schell-model beams in optical systems.
- 19-17** R. Simon, E.C. G. Sudarshan and N. Mukunda, *Phys. Rev. A* **29**, 3273 (1984), Generalized rays in first-order optics: Transformation properties of Gaussian Schell-model fields.
- 19-18** A. T. Friberg and R. J. Sudol, *Opt. Commun.* **41**, 383 (1982), Propagation parameters of Gaussian Schell-model beams.
- 19-19** E. Collett and E. Wolf, *Opt. Commun.* **32**, 27 (1980), Beams generated by Gaussian quasi-homogeneous sources.
- 19-20** A.C. Schell, *IEEE Trans. AP* **15**, 187 (1967), A technique for the determination of the

- radiation pattern of a partially coherent aperture.
- 19-21** R. Simon and N. Mukunda, *J. Opt. Soc. Am. A* **15**, 1361 (1998), Shape-invariant anisotropic Gaussian Schell-model beams: a complete characterization.
- 19-22** Y. Peng and B. Lü, *Optik* **115**, 1 (2004), Diagonalization and symmetrization of anisotropic twisted Gaussian Schell-model beams.
- 19-23** T. Asakura (Ed.), *International Trends in Optics and Photonics ICO IV* (Springer, Berlin, 1999).
- 19-24** Y. Qiu, H. Guo and Z. Chen, *Opt. Comm.* **245**, 21 (2005), Paraxial propagation of partially coherent Hermite-Gauss Beams.
- 19-25** B. Lü and L. Pan, *Opt. Comm* **205**, 7 (2002), Propagation of vector Gaussian-Schell-model beams through a paraxial optical ABCD system.
- 19-26** L. Mandel and E. Wolf, *Optical Coherence and Quantum Optics* (Cambridge University Press, Cambridge, 1995).
- 19-27** H. J. Caulfield (Ed.), *Optical Information Processing: A Tribute to Adolf Lohmann*, chapt. 4 (SPIE Press, Bellingham, 2002).
- 19-28** J. Tervo, T. Setälä and A. T. Friberg, *J. Opt. Soc. Am. A* **21**, 2205 (2004), Theory of partially coherent electromagnetic fields in the space-frequency domain.
- 19-29** R. Simon and N. Mukunda, *J. Opt. Soc. Am. A* **10**, 95 (1993), Twisted Gaussian Schell-model beams.
- 19-30** J. C. Dainty, *Laser Speckle and Related Phenomena* (Springer, Berlin, 1984).
- 19-31** T. G. Brown (Ed.), *The Optics Encyclopedia*, VOL 1, chapter *Speckle and Speckle Metrology* (Wiley-VCH, Weinheim, 2003).
- 19-32** R. G. Driggers (Ed.), *Encyclopedia of Optical Engineering*, Vol 1, (Marcel Dekker, New York, 2003).
- 19-33** D. Kermisch, *J. Opt. Soc. Am.* **67**, 1357 (1977), Principle of equivalence between scanning and conventional optical imaging systems.
- 19-34** D. J. Cronin, J.B. DeVelis and G. O. Reynolds, *Opt. Eng.* **15**, 276 (1976), Equivalence of annular source and dynamic coherent phase contrast viewing systems.
- 19-35** M. J. Bowman, *Appl. Opt.* **7**, 2280 (1968), Two new methods of improving optical image quality.
- 19-36** D. J. Cronin and A. E. Smith, *Opt. Eng.* **12**, 50 (1973), Dynamic coherent optical system.
- 19-37** R. F. van Ligten, *Appl. Opt.* **12**, 255 (1973), Speckle reduction by simulation of partially coherent object illumination in holography.
- 19-38** A. Büttner and U. D. Zeitner, *Opt. Eng.* **41**, 2393 (2002), Wave optical analysis of light-emitting diode beam shaping using micro-lens arrays.
- 19-39** V. L. Korneichik, *J. Opt. Technol.* **67**, 999 (2000), Diffraction effects in an illuminator with a lens raster.
- 19-40** S. N. Natarovskii, *Opt. Spectrosc.* **64**, 680 (1988), Degree of coherence of object illumination created by a raster illuminator.
- 19-41** S. L. Zlatina, I.P. Kozlova, S. N. Natarovskii, O. N. Menkova, T. F. Selezneva and N. B. Skobeleva, *J. Opt. Technol.* **71**, 764 (2004), Raster illuminators for microscopes.
- 19-42** B. Eppich, *Proc. SPIE* **4270**, 71 (2001), Definition, meaning and measurement of coherence parameters.

20

The Geometrical Optical Description and Incoherent Imaging

- 20.1 Introduction 188
- 20.2 Characteristic Functions 189
 - 20.2.1 Geometrical Optics and the Wave Equation 189
 - 20.2.2 The Characteristic Functions 191
 - 20.2.3 Geometrical-optical imaging 194
 - 20.2.4 The Canonical Pupil 196
 - 20.2.5 A Note on Diffractive Optical Elements 199
- 20.3 The Ideal Wave-optical Image of a Point and Geometrical-optical Image Formation 200
 - 20.3.1 The Scalar Luneburg Integral 200
 - 20.3.2 Energy Discussions for Optical Imaging 204
 - 20.3.3 The Airy Disc 206
 - 20.3.4 Incoherent Resolution 210
- 20.4 Aberrations of Optical Systems 211
 - 20.4.1 The Small-aberration Limit: The Strehl Ratio 211
 - 20.4.2 Expansion of the Wave-front Error into Zernike Polynomials 212
 - 20.4.3 Point Images for Different Aberrations 217
 - 20.4.4 Distortion, Defocus and Astigmatism 219
 - 20.4.5 Spherical Aberrations Z_9 , Coma Z_7 and Z_8 220
 - 20.4.6 Line of Sight 221
 - 20.4.7 Wave Aberrations for Annular Pupils 224
 - 20.4.8 Extended Zernike Expansion 227
- 20.5 Helmholtz–Lagrange Invariant and Phase-space Description 231
 - 20.5.1 The Phase Space 231
 - 20.5.2 The Resolution Limit in the Space Domain and in the Spatial Frequency Domain 234
 - 20.5.3 The Space–Bandwidth Product 236
- 20.6 Literature 237

20.1

Introduction

The design and analysis of optical imaging systems is based on ray-tracing and the geometrical optical description. The advantages of ray-tracing are numerous: it is powerful, fast, flexible and applicable over large scales, and there are a variety of commercially available software packages. With certain approximations and, e.g., statistical methods, even scattering or diffractive optical elements can be considered. However, ray-tracing alone is not sufficient to describe optical imaging. A wave-optical description is necessary in order to consider diffraction and interference effects. Diffraction effects from boundaries like the pupil stop can be neglected in most cases, but diffraction effects at the object and intensity formation in the image plane by interference, are not considered by ray-tracing.

The description of optical imaging is based on the, at first glance, abstract characteristic functions. The characteristic functions, which are in general four-dimensional functions, are based on fundamental geometrical-optical principles such as the minimum principle for the optical light path length, and are usually obtained by ray-tracing, measurement, or – in selected examples – analytically. The transition from geometrical optics to the wave-optical description with the help of the characteristic function allows a wave-optical treatment of optical imaging systems. For example, the calculation of the image intensity of a single point gives the point-spread function. In the case of incoherent imaging, when coherence effects can be neglected, the image is obtained by a convolution of object transmission with the wave-optical point-spread function. Partial coherent or coherent image formation requires a wave-optical treatment of diffraction at the object and interference in the image plane. In any case, however, the description of the imaging system is based on the characteristic functions. The concept of the characteristic function in combination with the transition to the wave-optical description allows comprehensive interpretation of optical imaging.

In section 20.2, the geometrical-optical description is introduced and some fundamental properties of geometrical optics and optical imaging systems are outlined. The characteristic functions are derived using the variation principle, and the transition to wave optics is performed. The point-spread function of ideal incoherent imaging systems is derived in section 20.3. As an example of two-point resolution, incoherent imaging is considered, further examples of incoherent imaging are given in comparison with partial coherent imaging in chapter 24. In section 20.4 aberrations are considered. In section 20.5, some remarks on the Helmholtz invariant and information theory are included. The Fourier-optical description of optical imaging according to Abbe is described in chapter 21. In chapter 22, several mathematical descriptions of partial coherent imaging theories are compared. Three-dimensional image formation is described in chapter 23.

20.2

Characteristic Functions

20.2.1

Geometrical Optics and the Wave Equation

The development of classical optical instruments usually takes place within the geometrical optical approximation. The design of optical systems may be carried out using first-order optical equations such as Newton's equations or third-order Seidel coefficients. Optical systems are further evaluated and optimized by means of ray-tracing. Here only the formal aspects of geometrical optics are introduced in so far as they are useful for comprehension, a detailed description is given in volume 3 of this book series. The reader familiar with the Hamiltonian description of classical mechanics will recognize the terms since, like classical mechanics, geometrical optics is based on a minimum principle and allows application of the calculus of variations. Incidentally, William Rowan Hamilton (1805–1865; publication of his theory from 1828–1837) originally developed his formalism for optics rather than for mechanics. For more details of variation calculus and the formal theory of geometrical optics, the reader is referred to [20-1] and [20-2].

Geometrical optics is based on a fundamental principle, Fermat's law. Fermat repeated an idea of Hero of Alexandria according to which light takes the path with the shortest time of travel. This minimum principle for the time of travel can be converted into a minimum principle for the light path between two (infinitesimally adjacent) points:

$$\delta V = 0, \text{ with } V = c \cdot T = \int_{\nu}^c dt = \int_{\nu}^c ds = \int n(\vec{r}) ds \quad (20-1)$$

where ds is the infinitesimal length element $ds = \sqrt{dx^2 + dy^2 + dz^2}$. The function $V(\vec{r}_0, \vec{r}_1)$ describes the light path between the two points r_0 and r_1 and is called the *point characteristic*. The point characteristic describes the path length of an optical system by the object and image points, \vec{r}_0 and \vec{r}_1 as parameters. The light path between these two points is not otherwise specified in more detail. In a perfectly imaging optical system there can be infinitely many light paths of the same length between object and conjugated image point.

For the description of the light path through optical systems, the infinitesimal length element ds is impractical. With the z-coordinate as parameter instead it follows from (20-1) that

$$V(x_0, y_0; x_1, y_1; z_0, z_1) = \int n ds = \int L(x, y; \dot{x}, \dot{y}; z) dz. \quad (20-2)$$

Equation (20-2) is applicable as long as the object and image space possess axes of symmetry with coordinates z_0 and z_1 . The function L is the Lagrange function, given by

$$L = n \sqrt{\dot{x}^2 + \dot{y}^2 + 1} \quad (20-3)$$

with $\dot{x} = \partial x / \partial z$ and $\dot{y} = \partial y / \partial z$. The canonical conjugated variables to the space coordinates x and y together with the Hamilton function H are given by the Hamilton–Jacobi equations:

$$p = \frac{\partial L}{\partial \dot{x}} = \frac{\partial V}{\partial x} = \frac{n\dot{x}}{\sqrt{1 + \dot{x}^2 + \dot{y}^2}}, \quad (20-4a)$$

$$q = \frac{\partial L}{\partial \dot{y}} = \frac{\partial V}{\partial y} = \frac{n\dot{y}}{\sqrt{1 + \dot{x}^2 + \dot{y}^2}}, \quad (20-4b)$$

$$m = \frac{\partial V}{\partial z} = \frac{n\dot{z}}{\sqrt{1 + \dot{x}^2 + \dot{y}^2}} = -H(x, y; p, q). \quad (20-4c)$$

The canonical conjugates p_x and p_y to the space coordinates x and y are equivalent to the x - and y -components of the tangent vector to a light-ray of length n , also called the optical direction cosine vector. For convenience, the optical direction cosine vector is also written as $\vec{p} = (p, q, m)$. The Hamiltonian is equivalent to the negative of the z -component of this vector. Due to the analogy with classical mechanics, the optical direction cosine vector $\vec{p} = (p_x, p_y, p_z) = (p, q, -H)$ is also called the quasi momentum. Equation (20-4) represents the vector equivalent to the Eikonal equation

$$\vec{p} = \vec{\nabla} V = n\vec{s}. \quad (20-5)$$

From the square of eq. (20-5) the known Eikonal equation (17-103) follows with $|\vec{\nabla} V|^2 = \vec{p}^2 = n^2$. The optical direction cosine vector \vec{p} plays an important role in optics and imaging systems. As will be shown, ideal optical systems have to perform linear transformations to transverse space coordinates x, y and also directions cosines q, p . As will be shown below, the transverse components of the optical direction cosine are also proportional to the canonical pupil coordinates of an imaging optical system.

By analogy with quantum mechanics, the optical direction cosines can be written as operators for a wave function by

$$\hat{p}_x = -i \frac{\lambda}{2\pi} \frac{\partial}{\partial x}, \quad \hat{p}_y = -i \frac{\lambda}{2\pi} \frac{\partial}{\partial y} \quad \text{and} \quad \hat{H} = i \frac{\lambda}{2\pi} \frac{\partial}{\partial z} \quad (20-6)$$

and after insertion into the Hamiltonian we obtain

$$\hat{H}U = i \frac{\lambda}{2\pi} \frac{\partial}{\partial z} U = -\sqrt{n^2 - \left(\frac{\lambda}{2\pi}\right)^2 \frac{\partial^2}{\partial x^2} - \left(\frac{\lambda}{2\pi}\right)^2 \frac{\partial^2}{\partial y^2}} U = 0. \quad (20-7)$$

From the square of eq. (20-7) the Helmholtz equation (17-17) follows:

$$\left[\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} + n^2 \left(\frac{2\pi}{\lambda}\right)^2 \right] U = \Delta U + n^2 k_0^2 U = 0. \quad (20-8)$$

For paraxial rays with small angles to the z-axis with $\dot{x} \ll 1$ and $\dot{y} \ll 1$, the Hamiltonian can be expanded

$$H_p = -\sqrt{n^2 - p^2 - q^2} \approx \frac{1}{2n}(p^2 + q^2) - n \quad (20-9)$$

and the paraxial wave equation is obtained as:

$$\hat{H}_p U = i \frac{\lambda}{2\pi} \frac{\partial}{\partial z} U = \frac{1}{2n} \left(\frac{\lambda}{2\pi} \right)^2 \left[\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right] U - nU = 0. \quad (20-10)$$

By analogy with classical mechanics, the paraxial Hamiltonian H_p can be written as a sum of the 'kinetic energy' T and a potential V , while the refractive index can be interpreted as a potential determining the trajectories of light rays.

20.2.2

The Characteristic Functions

The point characteristic V of eq. (20-2) is the most simple characteristic function although there are several others. Its function value is the optical light path length between two points in the object and the image plane z_0 and z_1 , and may be written as

$$V(\vec{r}_0, \vec{r}_1) = \int_{\vec{r}_0}^{\vec{r}_1} n ds = \int_{\vec{r}_0}^{\vec{r}_1} n \frac{d\vec{r}}{ds} \cdot d\vec{r} = \int_{\vec{r}_0}^{\vec{r}_1} \vec{p} \cdot d\vec{r} = \int_{\vec{r}_0}^{\vec{r}_1} (p dx + q dy - H dz). \quad (20-11)$$

The first integral is equivalent to the path length integral over the infinitesimal curve parameter $ds = |d\vec{r}|$ (eq. (20-1)). It can be converted with the help of the optical direction cosine $\vec{p} = (p, q, m)$ into the Hilbert integral over the scalar product of the direction cosine and position vector (eq. (20-11), right-hand side). The total differential of the point characteristic is thus given by

$$\begin{aligned} dV &= \vec{p}_1 \cdot d\vec{r}_1 - \vec{p}_0 \cdot d\vec{r}_0 \\ &= p_1 dx_1 + q_1 dy_1 + m_1 dz_1 - p_0 dx_0 - q_0 dy_0 - m_0 dz_0. \end{aligned} \quad (20-12)$$

Thus, in an existing optical system, there is a vector field of optical direction cosines to each pair of object and image points, which is given by the gradient of the path length or point characteristic V (eq. (20-5)). The optical path length or point characteristic V takes over the role of the potential in classical mechanics. As is known, the rotation of a gradient field vanishes, i.e., every closed path integral (Hilbert integral) yields the value 0:

$$\text{rot } \vec{p} = 0. \quad (20-13)$$

Equations (20-5) and (20-13) express the commonly known fact that light rays run perpendicular to the planes of equal light-path lengths. The planes of equal light-path lengths $V = \text{const.}$ are thus equivalent to the wave fronts of the wave-optical

description. This property, called orthotomy or normal congruence of rays, is illustrated in figure 20-1 for homogeneous object and image space. The orthotomy is valid throughout optical systems and is especially preserved at boundaries where, as a consequence, Snell's law of refraction can be derived from (20-13).

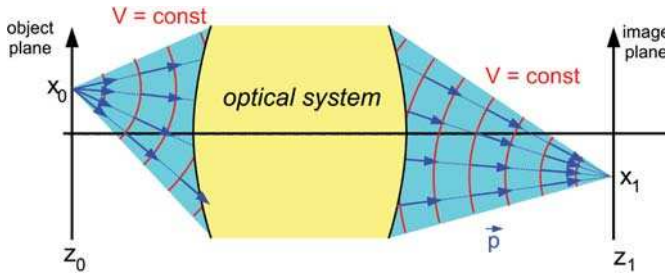


Figure 20-1: Normal congruence of rays or orthotomic system: the light rays run perpendicular to the wave fronts; any closed light path-integral has the length zero.

The conservation of the normal congruence in optical systems is used for energetic consideration of the optical imaging: from the equivalence of the density of the geometrical-optical rays to the energetic radiance the energy transfer from object to the image is given by the conservation of congruence. The radiant flux Φ_1 through a surface element dA_1 , limited by at least three rays in image space, is equivalent to the radiant flux Φ_0 through the corresponding surface element dA_0 limited by the same rays in object space, and may only be weakened by absorption or reflectance losses (figure 20-2).

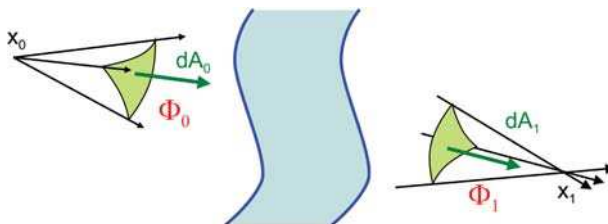


Figure 20-2: Conservation of energy as a consequence of conservation of congruence.

For constant z coordinates of the object and image plane the point characteristic is a four-dimensional function. The characteristic function $V(x_0, y_0; x_1, y_1)$ can be transformed into other characteristic functions by a Legendre transformation. By writing, e.g.,

$$\begin{aligned}
 dW_1 &= d(V - x_1 p_1 - y_1 q_1) \\
 &= -x_1 dp_1 - y_1 dq_1 + m_1 dz_1 - p_0 dx_0 - q_0 dy_0 - m_0 dz_0
 \end{aligned}
 \tag{20-14}$$

the mixed characteristic $W_1(x_0, y_0; p_1, q_1)$ is obtained as a function of the transverse object coordinate and the optical direction cosines in the image space. By further Legendre transformations the point characteristic can be converted further into the

second mixed characteristic $W_2(p_0, q_0; x_1, y_1)$ and into the Schwarzschild angular eikonal $T(p_0, q_0; p_1, q_1)$ [20-3]. So in total there are four characteristic functions:

$$V(x_0, y_0; x_1, y_1) \quad (20-15a)$$

$$W_1(x_0, y_0; p_1, q_1) = V - x_1 p_1 - y_1 q_1 \quad (20-15b)$$

$$W_2(p_0, q_0; x_1, y_1) = V + x_0 p_0 + y_0 q_0 \quad (20-15c)$$

$$T(p_0, q_0; p_1, q_1) = V - x_1 p_1 - y_1 q_1 + x_0 p_0 + y_0 q_0. \quad (20-15d)$$

A characteristic function parameterized in direction cosines gives the path length of a light ray with reference to the intersection point with the perpendicular line through the origin of coordinates. Figure 20-3 illustrates this for the example of the mixed characteristic W_1 . As will be shown below, the mixed characteristic W_1 is of special importance for optical imaging systems.

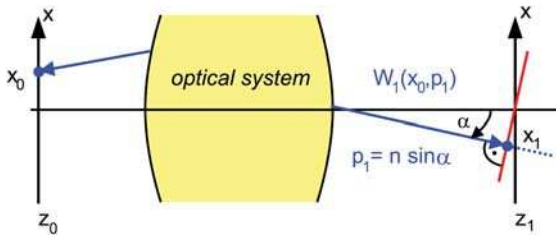


Figure 20-3: The function value of W_1 is equivalent to the light path between an object point and the perpendicular through the origin of coordinates to the light ray in image space.

The characteristic functions can, in general, be determined by ray-tracing software. The mixed characteristic for an object point, for example, is obtained by tracing a bundle of rays from the object point into image space, whereby the light path is determined up to an arbitrary plane z_1 in image space. In general, the plane z_1 may even be a virtual plane, i.e., obtained by a prolongation of the light rays in image space in the negative direction. From this plane the light rays propagate in free space and the mixed characteristic W_1 is changed according to

$$\begin{aligned} W_1(x_0, y_0; z_0; p_1, q_1; z_1 + \Delta z_1) &= W_1(x_0, y_0; z_0; p_1, q_1; z_1) + W_1^{\text{hom}}(\Delta z_1) \\ &= W_1(x_0, y_0; z_0; p_1, q_1; z_1) + \Delta z_1 \cdot \sqrt{n_1^2 - p_1^2 - q_1^2}. \end{aligned} \quad (20-16)$$

The preference of the mixed characteristic W_1 for the description of optical imaging systems becomes clear when considering the Hamilton–Jacobi equations. By differentiation of the parameters, the image coordinate x_1 , y_1 and the optical direction cosine in the object space p_0 , q_0 , respectively, are obtained:

$$\frac{\partial W_1}{\partial p_1} = -x_1, \quad \frac{\partial W_1}{\partial q_1} = -y_1, \quad (20-17a)$$

$$\frac{\partial W_1}{\partial x_0} = -p_0, \quad \frac{\partial W_1}{\partial y_0} = -q_0. \quad (20-17b)$$

The dependence of the lateral coordinates in image space from the position of the image plane z_1 , e.g., by a change of focus, can easily be obtained from eq. (20-16) by

$$x_1 = -\frac{\partial W_1^{\text{hom}}(\Delta z_1)}{\partial p_1} = \Delta z_1 \cdot \frac{p_1}{\sqrt{n^2 - p_1^2 - q_1^2}} = \Delta z_1 \cdot \tan \varphi \quad (20-18)$$

with φ as the propagation angle relative to the z_1 -axis of the ray projection into the x_1 - z_1 plane.

The mixed characteristic, however, is determined by taking the propagation angle in image space into account. The dependence on the exit pupil coordinates p_1 and q_1 at first glance seems impractical, since the exit pupil coordinates are in general not known beforehand for a light ray starting at object point x_0 and y_0 and have to be found by ray aiming. However, as will be shown below in more detail, in well-corrected optical systems there is a linear relation between exit and entrance pupil coordinates, allowing for a description in entrance pupil coordinates instead. Depending on the precision requirements the canonical exit pupil has to be scanned using an appropriate number of rays. For the description of rotational-symmetric pupils by eigenfunctions (Zernike polynomials) occasionally scanning in polar coordinates is preferred.

Beside numerical computation by ray-tracing, it is possible to derive the characteristic functions analytically. For systems comprising only spherical surfaces of radius R_i with distances d_i , for example the mixed characteristic can be written by a sum over contributions from the individual surfaces:

$$\begin{aligned} W_1(x_0, y_0; p_1, q_1) = & -z_0 m_0 - x_0 p_0 - y_0 q_0 + \sum_{i=1}^k d_i m_i \\ & + \sum_{i=1}^k \left[R_i (m_{i-1} - m_i) + R_i \text{sign}(n_i - n_{i-1}) \sqrt{(p_i - p_{i-1})^2 + (q_i - q_{i-1})^2 + (m_i - m_{i-1})^2} \right]. \end{aligned} \quad (20-19)$$

In general, it is possible to decompose the characteristic functions to the contributions from individual surfaces. Since the direction cosine components at the interfaces are in general not known, the computational effort to solve eq. (20-19) is rather complex, and becomes even more complex, if aspherical surfaces have to be considered.

20.2.3

Geometrical-optical imaging

For a geometrical-optical description of the optical imaging it is required that all light rays emitted from an object point meet in an image point at the ideal image

plane z_1 . So with eq. (20-17) the mixed characteristic of an ideal system – in addition to an arbitrary dependency on x_0 and y_0 – has to have only linear terms in p_1 and q_1 :

$$W_1(x_0, y_0; p_1, q_1) = -\beta(x_0 p_1 + y_0 q_1) + C(x_0, y_0). \quad (20-20)$$

Due to the orthotomy, all light paths between object and conjugated image points have the same length. Ultimately, a phase-synchronized superposition of the elementary wave emitted from the object point has to take place at the image point. A segment of a divergent spherical wave has to be transformed into a segment of a convergent spherical wave. This is fulfilled by the characteristic function of the ideal system with eq. (20-20). Deviations or wave-front errors W of the mixed characteristic W_1 from the ideal form eq. (20-20) with

$$W_1(x_0, y_0; p_1, q_1) = -\beta(x_0 p_1 + y_0 q_1) + C(x_0, y_0) + W(x_0, y_0; p_1, q_1) \quad (20-21)$$

lead via the Hamilton–Jacobi equations eq. (20-17) to transverse aberrations. The wave-front aberration $W(x_0, y_0; p_1, q_1)$ gives the optical path length difference or wave-front aberration for an object point x_0, y_0 in the direction cosines or canonical pupil coordinates in image space p_1, q_1 , thus also called the canonical exit pupil.

Ultimately it is the goal of the optics designer to find an optical system described by as many parameters as required to make the four-dimensional function W invariant within a specified field and aperture described by the direction cosines. An invariant wave-front error W yields a minimum transverse aberration given by:

$$\delta x_1 = -\frac{\partial W}{\partial p_1} = -\frac{\partial W_1}{\partial p_1} + \beta x_0 \ll \varepsilon. \quad (20-22)$$

In real systems, after summation over all ray angles in the image space, a spot diagram is obtained as a distribution of ray intercept points with the image plane whose center of mass generally deviates from the ideal image position; furthermore the image point is not stigmatic but faded. In general the distortion can be defined by the deviation of the center of mass from the ideal image point with:

$$\Delta x = \frac{\int_{p_1, q_1} A(p_1, q_1) \delta x_1 dp_1 dq_1}{\int_{p_1, q_1} A(p_1, q_1) dp_1 dq_1} = -\frac{\int_{p_1, q_1} A(p_1, q_1) \frac{\partial W_1}{\partial p_1} dp_1 dq_1}{\int_{p_1, q_1} A(p_1, q_1) dp_1 dq_1}. \quad (20-23)$$

Assuming $A(p_1, q_1) = 1$ for the amplitude of each ray within the required aperture, it follows that

$$\Delta x = -\frac{1}{C} \int_{\text{aperture}} \frac{\partial W_1}{\partial p_1} dp_1 dq_1. \quad (20-24)$$

The point image diameter \varnothing of the dispersion figure is determined by the second moment of the distribution:

$$\mathcal{O} = -\frac{1}{C} \int_{\text{aperture}} \left| \frac{\partial W_1}{\partial p_1} + \frac{\partial W_1}{\partial q_1} \right|^2 dp_1 dq_1 . \quad (20-25)$$

20.2.4

The Canonical Pupil

In the formal theoretical description derived above, the entrance pupil of an optical system describes for a given object, the angular spectrum passing through the aperture stop of an optical system, while the exit pupil is the corresponding angular spectrum forming the image point (figure 20-4). Since the optical direction cosines are the canonical conjugates to the space coordinates, both pupils are also called the canonical pupils of an optical system.

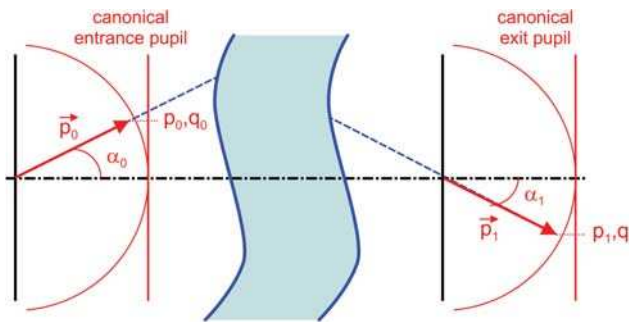


Figure 20-4: Canonical entrance and exit pupils.

The entrance pupil is the projection of the physical limits of the imaging system, limiting the angular spectrum passing the optical system, and the exit pupil is given by the projection of the physical limits in image space. The transverse components of the optical direction cosine \vec{p}_0 and \vec{p}_1 are proportional to the coordinates x_p and y_p in the aperture plane of the optical system. From eq. (20-17) and (20-20) it follows for ideal systems that the canonical coordinate of the exit pupil p_1 for the axial point $x_0=0$ is obtained from the canonical coordinate of the entrance pupil p_0 by a linear transformation:

$$p_1 = n \sin \alpha_1 = \lambda \nu_1 = \frac{x_p}{S_1} = \frac{1}{\beta} p_0 = \frac{1}{\beta} n \sin \alpha_0 . \quad (20-26)$$

Equation (20-26) is equivalent to Abbe's sine condition (see chapter 21). The canonical pupils are thus given in the angular domain and may be illustrated in the space domain for each object of the field point by spheres of arbitrary diameter. Figure 20-5 illustrates the definition of the canonical pupil coordinates according to eq. (20-26) with a positive z -component in the object and image space. Beside this definition, a diverse definition with an inversion of pupil coordinates in image space is common, which is derived from the unit planes of an aplanatic system (see also chapter 21,

section 21.3.2). The canonical pupils for a telecentric imaging system are illustrated in figure 20-5 in comparison with the unit planes for the axial object–image point pair. Pupil coordinates are frequently scaled in units of the maximum numerical aperture NA.

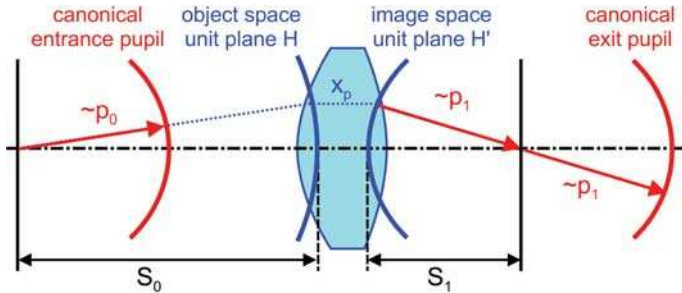


Figure 20-5: Canonical pupils in comparison with unit planes for an aplanatic system.

As a simplification, the pupil plane of an optical system is frequently described as the plane where the aperture stop is placed. The role of the aperture stop is to limit the ray bundles propagating through the optical system, and thus to limit the optical direction cosines passing the optical system from object space to image space. For identical imaging conditions, for all field points, it is essential that the aperture stop limits the ray bundles for all pairs of object–image points in a similar way. In addition to an offset of the center of gravity, given by the chief ray with optical direction cosine \vec{p}_c , the effective stop has to limit the optical direction cosines identically. Therefore, the aperture stop is typically placed at a position in the optical system, where the chief rays for all the object or image points $\vec{p}_c(x, y)$ intersect the optical axis. In the case of a telecentric imaging, the canonical pupils are centered around parallel chief rays $p_c(x, y) = 0$ and $q_c(x, y) = 0$ for all pairs of object and image points. This case is illustrated in figure 20-6 for the example of a lithographic imaging system [20-4], [20-5].

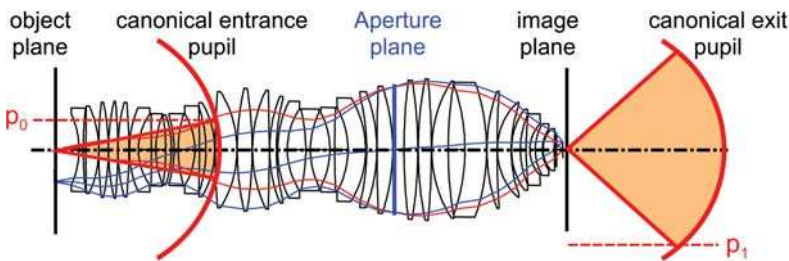


Figure 20-6: Canonical pupils for a telecentric imaging system.

Since the angular spectrum in the object or image space generally varies over the field for every object–image point pair, the canonical pupil may be different for each field point. However, for space-invariant optical imaging the canonical pupil for each image point must be identical, and the pupil planes of the image points can be

mapped onto each other. Only for telecentric imaging, are the pupils identical for all field points. For a homocentric pupil position the chief rays p_c are tilted accordingly (figure 20-7). The ideal imaging of the entrance and the exit pupil follows according to the linear imaging equation for optical direction cosines [20-1]

$$p_1 = -\frac{1}{f}x_0 + \frac{1}{\beta}p_0. \quad (20-27)$$

For an imaging system with homocentric entrance pupil and parallel chief rays, where $p_{c1} = 0$ in the image plane, the chief rays are thus given by $p_{c0} = \beta/f \cdot x_0$.

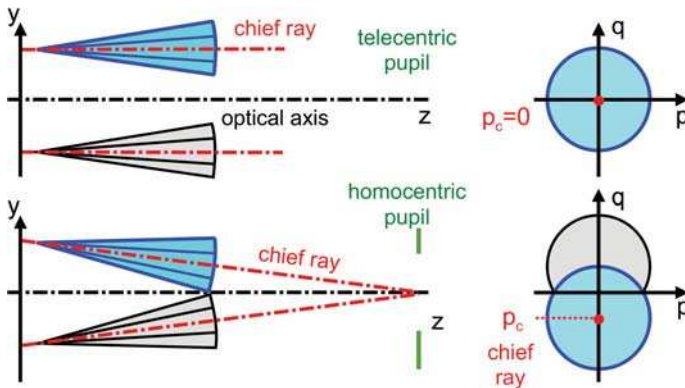


Figure 20-7: Chief rays for telecentric (top) and homocentric (bottom) pupil position for the example of the entrance pupil.

In the general case, however, there is a slight deviation from the ideal telecentric or homocentric pupil position, respectively. This has to be compensated for by suitable aiming of the chief ray. Starting with a ray bundle centered around the chief ray, the wave-optical image of an object point can be evaluated numerically, as will be shown in chapter 20.3. For the computation and scaling of the wave front in the exit pupil a mesh of rays over the entrance pupil and linear scaling according eq. (20-26) is performed for the corresponding pair of object and image point.

It can be shown that wave front aberrations according eq. (20-21) lead in general also to aberrations of the imaging of entrance pupil into exit pupil. The image error coma, for example, corresponds to a distortion of coordinates from entrance to exit pupil coordinates. Since in case of weak aberrations the wave-front errors is a slowly varying function, these effects can be neglected in typical image simulations. For exact computation or large aberrations, the linear transformation has to be replaced by an exact pupil transformation derived from eq. (20-17) [20-1].

It has to be noted that the linear relationship between entrance and exit pupil according to the sine condition in eq. (20-26) is valid only for aplanatically corrected optical systems, which is the common case for lateral imaging systems. It has however been proposed to optimize for other conditions instead the sine condition, such as Herschel or Lagrange-conditions. These conditions, eventually illustrated by non-spherical unit planes, are of less importance in optical imaging [20-6].

20.2.5

A Note on Diffractive Optical Elements

For conventional refractive optical systems it has been shown that the point characteristic V is equivalent to a conservative potential, which defines a vector field of optical direction cosines with vanishing rotation. This property is a necessary condition for image formation, since according to the physical interpretation at an ideal image point all interfering waves have to add in phase, i.e. all light paths between object and image point pairs have to be identical. For monochromatic fields, however, a superposition in phase is in general also achievable by path length differences varying by multiples of the wavelength. The monochromatic point characteristic can thus be generalized by

$$V = \int n ds \pm m\lambda . \quad (20-28)$$

In general, the rotation of the vector field is now given by a multiple of the wavelength

$$\text{rot } \vec{p} = m\lambda . \quad (20-29)$$

In consequence there exist optical elements for monochromatic applications, which allow for perfect imaging conditions although they violate the principle of Malus and Fermat. Such elements are given by diffractive optical elements with phase steps of multiple of 2π . The variation δV of the point characteristic between ideal object-image point pairs has to vanish, which is in consistence with the definition of eq. (20-28) affords equal path lengths modulo $m\lambda$ for all light rays. Equation (20-29) has to be considered very carefully, since in general it allows phase singularities, which are not consistent with the condition $\delta V=0$. It can be derived that in the plane of the diffractive optical element any closed Hilbert integral according eq. (20-11) or $\text{rot } \vec{p}$ must vanish to obtain $\delta V=0$. With the Hamilton–Jacobi-equations applied in object and image space

$$\vec{p}_0 = \vec{\nabla} V_0 \quad \text{and} \quad \vec{p}_1 = \vec{\nabla} V_1 \quad (20-30)$$

it follows for the grating vector and phase profile ϕ_{DOE} of the diffractive optical element

$$\vec{\gamma} = \vec{p}_1 - \vec{p}_0 = \vec{\nabla} V_{\text{DOE}} = \frac{\lambda}{2\pi} \vec{\nabla} \phi_{\text{DOE}} \quad (20-31)$$

Equation (20-31) is equivalent to the well known Laue-equation of grating diffraction. The local grating period is thus corresponding to the gradient of the phase difference introduced by the diffractive element. Equation (20-31) describes the implementation of diffractive elements in the optical system description by characteristic functions. For known V_0 and V_1 , eq. (20-31) further offers a convenient method for the design of diffractive optical elements. In the plane of the element the following condition must hold:

$$\frac{2\pi}{\lambda} \operatorname{rot} \vec{\gamma} = \operatorname{rot} (\vec{\nabla} \phi_{DOE}) = 0. \quad (20-32)$$

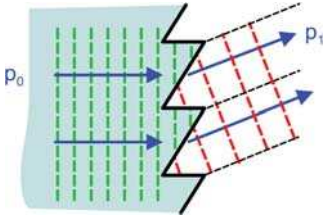


Figure 20-8: Diffractive optical element according to the definition of eq. (20-27)

Figure 20-8 illustrates the transition of the optical direction cosines at the interface of a diffractive optical element according eq. (20-31). Diffractive elements, however, suffer from several disadvantages. First, in general the phase function of a diffractive element according to eq. (20-31) depends on the angle of incidence of a light ray and thus can not be realized for large angular divergences. It has further to be considered that the description of diffractive elements given here is a simplification within the scope of geometrical optics. In general, diffractive optical elements show also higher diffraction orders with disturbing effects on the image quality.

20.3

The Ideal Wave-optical Image of a Point and Geometrical-optical Image Formation

20.3.1

The Scalar Luneburg Integral

From the principle of Malus and Fermat it follows that for the optical imaging not only all light rays emitted by one object point have to meet in one image point but also there has also to be the equality of all light paths between object and image point. This is equivalent to the demand for the phase-synchronized superposition of the elementary waves emitted from the object point in the image point. The reason for this is the wave nature of electromagnetic radiation implying the interference principle. In image points the geometrical-optical discussion fails for the calculation of the amplitude and the intensity, respectively. Usually, the calculation of the amplitude in an image point is obtained by solution of the Rayleigh-Sommerfeld diffraction integral (see paragraph 18). In a certain distance from the image plane a spherical wave is assumed, the so-called Gaussian reference sphere which converges towards the ideal image point. All rays with direction cosines \vec{p} intercept in the sphere's center. Aberrations of the wave front from the ideal spherical wave, given by deviations according to a phase delay, are given by the wave front aberration W (figure 20-9).

The approach of the Gaussian reference sphere leads, for small wave-front errors, to the correct result only [20-7]. With the Luneburg integral [20-1] a more general approach can be found, as will be shown in the following. A similar approach was given by [20-8].

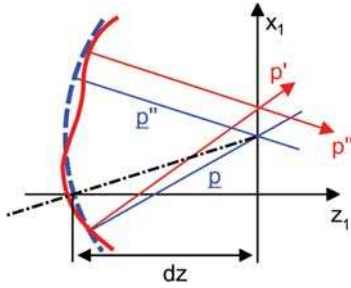


Figure 20-9: The ideal spherical wave or Gaussian reference sphere (blue) converges in one image point. Rays from the aberrated wave front with the same direction p'' or with the same ray height p' (red) do not necessarily intercept at the image point.

For the calculation of the image amplitude at the image point x_1 the Sommerfeld Ansatz is applied with the scalar field amplitude given by a harmonic function:

$$U(x_1, y_1) = A^* \cdot e^{ikV(\vec{r}_0, \vec{r}_1)} = A^* \cdot e^{ik \int_{r_0}^{r_1} \vec{p} \cdot d\vec{r}} . \quad (20-33)$$

According to Sommerfeld's Ansatz, the phase difference between two points is set equal to the point characteristic V . With this assumption the amplitude and the phase at a certain point are given for a light ray, although there might be, e.g., several solutions for the point characteristic V at the ideal image point. For this case the parameterisation using object coordinates $\vec{r}_0 = (x_0, y_0, z_0)$ and image coordinates $\vec{r}_1 = (x_1, y_1, z_1)$ is not useful. A parameterisation of eq. (20-33) is required which involves the ray angles in the image space. Such a characteristic function is given by the mixed characteristic $W_1(x_0, y_0; p_1, q_1)$. Inserting W_1 in the Sommerfeld approach, eq. (20-33), now delivers:

$$u(p_1, q_1) = A^* \cdot e^{ik[W_1(x_0, y_0; p_1, q_1) + p_1 x_1 + q_1 y_1]} . \quad (20-34)$$

In order to obtain the field amplitude at a point (x_1, y_1) we have to integrate over all direction cosines, i.e., the field amplitudes of the single 'rays' are added coherently or superimposed, respectively:

$$U(x_1, y_1) = \left(\frac{1}{\lambda}\right)^2 \int \int_{p_1, q_1} A(p_1, q_1) \cdot e^{ik[W_1(x_0, y_0; p_1, q_1) + p_1 x_1 + q_1 y_1]} dp_1 dq_1 . \quad (20-35)$$

It is quite easy to see that this step is a Fourier transformation. In general, by writing the field distribution in the 'exit pupil' as:

$$u(p_1, q_1) = A(p_1, q_1) \cdot e^{ikW_1(x_0, y_0; p_1, q_1)} \quad (20-36)$$

the scalar *Luneburg integral* is obtained as a Fourier transformation parameterised in optical direction cosines [20-1]:

$$U(x_1, y_1; z_1) = \left(\frac{1}{\lambda}\right)^2 \int \int_{p_1, q_1} u(p_1, q_1; z_1) \cdot e^{ik(p_1 x_1 + q_1 y_1)} dp_1 dq_1 . \quad (20-37)$$

After inserting $\vec{p} = \lambda \cdot \vec{\nu}$ the familiar expansion of plane waves follows:

$$U(x, y; z_1) = \int_{\nu_x, \nu_y} u(\nu_x, \nu_y; z_1) \cdot e^{i2\pi(\nu_x x + \nu_y y)} d\nu_x d\nu_y \quad (20-38)$$

where $u(\nu_x, \nu_y; z_1)$ is the transverse spectrum in the exit pupil. The *phase* of the spectrum in the image space is given by the angular characteristic, while the amplitude can be obtained – assuming radiance in the object space – by the radiance in the image space. The radiance in the object space can either be isotropic for an ideal point source, or described by the diffraction spectrum of the object pattern. In addition to this, the aperture limitation can be taken into account by a *pupil function* $P(p, q)$ that can also contain reflection and absorption losses. With this, eq. (20-35) can be converted into:

$$U(x, y; z_1) = \left(\frac{1}{\lambda}\right)^2 \iint_{p, q} P(p, q) A(p, q) \cdot e^{ik_0 W_1(x_0, y_0; p, q)} \cdot e^{ik_0(px+qy)} dpdq . \quad (20-39)$$

Neglecting absorption and reflection losses the ideal pupil function is given by:

$$P(p, q) = \begin{cases} 1 & \text{for } p^2 + q^2 < \text{NA} \\ 0 & \text{else .} \end{cases} \quad (20-40)$$

So far, no imaging system was required and eqs (20-35)–(20-40) are applicable also to non-imaging systems. For imaging systems, eq. (20-21) can be inserted for W_1 (neglecting the constant phase function $C(x_0, y_0)$):

$$U(x_1, y_1) = \left(\frac{1}{\lambda}\right)^2 \iint_{p_1, q_1} A(p_1, q_1) \cdot P(p_1, q_1) e^{ik[W(x_0, y_0; p_1, q_1) + p_1(x_1 - \beta x_0) + q_1(y_1 - \beta y_0)]} dp_1 dq_1 . \quad (20-41)$$

Equation (20-41) contains the change of coordinates from the object to the image coordinates under consideration of the magnification and the wave front aberration is considered for the calculation of the amplitude in image space.

Equations like eq. (20-38) and eq. (20-41) are equivalent to the so-called Debye approximation: outside a certain angular range limited by the numerical aperture the angular spectrum of the field is set equal to zero, inside it is set equal to the undisturbed wave. Furthermore, evanescent waves are neglected. Since one is usually at a large distance from diffracting apertures compared with the wavelength λ , the latter approximations do not involve any limitations for the calculation of image formation.

At no step of the development so far has an image point been necessary. So eq. (20-39) can be applied even when no imaging optical system, or when only a part of it, respectively, is discussed. Furthermore, the choice of the reference plane is arbitrary since by taking the free-space propagation into account in eq. (20-38) the amplitude in the image space can be calculated in any plane with

$$U(x, y; z_1) = \int \int_{v_x, v_y} u(v_x, v_y; 0) \cdot e^{i2\pi(v_x x + v_y y)} e^{i2\pi v_z z_1} dv_x dv_y. \quad (20-42)$$

All integrals (20-37)–(20-42) are equivalent to the integral over the canonical exit pupil (20-35). Occasionally it is preferable to use the solid angle integral over the Gaussian reference sphere [20-9]. Both descriptions can, of course, be converted into each other. After converting into spherical coordinates with $v_x = n/\lambda \cdot \cos\vartheta \cdot \sin\alpha$, $v_y = n/\lambda \cdot \sin\vartheta \cdot \sin\alpha$, $v_z = n/\lambda \cdot \cos\alpha$, and

$$\left| \frac{\partial(v_x, v_y)}{\partial(\alpha, \vartheta)} \right| = \left(\frac{n}{\lambda} \right)^2 \cos\alpha \sin\alpha \quad (20-43)$$

eventually, the surface integral over a sphere segment is obtained as

$$U(x, y, z) = \left(\frac{n}{\lambda} \right)^2 \int_0^{2\pi} \int_0^{\arcsin(NA)} u_\alpha(\alpha, \vartheta) \cdot e^{ik_0((x \cos\vartheta + y \sin\vartheta) \sin\alpha + z \cos\alpha)} \cos\alpha \sin\alpha \, d\alpha d\vartheta. \quad (20-44)$$

The integral ϑ in eq. (20-44) is equivalent to the integral over the solid angle $d\Omega = \sin\alpha d\alpha d\vartheta$:

$$U(x, y, z) = \left(\frac{n}{\lambda} \right)^2 \int_0^{2\pi} \int_0^{\arcsin(NA)} a(\alpha, \vartheta) \cdot e^{ik_0((x \cos\vartheta + y \sin\vartheta) \sin\alpha + z \cos\alpha)} d\Omega \quad (20-45)$$

with

$$u_\alpha(\alpha, \vartheta) = \frac{a(\alpha, \vartheta)}{\cos\alpha} = \frac{n}{\lambda} \frac{a(\alpha, \vartheta)}{v_z}. \quad (20-46)$$

One characteristic of a spherical wave is that the amplitude is constant over all angles. By inserting a spherical wave with $a(\alpha, \vartheta) = \lambda/n = \text{const.}$ we obtain

$$U(x, y, z) = \int \int_{v_x^2 + v_y^2 < \frac{NA^2}{\lambda^2}} \frac{1}{v_z} \cdot e^{i2\pi v_z z} \cdot e^{i2\pi(v_x x + v_y y)} dv_x dv_y \quad (20-47)$$

i.e., the illumination of the exit pupil has to increase towards the edge of the pupil with $1/\cos\alpha$ for a spherical wave. Equation (20-47) is equivalent to the Weyl expansion of the spherical wave into plane waves [20-10].

Both descriptions, namely the integral parameterised to solid angles and the integral over the canonical exit pupil, are identical. Which description should therefore be selected for a discussion of image formation? According to the linear system theory either the object has to be convoluted with the point response function or the frequency spectrum of the object has to be multiplied by a transfer function. For these operations the linearity must be fulfilled. For the object and image coordinates a linear correlation $x_1 = \beta x_0$ is always valid, although the (ideal) optical imaging is

only linear in the optical direction cosines but not in the angle itself, since for $x_0 = 0$, from eq. (20-26) it follows that:

$$p_1 = p_0 / \beta \quad \text{or} \quad v_1 = v_0 / \beta . \quad (20-48)$$

So the most effective way is to choose the simple description in frequency space or in optical direction cosines and canonical pupil coordinates, respectively. In general, though, a calculation of the Luneburg integral by using, for example, a Fast Fourier Transformation (FFT) in Cartesian coordinates is preferable.

If evanescent waves can be neglected, the scalar Luneburg integral is an exact treatment of optical imaging for scalar fields. However, electromagnetic waves are transverse vector waves and thus any scalar treatment is restricted to small aperture angles and to cases where polarization effects can be neglected. With increasing aperture angles, scalar treatment becomes more and more incorrect and vector theory has to be applied instead. In general, three ranges of the description of optical imaging can be defined as follows.

1. Scalar treatment is sufficient for small aperture angles.
2. In first order, the vector effects can be taken into account by scalar factors for the pupil function for apertures in the medium range and for unpolarized imaging.
3. In the case of very large numerical apertures, it is necessary to apply the exact vector treatment for optical imaging (see chapters 26–28).

The limits between the three ranges of the numerical aperture are a matter of the required accuracy. Frequently, the values 25° ($NA \approx 0.4$) and 40° ($NA \approx 0.6$), respectively, are used as limits. In optical lithography, for example, the detector medium is a photo-resist of refractive index $n \approx 1.7$, and vector effects can be neglected up to high numerical apertures.

Common to both the scalar and vector treatment of optical imaging, are photometric factors which are considered to conserve energy. Energy conservation is discussed in the next chapter.

20.3.2

Energy Discussions for Optical Imaging

Consider an area of size A_0 in the object plane, which is illuminated by a plane wave under the angle α_0 (figure 20-10). The radiant flux Φ (power) through the area A_0 is given by the scalar product of the Poynting vector \vec{S} and the surface normal vector:

$$\begin{aligned} \Phi &= \vec{S}_0 \cdot \vec{A}_0 = A_0 \cdot |u_0|^2 \cdot \cos \alpha_0 = \vec{S}_1 \cdot \vec{A}_1 = A_1 \cdot |u_1|^2 \cdot \cos \alpha_1 \\ &= \beta^2 A_0 \cdot |u_1|^2 \cdot \cos \alpha_1 . \end{aligned} \quad (20-49)$$

The projection of the propagation vector on the surface normal decreases with increasing angle α and for the energy flow through the area A_0 and A_1 , the cosine factor has to be taken into account.

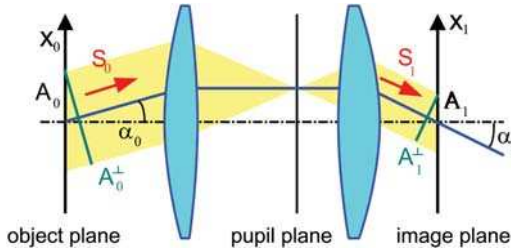


Figure 20-10: Energy conservation for optical imaging.

Equation (20-49) can be derived, in an analogous way, from the continuity equation for the irradiance (17-104):

$$\vec{\nabla} \cdot (I \cdot \vec{p}) = \oint_A I \cdot \vec{p} \cdot d\vec{a} = I \cdot A_1 \cdot n_1 \cos \alpha_1 - I \cdot A_0 \cdot n_0 \cos \alpha_0 = 0. \quad (20-50)$$

Any closed surface vector integral over the irradiance times the direction cosine must vanish.

Equations (20-49) and (20-50) illustrate a difficulty quite common in optics: angular information is required in the space domain. The radiant flux of a planar wave is proportional to the z -component of the wave vector. So to calculate the radiant flux in the object or image space, information about the propagation angles is necessary. However, they are generally not uniquely given since there is a superposition of many plane waves with different propagation directions. For the calculation of the intensity in the image plane, the field has to be decomposed into its spatial frequencies. In the frequency domain, the cosine factor can be taken into account and for the flux Φ we finally obtain:

$$\begin{aligned} \Phi(x_1, y_1) &= \int_{p, q < NA} \vec{S}_1 \cdot \vec{A}_1 dp_1 dq_1 \\ &\sim \left| \int_{p, q < NA} u_1(p_1, q_1) \cdot e^{ik_0(x_1 p_1 + y_1 q_1)} \cdot \sqrt{n^2 - p_1^2 - q_1^2} \cdot dp_1 dq_1 \right|^2. \end{aligned} \quad (20-51)$$

Of course, this is not very practical, and therefore it is advantageous to consider the cosine factor by using an additional factor directly in the pupil function of the optical system. Then the intensity can be calculated simply by forming the value square [20-11], [20-12], [20-13]. After insertion in eq. (20-41) and consideration of the $\sqrt{\cos}$ -factors in the object and image space, we obtain:

$$U(x_1, y_1) = \left(\frac{1}{\lambda}\right)^2 \int_{p_1, q_1} A(p_1, q_1) \cdot P(p_1, q_1) e^{ik[W(x_0, y_0; p_1, q_1) + p_1(x_1 - \beta x_0) + q_1(y_1 - \beta y_0)]} \cdot \frac{\sqrt{n_1 \cos \alpha_1}}{\sqrt{n_0 \cos \alpha_0}} \cdot dp_1 dq_1 \quad (20-52)$$

with

$$n_i \cos \alpha_i = \sqrt{n_i^2 - p_i^2 - q_i^2} \quad (20-53)$$

The energy factor thus depends on the magnification and cancels for 1:1 imaging.

20.3.3

The Airy Disc

An ideal point image is the aberration-free image of an object point with Lambertian emission characteristic illuminating the pupil homogeneously. In this case, in eq. (20-41), the amplitude $A(p, q)$ and the pupil function $P(p, q)$ within the circular aperture can be set as constant and equal to one. Choosing the coordinate of the geometrical-optical image point as the origin of the coordinate system, one obtains after converting eq. (20-41) into polar coordinates and scaling to the maximum aperture:

$$U(r, \theta) = \left(\frac{1}{\lambda}\right)^2 e^{-ik(p_c x + q_c y)} \int_0^1 \int_0^{2\pi} e^{ik_0 \cdot NA \cdot r \cdot \rho' \cos(\theta - \varphi)} \rho' d\rho' d\varphi \quad (20-54)$$

In eq. (20-54) it is taken into account that, for non-telecentric imaging, the center point of a pupil to a off-axis field point is decentered by the chief ray angle p_c, q_c (figure 20-11). The chief ray with p_c and q_c leads only to an additional phase term which vanishes during the intensity calculation.

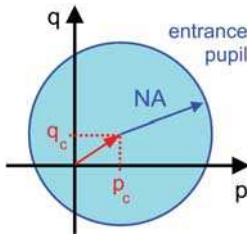


Figure 20-11: Pupil with chief ray direction p_c, q_c for an off-axis field point in a non-telecentric system.

For rotational-symmetric systems, the ideal amplitude distribution function is obtained from eq. (20-54). With the coordinates normalized to the maximum aperture angle described by the numerical aperture NA and neglecting the normalization constants and constant phase terms, we obtain:

$$U(r) = \int_0^1 \int_0^{2\pi} e^{i2\pi r \rho \cos(\theta - \varphi)} \rho d\rho d\varphi = 2\pi \int_0^1 J_0(2\pi r \rho) \rho d\rho = \frac{NA}{r \cdot \lambda} J_1\left(2\pi r \cdot \frac{NA}{\lambda}\right) \quad (20-55)$$

with the Bessel function $J_1(x)$. The intensity formation and normalization to the maximum value 1 gives the *Airy disc* [20-14]:

$$I_{\text{PSF}}(r) = \left(\frac{J_1(2\pi w)}{\pi w} \right)^2 \tag{20-56}$$

with normalized coordinates $w = NA/\lambda \cdot r$. The ideal point image intensity is called the point-spread function PSF (see chapter 21, *Imaging with Partial Coherence*). Figure 20-12 shows the normalized amplitude $U(w)$ and the intensity distribution $I(w)$ of the Airy disc with the first two zeros and secondary maxima. The radii of the first minima and the encircled energies are compiled in table 20-1, with the encircled energy defined by

$$\text{enc}E(r) = 2\pi \int_0^r I(r) r dr . \tag{20-57}$$

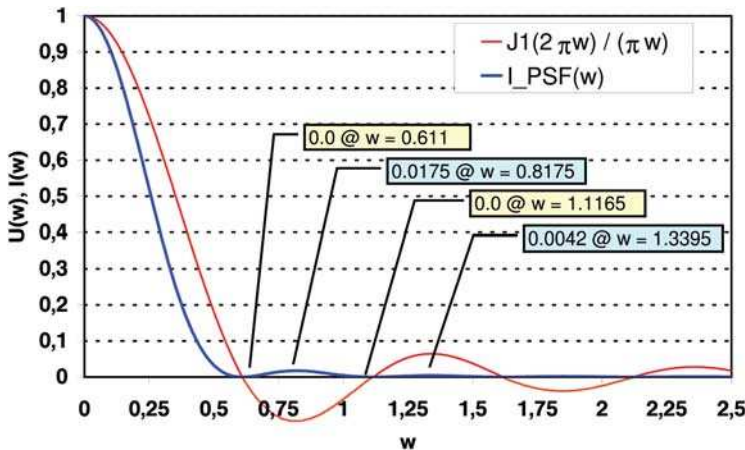


Figure 20-12: Cross-section through the Airy disc.

Table 20-1: Radii and encircled energy of first rings of the Airy intensity pattern

	Central spot	First ring	Second ring
r_n	$r_0 = 0.611 \lambda/NA$	$r_1 = 1.1165 \lambda/NA$	$r_2 = 1.62 \lambda/NA$
Encircled energy	83.8%	7.2%	2.8%

In a perfect optical imaging system, each object point $\delta(x-x_0)$ is imaged at an ideal image point blurred to an Airy disc, because of the finite wavelength λ of light and the limited aperture angle of the converging spherical wave, due to cutting back at the aperture.

The polychromatic point-spread function is generated by integration of the monochromatic point-spread function according to the Airy formula (eq. (22-56)), weighted by the power spectrum of the light source:

$$I_{\text{poly}}(r) = \int S(\lambda) \cdot \left[\frac{J_1(2\pi \cdot w)}{\pi \cdot w} \right]^2 d\lambda = \int S(\lambda) \cdot \left[\frac{J_1\left(2\pi \frac{NA}{\lambda} \cdot r\right)}{\pi \frac{NA}{\lambda} \cdot r} \right]^2 d\lambda . \quad (20-58)$$

Figure 20-13 shows polychromatic PSF for band-limited homogeneous spectra. With increasing relative spectral bandwidth $\Delta\lambda/\lambda$, the zeros of the Airy diffraction pattern are smoothed by the integration and finally disappear. For the visible part of the spectrum, the relative spectral bandwidth at which the zero disappears, is approximately at $\Delta\lambda/\lambda=0.5$. Here the first diffraction minimum drops to a value of approximately 1% of the peak value.

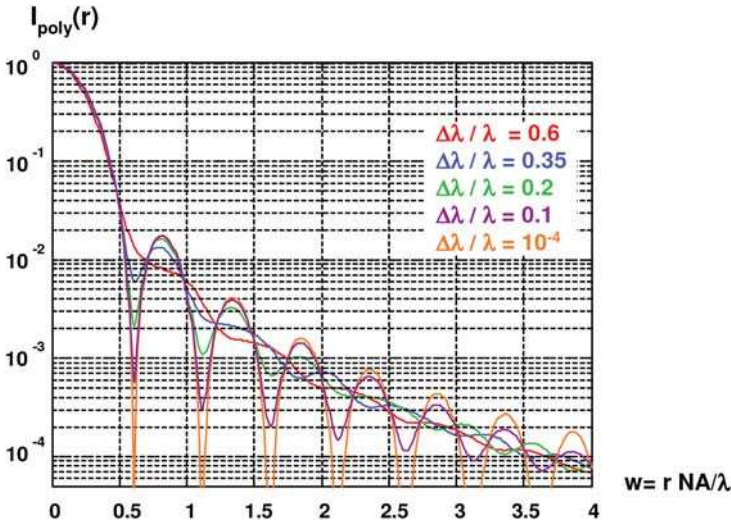


Figure 20-13: Transverse intensity distribution of the polychromatic point-spread function for different sizes of the spectral interval.

Particularly for imaging with high numerical aperture and large magnification, the amplitude distribution in the exit pupil may be non-homogeneous due to the energetic apodization factors of eq. (20-52). Figure 20-14 illustrates the effect of the apodization factor for a collimating optical system ($\beta \approx 0$).

As shown in more detail below, the point-spread function limits the resolution of imaging. The Airy disc according eq. (20-56) is valid only for conventional systems with a circular pupil, while cylindrical systems or rectangular apertures, for example, show different point-spread functions. Examples of point-spread functions for annular pupils, pupil filters and apodization filters are given in chapter 25. The three-dimensional shape of the point-spread function is discussed in chapter 23 under the generalized three-dimensional aperture.

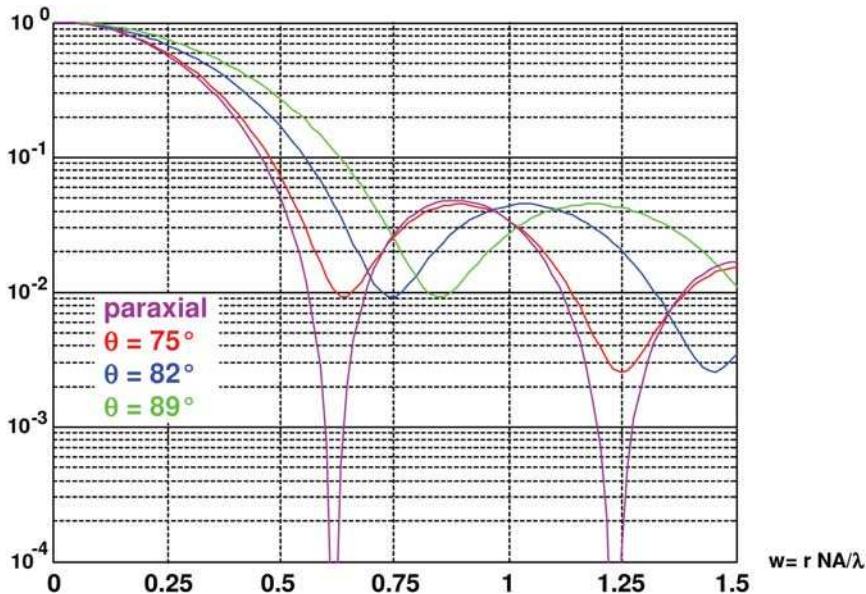


Figure 20-14: Comparison of the impact of energetic apodization factors for different apertures (for small $\beta \approx 0$).

According to a simplified geometrical-optical interpretation of optical imaging, each object point is imaged into an image point. The geometrical image point of a perfect optical system is stigmatic, and the wave-optical description adds an additional wave-optical blurring to each image point. In the presence of aberrations, the geometrical image point is additionally blurred by ray aberrations according to eq. (20-21), and, as will be shown in paragraph 20.4, the corresponding wave-optical point image is determined in accordance with the wave-front aberration W . In first order the image intensity $I_1(x)$ in the image plane is thus given by of the intensity distribution $I_0(x)$ in the object plane blurred by the point-spread function. Consequently, the image intensity is given by a linear superposition of single point-spread functions according to the single object points of intensity $I_0(x)$ in the object plane, or, respectively, by a convolution of the object intensity $I_0(x)$ with the point-spread function. Optical imaging may thus be considered as a linear system for linear transformation of the intensity. As will be shown in chapter 21, this approximation is only valid for the case of perfectly incoherent optical imaging, when the wave-optical image points of neighbouring object points do not interfere. Incoherent imaging approximation is always applicable when an effective light source is used for an illumination whose image fills the pupil of the imaging optical system completely, or when self-luminous objects like stars are discussed.

20.3.4

Incoherent Resolution

As we have shown above, for the case of rotational-symmetric imaging with a circular pupil according to eq. (20-56), the ideal image of an object point is given by the Airy disc. Considering the incoherent imaging of two points, the image cannot be distinguished from the imaging of a line element until the points are at a certain distance. The resolution limit defined in this way is determined by the required contrast of the detector – and sometimes from the fantasy or preliminary information of the observer. Figure 20-15 shows the cross-section through the intensity of the image of two point-like objects with increasing normalized distance dw .

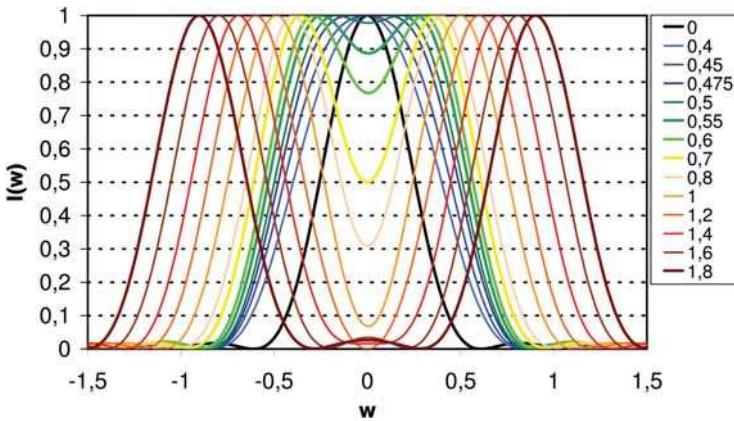


Figure 20-15: Superposition of two point images (normalized representation) at distances dw .

The point images of single points cannot be separated at distances $dw = \lambda / NA \cdot dr < 0.47$, since only for $dw > 0.47$ does a minimum form between the two point images. The borderline case with $dw = 0.47$ is also called the Sparrow criterion of resolution. At this distance of two points, the second derivative of the intensity distribution vanishes at the maximum of the intensity distribution. For larger separations of two image points, a minimum I_{\min} is formed between the two image points resulting in contrast or visibility $V = (I_{\max} - I_{\min}) / (I_{\max} + I_{\min})$. In figure 20-16 the visibility of the double point image is plotted versus the point distance. The possible maximum visibility of $V = 100\%$ is reached when the first minima (1. zero) of the diffraction figures cover each other according to a distance $dw \sim 1.2$. According to the Rayleigh resolution limit, two points can be separated from each other if the maximum of one point image coincides with the minimum of the second point at $\delta x = 0.611 \cdot \lambda / NA$. This is equivalent to a visibility of $V \approx 15\%$, when the minimum between the two maxima then has a height of 74% of the maximum intensity. Assuming a visibility requirement of 80% only points at a distance of at least $dw > 0.95$ can be distinguished from each other [20-15], [20-16], [20-17].

For the observation of the starry sky, the spatial resolution delivers quite unwieldy numbers. That is why the angular resolution is preferred. The telescope resolution

is given in sexagesimal minutes with $\theta = r/z = \lambda/2D$ with D the diameter of the telescope. For the first minimum one then gets $\theta_0 = 1.22 \cdot \lambda/D$. For younger persons the accuracy of vision with the naked eye is about $10 \mu\text{m}$ corresponding to about one sexagesimal minute. The standard diameter of the eye is 18 mm . According to this a young person should be able to distinguish two points at a distance of 20 m that are 6 mm apart.

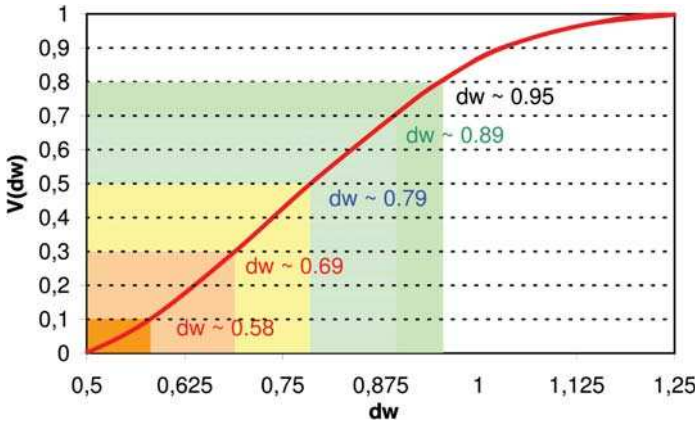


Figure 20-16: Visibility V versus point image distance dw .

20.4
Aberrations of Optical Systems

20.4.1
The Small-aberration Limit: The Strehl Ratio

In the case of wave-front aberrations, the image amplitude is given by eq. (20-41). In image coordinates, we obtain

$$U(x, y) = \left(\frac{1}{\lambda}\right)^2 \int_{p_1, q_1} A(p, q) \cdot P(p, q) e^{ik[W(x_0, y_0, p, q) + px + qy]} dpdq. \tag{20-59}$$

For small wave-front errors the phase term of eq. (20-59) can be expanded with

$$e^{i\Phi} = 1 + i\Phi - \frac{1}{2}\Phi^2 + \dots + \frac{1}{n!}(i\Phi)^n + \dots \tag{20-60}$$

Thus the image intensity of an object point can be written as:

$$I(x, y) = \left(\frac{1}{\lambda^2}\right) \left| \int_{p^2+q^2 < NA^2} \left(1 + ik_0 W - \frac{1}{2}k_0^2 W^2\right) \cdot e^{ik_0(px+qy)} dpdq \right|^2 \tag{20-61}$$

with $A(p,q) = 1$ and $P(p,q)$ given according (20-40). The normalized intensity for the position of an ideal image point on the axis is thus reduced by the square of the standard deviation ΔW of the wave-front error with

$$I_n(0) \approx \left| 1 + ik_0 \overline{W} - \frac{1}{2} k_0^2 \overline{W^2} \right|^2 \approx 1 + k_0^2 \overline{W^2} - k_0^2 \overline{W}^2 = 1 - k_0^2 (\Delta W)^2 . \quad (20-62)$$

The intensity at the center of mass of the point image is called Strehl's definition. According to eq. (20-62) the deviation of Strehl's definition from 1 is given to a good approximation by the root-mean-square deviation of the wave front $\Delta W = \text{RMS}$. The decrease of Strehl's definition is equal to the *contrast loss* of the aberrated optical image. So the contrast loss is proportional to the square of the RMS value of the wave-front error.

From eq. (20-62) an estimation for the maximum allowed wave-front error can be derived. Allowing a decrease of Strehl's definition to about 80% of the ideal radiance the *Maréchal criterion* for the maximum allowable wave-front error is obtained

$$\text{RMS}(W) = \Delta W < \frac{\lambda}{14} . \quad (20-63)$$

Particularly for modern high-resolution optical systems with deviations of the definition of less than 10^{-3} Strehl's definition and the Maréchal criterion are not sensible criteria and thus are only useful in a limited way.

20.4.2

Expansion of the Wave-front Error into Zernike Polynomials

The mixed characteristic W_1 contains a description of the wave-front error using the optical direction cosines. Usually for the description of an optical system the wave-front error and the spatial distribution of the wave-front error is expanded into polynomials over the image field. From this the known power series expansion of the characteristic polynomial follows [20-1]:

$$W_1(x_0, y_0; p_1, q_1) = \sum_{n,m,k,l} A_{nmkl} x_0^n y_0^m p_1^k q_1^l . \quad (20-64)$$

The polynomial expansion is frequently ordered according to the sum of the powers $O = n+m+k+l$, whereby the design aberrations are represented by the even powers, and the odd powers are contributions due to fabrication errors such as decentering of optical elements. For isotropic imaging properties the pupil assigned to each field point has to be identical. For real imaging systems with eq. (20-64) only in small areas is a constancy of the pupil given approximately, so that only in small areas can the linear approach of the image formation, with a spatially invariant point response function, be applied.

For the description of rotational-symmetric systems, an orthogonal expansion in polar coordinates is preferred to the expansion in Cartesian coordinates. One set of orthogonal functions for a circular aperture are the Zernike polynomials. If the image field is limited to circular area the wave-front error can be described by a double power series expansion in polar coordinates [20-18], [20-19], [20-20], resulting in a system of the aberrations over the image field which is important for the correction and adjustment of an optical system. In the following the discussion is limited to a certain image field with locally constant wave-front errors. This condition is also called the isoplanatic condition for optical systems.

From eq. (20-59) one obtains in coordinates normalized to the maximum aperture:

$$U(r, \theta, z) = \left(\frac{NA}{\lambda} \right) e^{-ik(p_c \cdot x + q_c \cdot y)} \int_0^1 \int_0^{2\pi} A(\rho, \varphi) \cdot e^{ik_0 W} \cdot e^{ik_0 NA \cdot r \cdot \rho \cos(\theta - \varphi)} \rho d\rho d\varphi. \quad (20-65)$$

Again the chief ray angle p_c , q_c leads only to an additional phase term which vanishes during the intensity formation. However, it has to be noticed that the chief ray has to be chosen properly since its selection as center of expansion influences the series expansion coefficients of the wave front aberrations, i.e. a wrong chief ray aiming at the calculation leads to different wave-front errors.

According to Zernike the wave-front error for an image point can be expanded into a set of circular orthogonal polynomials:

$$\begin{aligned} W(x_0; \rho, \varphi) &= \sum_{n=0}^{\infty} \sum_{m=0}^n C_{n,m}(x_0) R_n^m(\rho) e^{im\varphi} \\ &= \sum_{n=0}^{\infty} \sum_{m=0}^n A_{n,m}(x_0) R_n^m(\rho) \cos(m\varphi) + \sum_{n=0}^{\infty} \sum_{m=0}^n A'_{n,m}(x_0) R_n^m(\rho) \sin(m\varphi). \end{aligned} \quad (20-66)$$

According to common conventions the Zernike aberrations are usually ordered by a single index j :

$$W(x_0; \rho, \varphi) = \sum_{j=0}^{\infty} c_j(x_0) Z_j(\rho, \varphi). \quad (20-67)$$

The radial polynomial R of maximum radial power order n is given by

$$R_n^m(\rho) = \sum_{s=0}^{\frac{n-m}{2}} (-1)^s \frac{(n-s)!}{s! \left(\frac{n+m}{2} - s \right)! \left(\frac{n-m}{2} - s \right)!} \rho^{n-2s}. \quad (20-68)$$

The radial polynomial always has the value 1 at the pupil edge. The norm of the radial polynomials is given by

$$\int_0^1 R_n^m R_l^m \rho d\rho = \frac{1}{2(n+1)} \delta_{nl} . \quad (20-69)$$

With this the Zernike polynomials are normalized over

$$\int_0^{2\pi} \cos(kx) \cos(mx) dx = \begin{cases} \pi \delta_{km} & m = k \geq 1, \\ 2\pi & m = k = 0. \end{cases} \quad (20-70)$$

It can be shown that the average integral over radial polynomials vanishes for $n > 0$, thus the averages over Zernike polynomials with $m > 0$ and $n > 0$ vanish:

$$\int_0^1 \int_0^{2\pi} Z_j \rho d\rho d\varphi = \int_0^1 R_n^m(\rho) \rho d\rho \int_0^{2\pi} \cos(m\varphi) d\varphi = \delta_{m0} \delta_{n0} . \quad (20-71)$$

The term with $n=m=0$ is equivalent to a constant wave-front error which can be neglected since a constant phase is insignificant – only phase differences play a role. For the square of the RMS wave-front error, we have:

$$\begin{aligned} \Delta W^2 &= \frac{1}{2\pi} \int_0^1 \int_0^{2\pi} [W(x_0; \rho, \varphi) - \overline{W}]^2 \rho d\rho d\varphi \\ &= \sum_{n=1}^{\infty} \frac{A_{n,0}^2}{n+1} + \frac{1}{2} \sum_{n=1}^{\infty} \sum_{m=1}^n \frac{A_{n,m}^2 + A_{n,m}'^2}{n+1} . \end{aligned} \quad (20-72)$$

The most important Zernike polynomials are compiled in table 20-2 and in figure 20-17. It should be noted that other normalizations are used beside the normalization of the Zernike polynomial Z_j to a maximum value of one at the pupil edge. According to Noll, the Zernike polynomials are normalized according to a RMS definition. The difference with respect to the definition above lies on the normalization factor $\sqrt{n+1}$ for the rotationally symmetric terms and $\sqrt{2(n+1)}$ for the non-rotationally symmetric terms.

The mathematical appendix contains a discussion of orthogonal expansions for rectangular regions in section A.8.

Table 20-2: The Fringe-Zernike aberrations from $j=1$ to $j=36$

j	n	m	Polynomial term	Image aberrations
1	0	0	1	Constant phase
2	1	1	$\rho \cos\varphi$	Distortion in x
3	1	1	$\rho \sin\varphi$	Distortion in y
4	2	0	$2\rho^2 - 1$	Defocus (parabolic part)
5	2	2	$\rho^2 \cos(2\varphi)$	Astigmatism 3rd order
6	2	2	$\rho^2 \sin(2\varphi)$	Astigmatism 3rd order
7	3	1	$(3\rho^3 - 2\rho) \cos\varphi$	Coma 3rd order
8	3	1	$(3\rho^3 - 2\rho) \sin\varphi$	Coma 3rd order
9	4	0	$6\rho^4 - 6\rho^2 + 1$	Spherical aberration 3rd order
10	3	3	$\rho^3 \cos(3\varphi)$	Trifoil
11	3	3	$\rho^3 \sin(3\varphi)$	Trifoil
12	4	2	$(4\rho^4 - 3\rho^2) \cos(2\varphi)$	Astigmatism 5th order
13	4	2	$(4\rho^4 - 3\rho^2) \sin(2\varphi)$	Astigmatism 5th order
14	5	1	$(10\rho^5 - 12\rho^3 + 3\rho) \cos\varphi$	Coma 5th order
15	5	1	$(10\rho^5 - 12\rho^3 + 3\rho) \sin\varphi$	Coma 5th order
16	6	0	$20\rho^6 - 30\rho^4 + 12\rho^2 - 1$	Spherical aberration 5th order
17	4	4	$\rho^4 \cos(4\varphi)$	Four wave
18	4	4	$\rho^4 \sin(4\varphi)$	Four wave
19	5	3	$(5\rho^5 - 4\rho^3) \cos(3\varphi)$	Three wave 7th order
20	5	3	$(5\rho^5 - 4\rho^3) \sin(3\varphi)$	Three wave 7th order
21	6	2	$(15\rho^6 - 20\rho^4 + 6\rho^2) \cos(2\varphi)$	Astigmatism 7th order
22	6	2	$(15\rho^6 - 20\rho^4 + 6\rho^2) \sin(2\varphi)$	Astigmatism 7th order
23	7	1	$(35\rho^7 - 60\rho^5 + 30\rho^3 - 4\rho) \cos\varphi$	Coma 7th order
24	7	1	$(35\rho^7 - 60\rho^5 + 30\rho^3 - 4\rho) \sin\varphi$	Coma 7th order
25	8	0	$70\rho^8 - 140\rho^6 + 90\rho^4 - 20\rho^2 + 1$	Spherical aberration 7th order
26	5	5	$\rho^5 \cos(5\varphi)$	Five wave
27	5	5	$\rho^5 \sin(5\varphi)$	Five wave
28	6	4	$(6\rho^6 - 5\rho^4) \cos(4\varphi)$	Four wave 9th order
29	6	4	$(6\rho^6 - 5\rho^4) \sin(4\varphi)$	Four wave 9th order
30	7	3	$(21\rho^7 - 30\rho^5 + 10\rho^3) \cos(3\varphi)$	Three wave 9th order
31	7	3	$(21\rho^7 - 30\rho^5 + 10\rho^3) \sin(3\varphi)$	Three wave 9th order
32	8	2	$(56\rho^8 - 105\rho^6 + 60\rho^4 - 10\rho) \cos(2\varphi)$	Astigmatism 9th order
33	8	2	$(56\rho^8 - 105\rho^6 + 60\rho^4 - 10\rho) \sin(2\varphi)$	Astigmatism 9th order
34	9	1	$(126\rho^9 - 280\rho^7 + 210\rho^5 - 60\rho^3 + 5\rho) \cos\varphi$	Coma 9th order
35	9	1	$(126\rho^9 - 280\rho^7 + 210\rho^5 - 60\rho^3 + 5\rho) \sin\varphi$	Coma 9th order
36	10	0	$252\rho^{10} - 630\rho^8 + 560\rho^6 - 210\rho^4 + 30\rho^2 - 1$	Spherical aberration 9th order

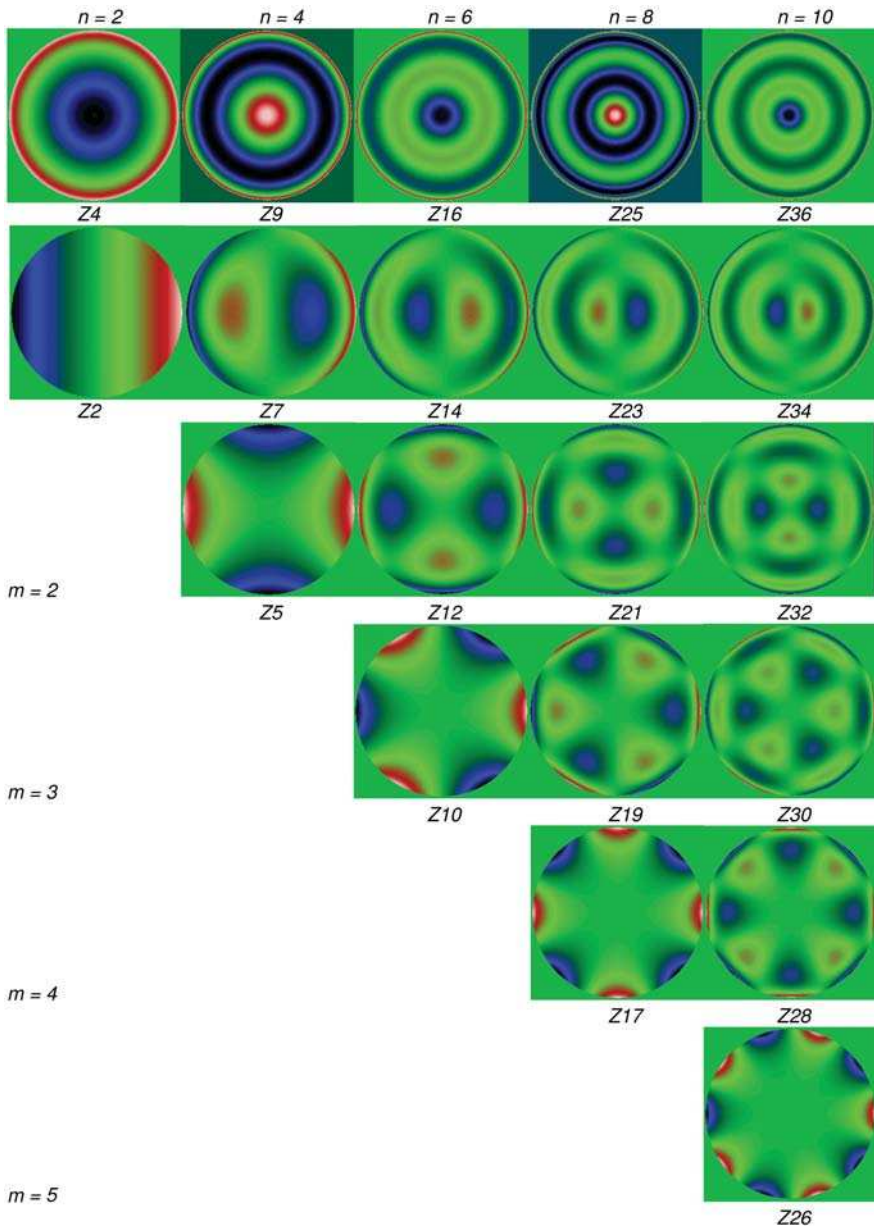


Figure 20-17: Comparison of Zernike wave-front errors over the power of the radial polynomial order n and azimuthal order m ; map across the pupil with a color coding of the amplitude. In each line the powers n are reduced by $2m$.

20.4.3

Point Images for Different Aberrations

The following figures 20-18 – 20-23 illustrate point images for the wave-front aberrations of lowest order. A detailed analysis of selected aberrations is given in 20.4.4– 20.4.6. Figure 20-18 illustrates the effect of defocus Z_4 on the lateral intensity distribution.

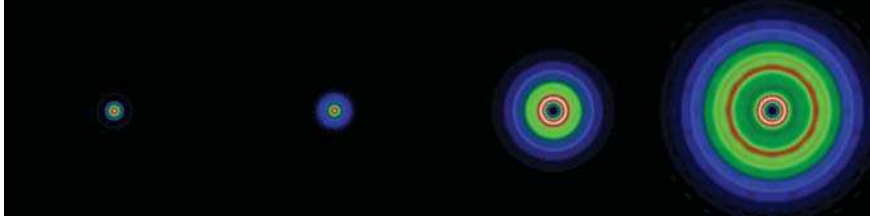


Figure 20-18: Point images at $Z_4 = 0, \lambda/5, \lambda/2$ and λ .

After defocus, spherical aberration Z_9 is the next higher aberration of rotational symmetry. The intensity distribution is consequently rotationally symmetric. Figure 20-19 compares the intensity distributions in the focal plane for different amounts of Z_9 , figure 20-20 compares the intensity distributions in the focal plane for different spherical aberrations of higher order. In contrast to spherical aberrations of lower order, the central intensity peak vanishes for higher spherical aberration Z_{36} .

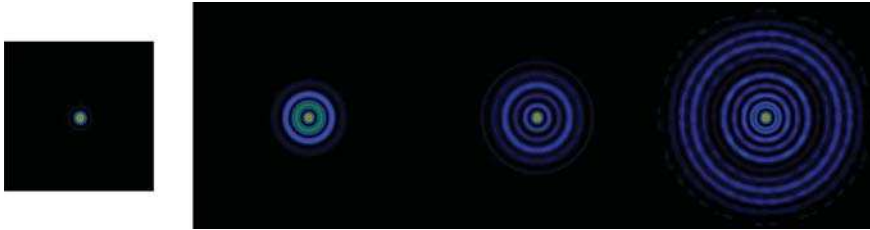


Figure 20-19: Point images in the ideal case and with spherical aberration $Z_9 = 0, \lambda/2, \lambda$ and 2λ (figure scale changed in comparison with figure 20-18!).

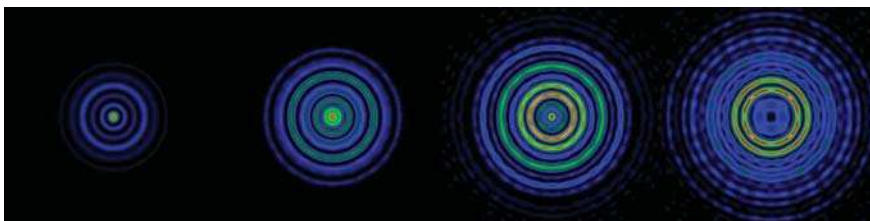


Figure 20-20: Point images with 1λ spherical aberration Z_9 in comparison with spherical aberration of higher order Z_{16}, Z_{25} and Z_{36} .

Diffraction images of different amounts of astigmatism Z_5 and Z_{12} are compared to coma Z_7 and Z_{14} in figure 20-21. Figure 20-22 illustrates the effect of three-wave and four-wave aberrations of higher order on the intensity distribution.

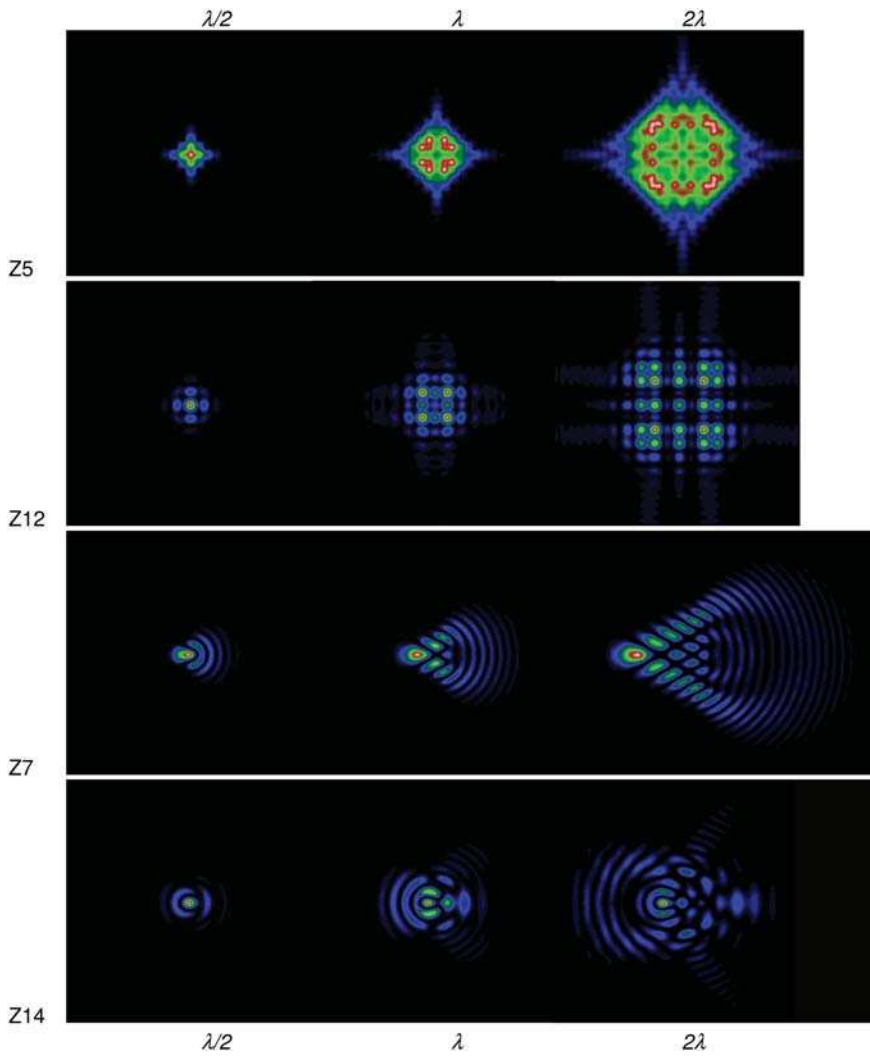


Figure 20-21: Diffraction-optical coma figures for Z_7 and $Z_{14} = \lambda/2, \lambda$ and 2λ .

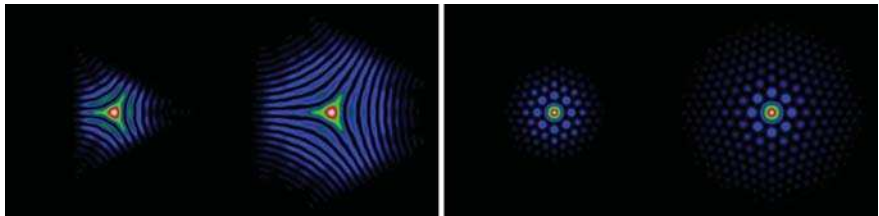


Figure 20-22: Point image at three-wave error $Z_{10} = \lambda$ and 2λ and four-wave error $Z_{17} = \lambda$ and 2λ .

20.4.4

Distortion, Defocus and Astigmatism

After the trivial aberration Z_1 the first two aberrations Z_2 and Z_3 are linear functions in x and y , respectively. A linear phase causes 'only' a shift in an image plane according to the shift theorem of the Fourier transformation:

$$U(x, y) = \frac{NA}{\lambda} \int \int_{a^2+b^2 < 1} e^{ik_0 c_2 a} \cdot e^{ik_0 \cdot NA \cdot (ax+by)} da db = U_{w=0} \left(x + \frac{c_2}{NA}, y \right) \quad (20-73)$$

with the normalized pupil coordinates a and b introduced. So the distortion by a wave-front error Z_2 with the amplitude c_2 is given by

$$\delta x = -\frac{1}{NA} c_2 . \quad (20-74)$$

The Zernike term Z_4 is equivalent to defocusing. For the edge of the pupil with the canonical equation a ray aberration of

$$\delta x = -\frac{4}{NA} c_4 \quad (20-75)$$

is obtained. The position of the axial intercept of a ray with maximum aperture thus is at the defocus position δz with

$$\delta z = \frac{\sqrt{n^2 - NA^2}}{NA} \delta x = 4 \frac{\sqrt{n^2 - NA^2}}{NA^2} c_4 . \quad (20-76)$$

An exact defocusing, however, is equivalent to a hyperbolic phase term and not to a parabolic one. The development of the defocusing from Z_4 using eq. (20-76) is therefore correct in the paraxial approximation only. In general the defocusing or change of the reference plane is taken into account in the characteristic function with (for $n = 1$):

$$\begin{aligned} W_{\text{defocus}} &= \delta z \cdot \sqrt{1 - NA^2 \rho^2} \\ &= \delta z \\ &\cdot \left(1 - \frac{1}{2} NA^2 \rho^2 - \frac{1}{8} NA^4 \rho^4 - \frac{1}{16} NA^6 \rho^6 - \frac{5}{128} NA^8 \rho^8 - \dots \right) . \end{aligned} \quad (20-77)$$

For the calculation of the defocusing from Z_4 the hyperbolic series has to be expanded to Zernike polynomials; so for the defocusing it follows approximately:

$$\delta z \approx \frac{4}{NA^2} c_4 . \quad (20-78)$$

The defocusing vanishes for $c_4 = 0$. If one defocuses according to eq. (20-77) into a new reference plane, so that Z_4 vanishes, the hyperbolic difference phase has to be

taken into account for all higher rotational-symmetric parts. So the coefficients of the spherical aberrations Z_9 , Z_{16} , Z_{25} have to be modified if the reference plane is changed. The same is of course valid for the terms Z_5 and Z_6 , which are equivalent to the astigmatism only in paraxial approximation. For the astigmatic difference it follows that

$$\delta z_{\text{ast}} = \frac{4}{NA^2} c_5 . \quad (20-79)$$

20.4.5

Spherical Aberrations Z_9 , Coma Z_7 and Z_8

Besides the phase term of 4th order the spherical aberration $Z_9 = c_9 (6\rho^4 - 6\rho^2 + 1)$ contains a defocusing with the phase $c_9 6\rho^2$. If first the z -position of the rays is calculated using the spherical aberration without the defocusing term ($c_9 6\rho^2$), i.e., without the defocus of longitudinal spherical aberration one obtains for the transverse ray aberration in polar coordinates

$$\delta r = -\frac{1}{NA} \frac{\partial W(4\text{th order})}{\partial \rho} = -\frac{24}{NA} c_9 \rho^3 . \quad (20-80)$$

Conversion into longitudinal aberrations gives the longitudinal spherical aberration

$$l(\rho) = \frac{\sqrt{1 - \rho^2 NA^2}}{\rho NA} \delta r = -24 \sqrt{1 - \rho^2 NA^2} \frac{\rho^2}{NA^2} c_9 . \quad (20-81)$$

The defocusing term $-c_9 6\rho^2$, however, yields

$$\delta r_{\text{defocus}} = -\frac{1}{NA} \frac{\partial W(O2)}{\partial \rho} = \frac{12}{NA} c_9 \rho \quad (20-82)$$

from which it follows that

$$\delta z_{\text{defocus}}(\rho) = \frac{\sqrt{1 - \rho^2 NA^2}}{\rho NA} \delta r_{\text{defocus}} = 12 \frac{\sqrt{1 - \rho^2 NA^2}}{NA^2} c_9 . \quad (20-83)$$

The aerial image of spherical aberration is therefore longitudinally shifted. The transverse and longitudinal aberrations δr and δz vanish for $\rho = \sqrt{0.5}$. For $\rho = 1$ the focus compensates half of the longitudinal spherical aberration $l(\rho = 1)$. Figure 20-23 compares the negative and positive spherical aberration at different intensity scans through the focus (x - z -plane).

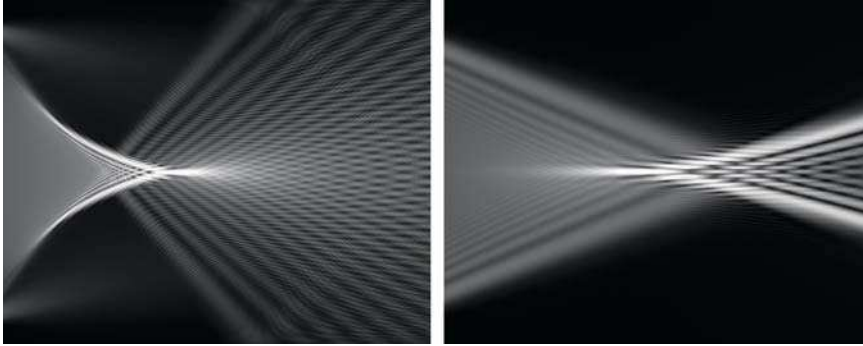


Figure 20-23: Typical distribution of the point- image intensity through the focus for spherical aberration, left negative, right positive spherical aberration (optical axis pointing to the right).

The average of all Zernike polynomials including that of coma $W=c_7(3\rho^3-2\rho)\cdot\cos\varphi$, vanishes over the aperture. Since Z_7 contains a linear term, a distortion of magnitude

$$\delta x = \frac{2}{NA}c_7 \quad (20-84)$$

occurs. So in the case of coma the center of mass of the point image is shifted by $2/NA \cdot c_7$. This is remarkable because the wave-optical shift of the point image with coma is not equivalent to the ray offset of $\delta x_{\text{go}}=3/NA \cdot c_7$ from the ray aberration. The reason for this can be found in the interference, which is neglected in the geometrical-optical discussion. This effect will be considered in more detail in the next section.

20.4.6

Line of Sight

The geometrical-optical consideration of the chief ray of an oblique ray bundle defines its center of mass. In wave optics, the centroid of an intensity distribution is the corresponding term, which is, however, not always equivalent to the chief ray. It can be shown by an elementary calculation, that the centroid of a beam is propagating in free space along a straight line. This is the so-called line of sight [20-21]. With an intensity distribution in image space $I(x,y,z)$, the transverse centroid coordinates in a plane z are given by the first moments according to

$$x_s(z) = \frac{1}{I_0} \cdot \iint x \cdot I(x, y, z) dx dy, \quad (20-85a)$$

$$y_s(z) = \frac{1}{I_0} \cdot \iint y \cdot I(x, y, z) dx dy, \quad (20-85b)$$

with total power I_0 . In the case of asymmetrical wave-front aberrations or inhomogeneous intensity distribution across the exit pupil of an optical system, the line of sight differs from the chief ray of the system. Therefore, in general, the chief ray, the centroid and the lateral peak position of the beam, must be distinguished for every value of z . The reason for the deviation is given by interference effects, which are not considered by simple spot diagrams and parameters introduced in eq. (20-23)–(20-25).

As shown above, the interference effect is particularly observable in systems with coma. In polar coordinates, the wave-front error of order n given by

$$W_n(\rho, \theta) = A_n \rho^n \cos \theta \quad (20-86)$$

with coefficient A_n generates a transverse offset of the centroid of size

$$x_s \approx \frac{2z}{D} \cdot A_n \quad (20-87)$$

where D is the diameter of the exit pupil. Figure 20-24 illustrates the principal behaviour of the centroid line in the case of coma aberrations.

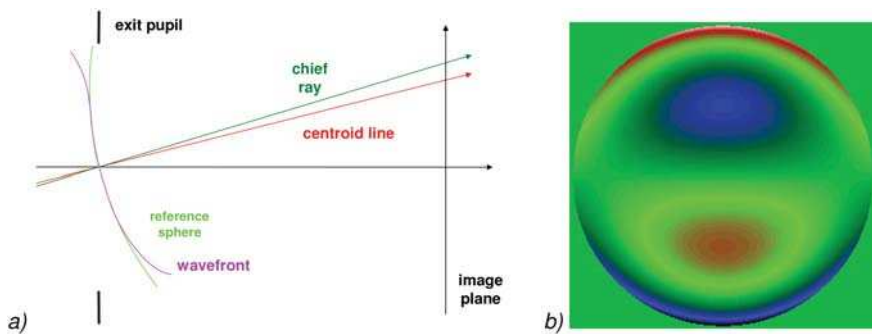


Figure 20-24: a) Line of sight or centroid in a system with b) coma-type wave-front aberration.

Figure 20-25 shows the line of sight from numerical calculations for a system with different amounts of coma according to the definitions of Zernike given above. Additionally, the corresponding paths of the intensity peaks are shown. These are propagating along curves of nearly parabolic form and the intensity maximum has a shape similar to a banana through the focus and therefore this is also known as bananicity. Figure 20-26 illustrates the corresponding transverse intensity profiles of the caustic distributions.

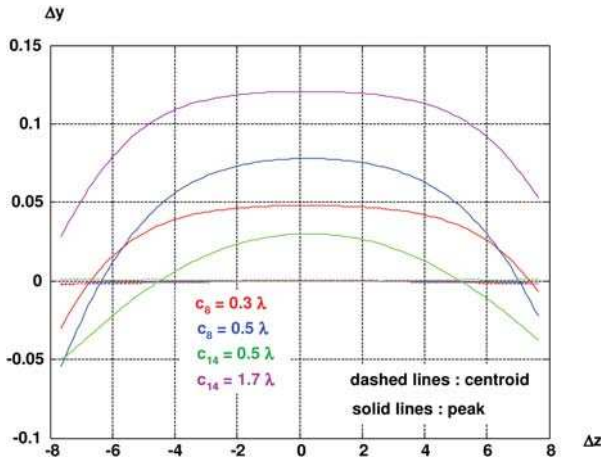


Figure 20-25: Line of sight in a system with coma-type wave aberrations of different size.

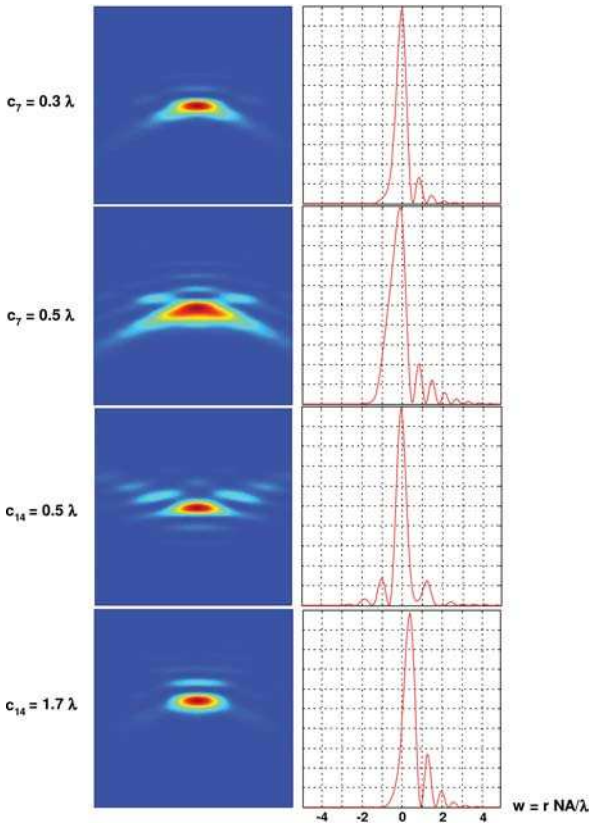


Figure 20-26: Lateral intensity profiles of the point-spread function of the coma aberrations of figure 20-23.

A significant difference in the chief ray and centroid can also be observed for systems with an asymmetric illumination distribution of the exit pupil. Figure 20-27 illustrates the principle and figure 20-28 the focal caustic and the line of sight.

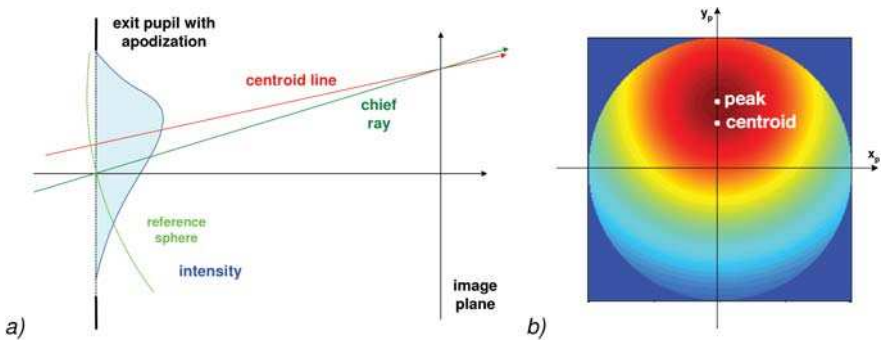


Figure 20-27: a) Line of sight for a system with b) an asymmetric intensity profile in the exit pupil.

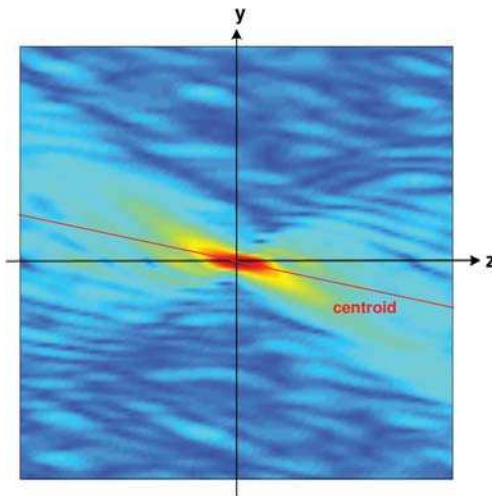


Figure 20-28: Line of sight for a system with an asymmetric intensity profile in the exit pupil in the focal region showing the whole intensity distribution.

20.4.7

Wave Aberrations for Annular Pupils

The Zernike representation is an orthogonal description of the wave-front aberrations only for circular pupils. For other shapes of pupil, adapted orthogonal sets of expansion functions should be applied. As the most relevant example, the expansion of wave-front errors for annular pupils is presented here. Annular pupils are given by systems with central obscuration such as, e.g., mirror telescopes. This polynomial expansion is given by the Tatian polynomials [20-20], [20-22], [20-23]. The azimuthal

part of the Tatian polynomials is identical to the Zernike polynomials. The radial polynomials \bar{R}_n^m depend on the obscuration parameter ε of the annular pupils, given by the ratio of the inner pupil radius NA_{in} and the maximum numerical aperture, with $\varepsilon = NA / NA_{in}$. For radial polynomials the orthogonality relation reads

$$\int_{\varepsilon}^1 \bar{R}_n^m(r, \varepsilon) \cdot \bar{R}_n^m(r, \varepsilon) r dr = \frac{1 - \varepsilon^2}{2(n+1)} \cdot \delta_{n'n} \quad (20-88)$$

and the explicit form of the radial polynomials is given by the complicated expression

$$\bar{R}_n^m(r, \varepsilon) = \approx \frac{\left(\frac{n-m}{2}\right)! \cdot \left(\frac{n+m}{2}\right)!}{(1 - \varepsilon^2)^{n/2}} \cdot \sum_{j=0}^{\frac{n-m}{2}} \frac{(r^2 - 1)^j \cdot (r^2 - \varepsilon^2)^{n/2-j}}{(j!)^2 \cdot \left(\frac{n-m}{2} - j\right)! \cdot \left(\frac{n+m}{2} - j\right)!} \quad (20-89)$$

The first few Tatian polynomials are listed in table 20-3. In figure 20-29, the polynomials of lowest order $n=1...3$ for different obscuration ratios ε are compared. With the obscuration parameter ε a scaling of the radial polynomials on the free part of the aperture ring is obtained. Figure 20-30 illustrates the Tatian polynomials of lowest order for an obscuration parameter $\varepsilon=0.5$ with order numbers corresponding to table 20-3.

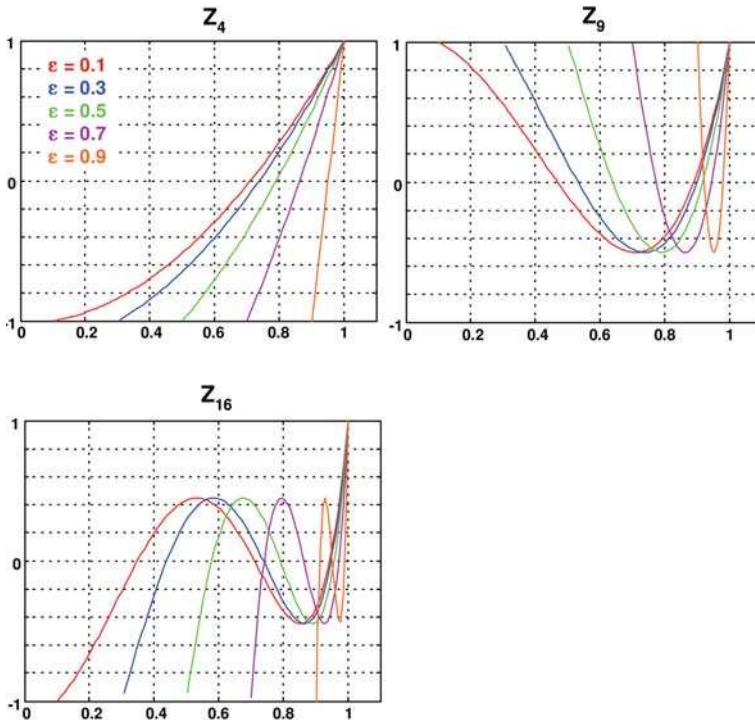


Figure 20-29: Representation of the first 3 Tatian polynomials for an annular pupil with rotational symmetry and five different obscuration ratios ε .

Table 20-3: Tatian Radial polynomials

n	m	Radial polynomial
0	0	1
1	1	$r/\sqrt{1+\varepsilon^2}$
2	2	$r^2/\sqrt{1+\varepsilon^2+\varepsilon^4}$
2	0	$[2r^2 - (1 + \varepsilon^2)]/1 - \varepsilon^2$
3	3	$r^3/\sqrt{1+\varepsilon^2+\varepsilon^4+\varepsilon^6}$
3	1	$[3r^3 \cdot (1 + \varepsilon^2) - 2r \cdot (1 + \varepsilon^2 + \varepsilon^4)]/[(1 - \varepsilon^2) \cdot \sqrt{(1 + \varepsilon^2)(1 + 4\varepsilon^2 + \varepsilon^4)}]$
4	0	$[6r^4 - 6r^2 \cdot (1 + \varepsilon^2) + (1 + 4\varepsilon^2 + \varepsilon^4)]/(1 - \varepsilon^2)^2$
4	2	$\left[4r^4 \cdot (1 - \varepsilon^2) - 3r^2 \cdot \frac{1 - \varepsilon^8}{1 - \varepsilon^6} \cdot (1 - \varepsilon^2)\right] / \sqrt{16 \cdot (1 - \varepsilon^{10}) - 15 \cdot (1 - \varepsilon^8)^2 / (1 - \varepsilon^6)}$
4	4	$r^4/\sqrt{1+\varepsilon^2+\varepsilon^4+\varepsilon^6+\varepsilon^8}$
5	1	$\frac{10r^5 \cdot (1 + 4\varepsilon^2 + \varepsilon^4) - 12r^3 \cdot (1 + 4\varepsilon^2 + 4\varepsilon^4 + \varepsilon^6) + 3r \cdot (1 + 4\varepsilon^2 + 10\varepsilon^4 + 4\varepsilon^6 + \varepsilon^8)}{(1 - \varepsilon^2)^2 \cdot \sqrt{(1 + 4\varepsilon^2 + \varepsilon^4) \cdot (1 + 9\varepsilon^2 + 9\varepsilon^4 + \varepsilon^6)}}$
5	3	$\left[5r^5 \cdot (1 - \varepsilon^2) - 4r^3 \cdot \frac{1 - \varepsilon^{10}}{1 - \varepsilon^8} \cdot (1 - \varepsilon^2)\right] / \sqrt{25 \cdot (1 - \varepsilon^{12}) - 24 \cdot (1 - \varepsilon^{10})^2 / (1 - \varepsilon^8)}$
5	5	$r^5/\sqrt{1+\varepsilon^2+\varepsilon^4+\varepsilon^6+\varepsilon^8+\varepsilon^{10}}$

One of the drawbacks of the application of Tatian polynomials to the description of aberrations is the dependence on the parameter ε . It is therefore difficult to compare the aberration coefficients of systems with different obscuration parameters ε .

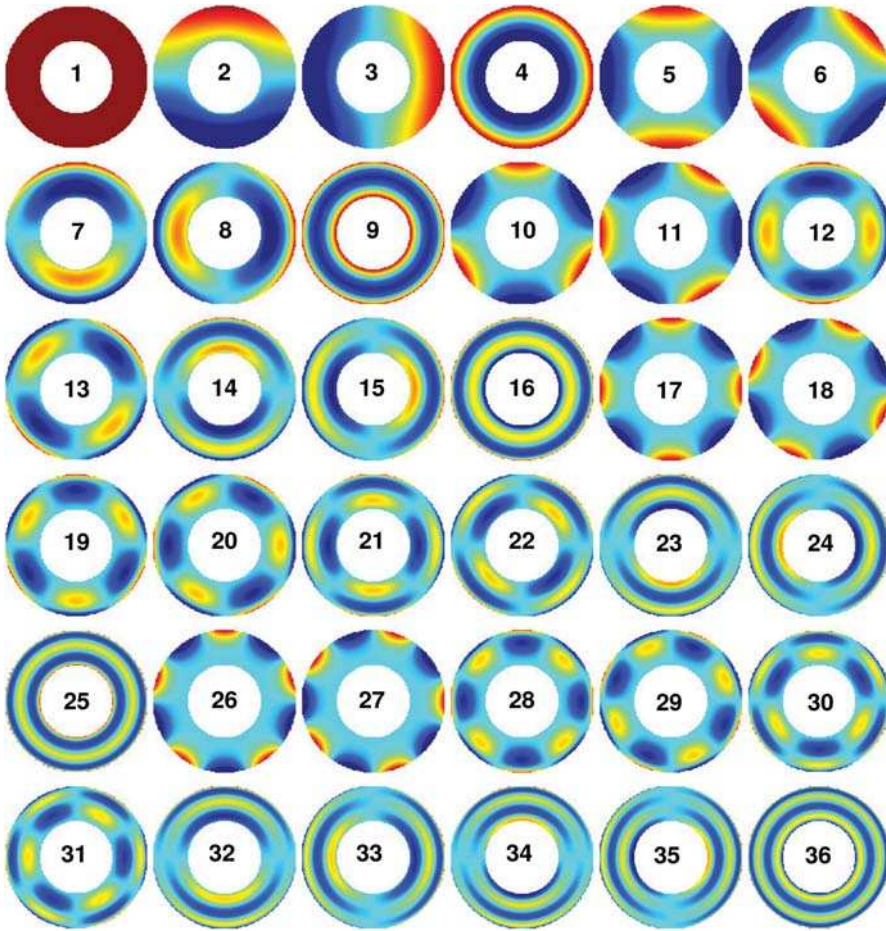


Figure 20-30: Representation of the first 36 Tatian polynomials for an annular pupil with an obscuration ratio of $\varepsilon = 0.5$.

20.4.8

Extended Zernike Expansion

The calculation of the intensity is usually done by solution of the Kirchhoff diffraction integral (see chapter 18) or the corresponding Fourier approximations described in this chapter. The intensity in the image plane is thus obtained by Fourier transformation, with aberrations of the optical system described by Zernike expansion. For defocused positions, the corresponding intensity distribution is either obtained by free-space propagation (see chapter 17), or by consideration of the generalized three-dimensional aperture, as is described in chapter 23. In paraxial approximation, however, the intensity distribution through the focus can be described by a series expansion, the extended Zernike expansion [20-24]. According to the transport of the intensity equation

(17-104), the phase surface in the exit pupil of an optical system approximately determines the intensity distribution in the focal region. After writing (17-100) with

$$\vec{\nabla} \left(I \cdot \vec{\nabla} W \right) = \frac{\partial}{\partial z} \left(I \cdot \frac{\partial}{\partial z} W \right) + \vec{\nabla}_{\perp} \left(I \cdot \vec{\nabla}_{\perp} W \right) \approx \frac{\partial}{\partial z} I + \vec{\nabla}_{\perp} \left(I \cdot \vec{\nabla}_{\perp} W \right) \quad (20-90)$$

the paraxial equation of intensity transport is obtained

$$\frac{\partial}{\partial z} I = -\vec{\nabla}_{\perp} \left(I \cdot \vec{\nabla}_{\perp} W \right). \quad (20-91)$$

This direct dependence of the intensity on the wave-front error in paraxial approximation, motivated an expansion of the general three-dimensional field distribution into Zernike terms with a Zernike representation in the transverse directions and a Taylor expansion along the optical axis. The zero-order term is given by the ideal Airy profile in the image plane with vanishing defocus. A defocus or a degradation of the ideal spherical wave front by aberrations is considered as a disturbance and is taken into account by corresponding correction terms in the expansion. This expansion of the extended Zernike polynomials allows the direct determination of the intensity in the focal region as a function of the Zernike coefficients in the exit pupil in the limit of the paraxial approximation. This relationship is illustrated in figure 20-31.

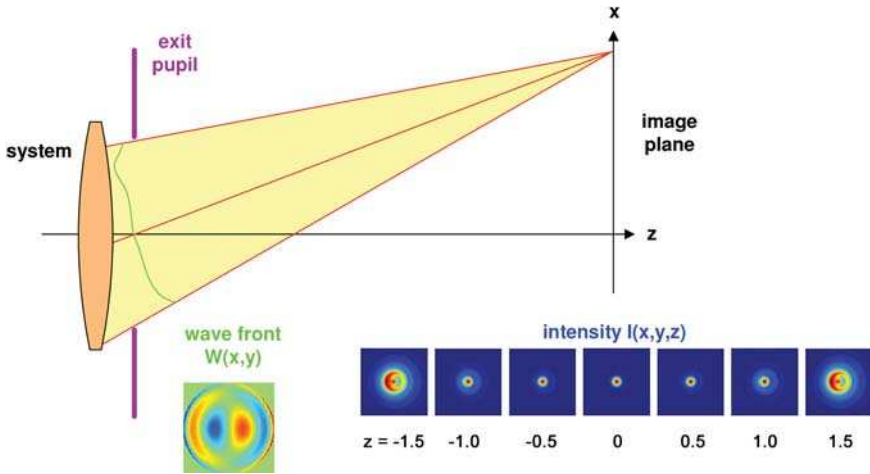


Figure 20-31: Intensity profiles of the focus of an optical system as a function of the wave front in the exit pupil.

For conventional systems with circular pupils, the extended Zernike representation below is formulated in cylindrical coordinates r and φ . The field amplitude in image space is represented by the expression

$$E(r, \varphi, z) = 2 \cdot B_{00}(r, z) + 4\pi i \sum_{n,m \geq 0} A_{nm} \cdot i^m \cdot B_{nm}(r, z) \cdot \cos(m\varphi) + 4\pi i \sum_{n,m < 0} A'_{nm} \cdot i^m \cdot B_{nm}(r, z) \cdot \sin(m\varphi). \quad (20-92)$$

The axial dependence is considered by the transverse expansion coefficients B_{nm} . The coefficients A_{nm} and A'_{nm} are identical to the Zernike coefficients, according to eq. (20-66). In the case of vanishing aberrations, the expression reduces in the image plane to the Airy distribution with

$$B_{00}(r, 0) = 2J_1(r)/r \quad (20-93)$$

It has to be noted that, although the motivation for an expansion of the intensity was restricted to paraxial approximation, a general expansion of the field amplitude in image space is not restricted to paraxial approximations. The derivation given below, however, is limited to the paraxial domain, since the defocus is described by Z_4 [20-24]. The range of validity, especially in the z -direction depends on the number of terms taken into account. The extended Zernike polynomial expansion was generalized for pupils with apodization, large defocus values [20-25] and vectorial effects [20-26]. In particular the apodization can be very easily taken into account. For this purpose, it is only necessary to allow the Zernike coefficients to have complex values. The imaginary part of the coefficients then describes a variation of the amplitude over the pupil area.

For simplification, in the following, the new coordinates (old coordinates have primes) are scaled according to

$$x = x' \cdot 2\pi \cdot \frac{n \cdot \sin u}{\lambda}, \quad (20-94a)$$

$$y = y' \cdot 2\pi \cdot \frac{n \cdot \sin u}{\lambda}, \quad (20-94b)$$

$$z = \frac{\pi \cdot z' \cdot n^2 \cdot \sin^2 u}{\lambda} = \pi \cdot \frac{z'}{R_E}. \quad (20-94c)$$

After a lengthy derivation, we obtain [20-24]:

$$E(r, \varphi(z)) = 2 \cdot \frac{2J_1(r)}{r} + 4\pi i \cdot \sum_{n,m \geq 0} B_{nm} \cdot i^m \cdot \cos(m\varphi) \cdot e^{iz} \cdot \sum_{l=1}^{\infty} (-2iz)^{l-1} \sum_{j=0}^p b_{lj}^{nm} \cdot \frac{J_{m+l+2j}(r)}{l \cdot r^l} \quad (20-95)$$

with the factors

$$b_{lj}^{nm} = (-1)^p \cdot (m+l+2j) \cdot \binom{m+j+l-1}{l-1} \cdot \binom{j+l-1}{l-1} \cdot \frac{\binom{l-1}{p-j}}{\binom{q+l+j}{l}} \quad (20-96)$$

and with $p = (n-m)/2$ and $q = (n+m)/2$. In practice, the axial expansion is carried out up to a maximum index l_{\max} . It is, however, difficult to estimate the re-

quired number of terms for the z -expansion in order to achieve a desired accuracy [20-27]. The maximum number l_{\max} depends on the radial order n_{\max} and the defocus z . The behaviour and the contribution of the terms for the sum is illustrated in figure 20-32a. The left picture shows the size of the individual terms and the sum over all terms for increasing l . The right picture shows the resulting error when truncating the series expansion at the corresponding number of terms. In the special case shown here, approximately 90 terms are necessary to get a relative accuracy of 10^{-5} . As shown in figure 20-32b, the dependence of the required number of series terms as a function of the defocus z is nearly linear and can be approximated by the empirical equation

$$l_{\max} = 8.5 \cdot |z_{\max}| + 15 \tag{20-97}$$

with the scaling according to eq. (20-94c). Figure 20-33 illustrates the point-spread function obtained by the solution of eq. (20-95) for spherical aberration, coma and astigmatism for different focus positions.

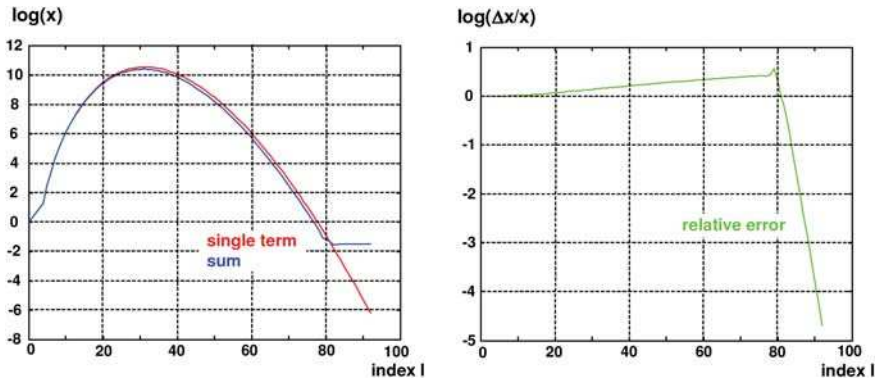


Figure 20-32: a) Size of the individual terms and the complete sum for truncating the l -sum and resulting error of the expansion.

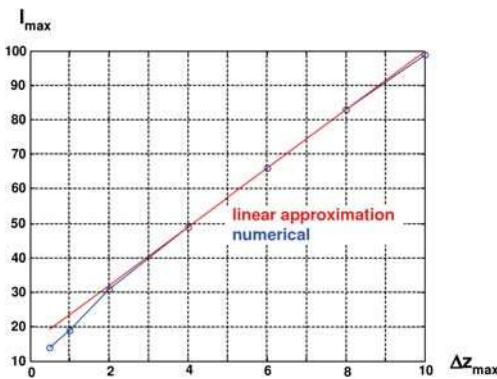


Figure 20-32: b) Required number l_{\max} as a function of the defocus range z .

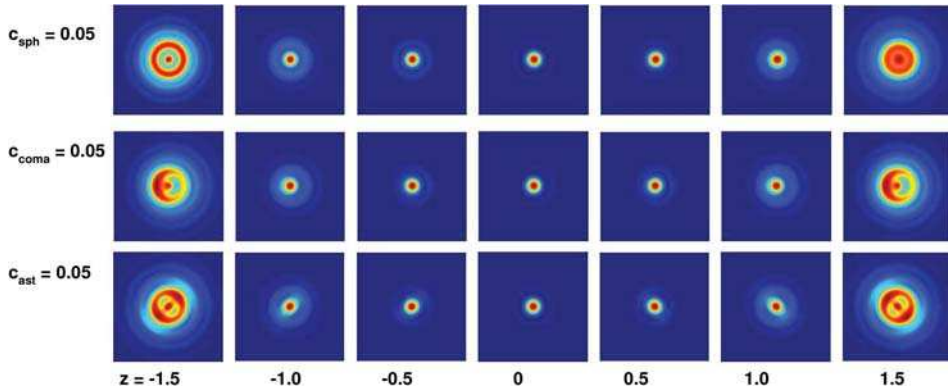


Figure 20-33: Point-spread functions calculated by the extended Zernike expansion for three different aberrations and different focus positions z .

20.5 Helmholtz–Lagrange Invariant and Phase-space Description

20.5.1 The Phase Space

The phase space description of geometrical optics is a convenient method for the illustration of ray curves and properties of optical systems. As shown in this chapter, the phase-space volume corresponds to the Helmholtz–Lagrange invariant of optical systems, which is a useful quantity to describe the complexity of optical systems as well as for light flux calculations, for example. As will be shown, incoherent optical imaging can be transferred to the phase space in a simple way.

The phase space is given by the representation of the optical direction cosines p_x versus the position x and p_y versus y , respectively. It is equivalent to the Wigner distribution function description, if

$$\vec{p} = \lambda \cdot \vec{\nu} \quad (20-98)$$

is taken into account. A Fourier transformation exchanges the roles of position and frequency coordinates. Therefore the phase space representation in the pupil plane as the Fourier plane to the object plane follows from the phase space representation in the object plane by a rotation of 90° and a scaling with focal length f [20-28], [20-29]. Figure 20-34 illustrates the rotation property of the field distribution given in the front and back focal planes of a lens in phase space for the example of a single oblique bundle of parallel rays focused to x_p .

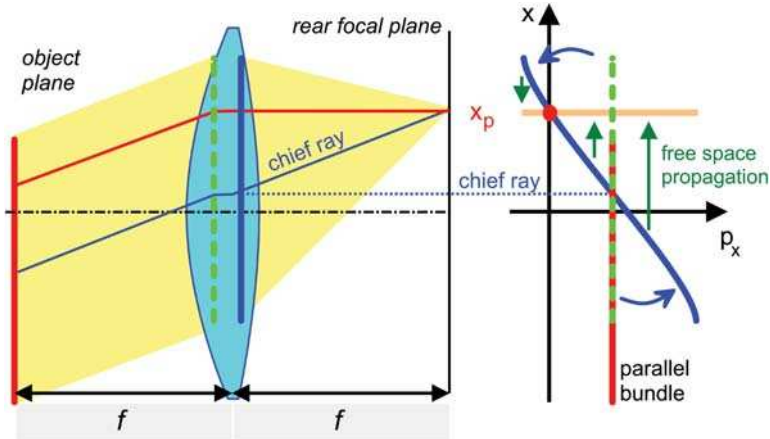


Abb. 20-34: Rotation of a ray bundle in phase space.

Figure 20-35 illustrates the phase-space transformation for an imaging set-up. From each object point a bundle of rays, according to the numerical aperture, is received by the objective. The phase space for the object plane is given by a rectangle of width L and the height $2NA$. In the pupil plane the bundle of rays to an object point completely fills the pupil with width $D = 2f_1 \cdot NA$, with the bundle assuming the angle $\sin \gamma = L/2f_1$ to the field edge point. The extension in the direction of the axis of the direction cosine is thus $F = L/f_1$. By repeating the rotation and scaling with the focal length f_2 , once again a rectangle is obtained of width $L' = \beta L$ and height $2NA' = 2NA/\beta$ in the image plane.

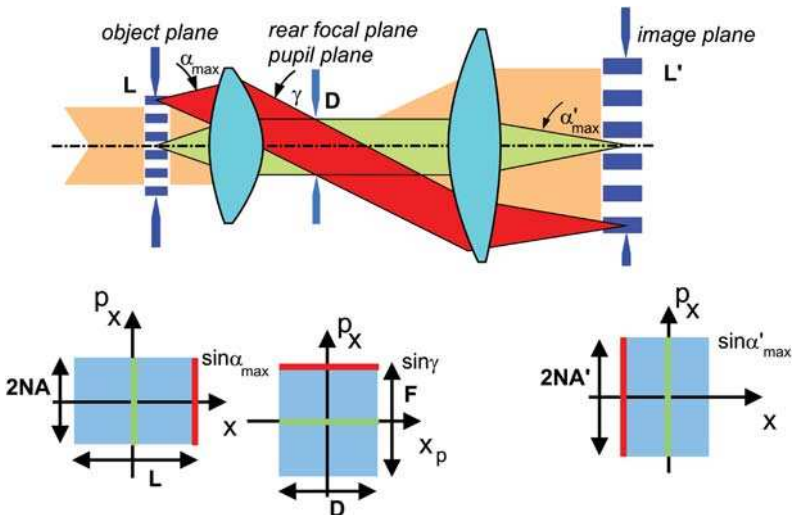


Figure 20-35: Phase space representation of optical imaging.

As it can be easily shown, in each plane the area and the phase-space volume, respectively, are equal:

$$G_{1D} = 2NA \cdot L = D \cdot F = \frac{D \cdot L}{f} . \quad (20-99)$$

The one-dimensional phase-space volume above is also known as the *Helmholtz–Lagrange Invariant*. In many textbooks of optics the Helmholtz–Lagrange Invariant is derived as a paraxial principle; for rotational-symmetric systems, though, the principle is generally valid. It is equivalent to an invariant Wigner distribution function density.

In any imaging optical system, the phase-space volume is an invariant quantity comparable to the energy conservation of conservative systems. In general, the phase-space volume is given by:

$$G(z) = \int g d\Gamma = \iiint g(x, y, p, q; z) dx dy dp_x dp_y = \int \int_{A, \Omega} dA \cos \alpha d\Omega = \text{const} \quad (20-100)$$

with infinitesimal surface area dA and infinitesimal direction cosines dp_x , dp_y , expressed by the solid angle element $d\Omega$. The integral has to be carried out over the area of the object plane and the area of the pupil or aperture. This can be expressed in a more general way by the phase-space function $g(x, y, p, q; z)$ as a function of the z -coordinate [20-30]. For this it is $g(z) = 1$, if a ray with these phase-space coordinates can occur, otherwise it is $g = 0$. Because of the resolution limits the smallest phase-space cell or a phase-space point in the image plane, is given by the volume $\delta g = \delta x \cdot \delta v$. This is approximately equivalent to a single geometrical light ray. The invariance of eq. (20-99) has already been assumed by Straubel [20-31]. Between the object and the image plane of an aplanatically corrected system, Straubel's theorem is always trivially fulfilled, and it can in general be shown that it holds for any system and any area in between object and image planes [20-32]:

$$dA_0 \cos \alpha_0 d\Omega_0 = dA_1 \cos \alpha_1 d\Omega_1 . \quad (20-101)$$

A direct consequence of the conservation of the phase-space volume for magnifying microscopic imaging was described by Abbe [20-33]:

‘Under no conditions can the radiance of the microscopic image exceed that one with which the object would appear to the naked eye.’

Either the free aperture of the lens or the illuminating conic bundle of light, limits the brightness of the image. In general, the brightness depends on the magnification β with $\sim 1/\beta^2$.

While Helmholtz–Lagrange Invariant or space–bandwidth products are used in imaging optical systems, the equivalent terminus geometrical flux or étendue is employed in the characterization, design and simulation of light sources and illumination of optical systems [20-34], [20-35], [20-36], [20-37], [20-38]. As a consequence of the conservation of étendue, it cannot be reduced with light losses. This has par-

ticular meaning for the illumination and design of illumination systems. In the Abbe explanation of image formation it is assumed that all light-source points can be considered as incoherent. For incoherent illumination the geometrical flux of a light source filling the entrance pupil of an optical system is thus identical to the phase-space volume of the optical system. If the effective source is too large, light will be lost, if it is too small, it may have a severe impact on the imaging. These impacts and more will be treated in the next paragraph in more detail.

20.5.2

The Resolution Limit in the Space Domain and in the Spatial Frequency Domain

The resolution limit, written as a product from canonical conjugates has often been compared with the Heisenberg principle of uncertainty. Writing δx for the smallest resolvable structure we obtain

$$\delta x \cdot \Delta p \geq \lambda \tag{20-102}$$

with Δp equivalent to the maximum difference of the optical direction cosines or quasi-momenta which can be related to the spatial frequency by eq. (20-97), and it follows that $\delta x \Delta \nu \geq 1$. The difference from the Heisenberg principle of uncertainty lies in the Heisenberg constant, which is to be replaced by the wavelength.

In optics, the position and spatial frequency or direction cosine, respectively, are Fourier-inverse quantities. From this, eq. (20-102) becomes clear immediately. By inserting, e.g., a Gaussian function of the width δx as the minimum object, one obtains again a Gaussian function of width $1/\delta x$ as the spectrum (figure 20-36):

$$u(\nu) = \int e^{-\frac{\pi x^2}{\delta x^2}} e^{-2\pi i x \nu} dx = |\delta x| e^{-\delta x^2 \pi \Delta \nu^2} \tag{20-103}$$

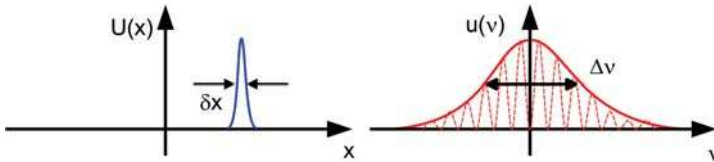


Figure 20-36: The product of the peak-widths of two Fourier-inverse Gaussian functions is equal to 1.

With $1/\delta x = \Delta \nu$ one obtains eq. (20-102) immediately. For optical imaging systems with a finite aperture and with pupil diameter D one obtains

$$\Delta \nu = \frac{1}{\delta x} = \frac{2NA}{\lambda} = \frac{D}{\lambda f} \tag{20-104}$$

with the equivalent point images instead of the Gaussian function, and in the one-dimensional case, the sinc function of the width $2NA$ (figure 20-37).

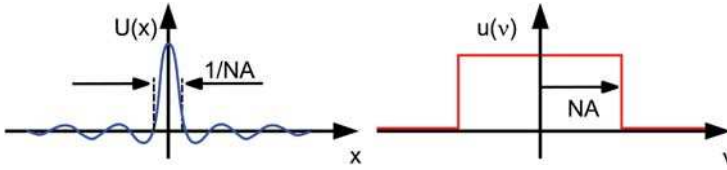


Figure 20-37: Aperture limitation and point image for one-dimensional imaging.

On the other hand, the direction of a wave can only be determined with limited accuracy. Starting with an incident finite planar wave limited by a field stop of extent L , one obtains a blurred focus in the rear focal plane. By writing ‘true’ coordinates in the object plane as x and in the pupil plane as x_p instead of spatial frequencies, one obtains from eq. (20-104), the optical invariant

$$\delta x \cdot D = L \cdot \delta x_p = \lambda f. \quad (20-105)$$

The resolution limits are

$$\delta x = \frac{\lambda f}{D} = \frac{\lambda}{2NA} \quad (20-106)$$

and

$$\delta x_p = \frac{\lambda f}{L}. \quad (20-107)$$

The minimum frequency resolution $\delta\nu$ of a Fourier-transforming optical system is given by

$$\delta\nu = \frac{\delta p}{\lambda} = \frac{\delta x_p}{\lambda f} = \frac{1}{L}. \quad (20-108)$$

From eq. (20-108) from which the analogy to equation (20-102) follows:

$$L \cdot \delta p \geq \lambda. \quad (20-109)$$

Thus, the spatial resolution of an imaging optical system is limited first, and then the frequency resolution of a Fourier-transforming optical system. Since the pupil coordinates x_p scale with focal length f , the spatial frequency resolution is limited by the extension L of the field only. This consideration is of particular relevance for grating spectrometers, where incident waves of different wavelengths are split, e.g., by a grating of length L , into waves with different propagation angles (figure 20-38). For a given grating period, the frequency resolution is determined only by the grating length.

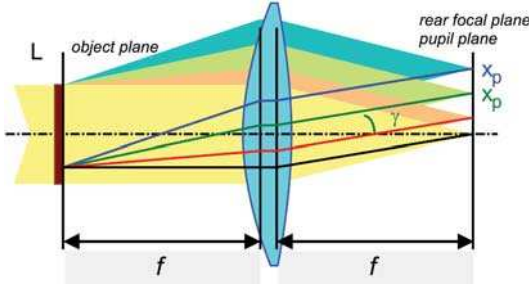


Figure 20-38: Illustration of the restricted spectral resolution for a Fourier-spectrometer. The spectral resolution is limited by the length L of the object plane.

20.5.3

The Space–Bandwidth Product

The possible minimum phase-space volume for an image point is given by the product of the minimum spatial resolution and the aperture, i.e., in the one-dimensional case according to eq. (20-102) by

$$\delta x \cdot \Delta \nu = \lambda . \tag{20-110}$$

From the ratio of the phase-space volume and the minimum phase-space step or information cell, the number of resolvable image points N follows

$$N = \frac{G_{1D}}{\lambda} = \frac{L}{\delta x} = L \cdot \Delta \nu = \frac{LD}{\lambda f} \approx 2 \cdot L \cdot \frac{NA}{\lambda} . \tag{20-111}$$

For two-dimensional imaging, of course, the image area A and the pupil area have to be taken into account. With the minimum phase-space area $\pi \lambda^2$ the phase-space volume follows:

$$G_{2D} = \pi NA^2 \cdot A . \tag{20-112}$$

The phase-space volume as a measure of resolvable image points is equivalent to the space–bandwidth product SBP of information theory and gives a rough number for the complexity of an optical system [20-39]:

$$SBP = \frac{G_2}{\pi \lambda^2} = A \cdot \left(\frac{NA}{\lambda} \right)^2 = \frac{A}{(\delta x)^2} . \tag{20-113}$$

According to eq. (20-111) and (20-112) the number of resolvable image points is thus given by the product of the field area A times the pupil area divided by λ^2 . The larger the space–bandwidth product, the better corrected is the optical system, the more effort has to be made to fabricate the optical system, and finally the higher the

price. The space–bandwidth product ranges from 1 for fiber coupling or beam-shaping optical systems, to the high-end lithography optical systems in the order of 10^{11} , corresponding to 10^{11} independently imaged single-image points! Due to this large space–bandwidth product, free-space imaging optics has frequently been recommended for interconnections in computing and telecommunication.

20.6

Literature

- 20-1 R. K. Luneburg, *Mathematical theory of optics* (University of Colorado Press, Berkeley and Los Angeles, 1966).
- 20-2 O. Stavroudis, *The Optics of Rays, Wave fronts and Caustics* (1972).
- 20-3 M. Born, *Optik*, Springer 1933.
- 20-4 K.-H. Schuster et al., *Mikrolithographisches Projektionsobjektiv*, German Patent Application DE 19855108.
- 20-5 W. Ulrich, S. Beiersdörfer, H.-J. Mann, Trends in optical design of projection lenses for UV- and EUV-Lithography, *SPIE* **4146**, 13–24 (2000).
- 20-6 C.J.R. Sheppard, M. Gu, Imaging by high aperture optical system, *Journ. Mod. Opt.* **40**, 1631–1651 (1993).
- 20-7 M. Born, E. Wolf, *Principles of Optics*, 6th ed. (Pergamon Press, Oxford, New York, 1980).
- 20-8 A. Brouwer, E.L. O’Neill, and A. Walther, The role of eikonal and matrix methods in contrast transfer calculations, *Appl. Opt.* **2**, 1239–1246 (1963).
- 20-9 E. Wolf, “Electromagnetic diffraction in optical systems I. An integral representation of the image field”, *Proc. Royal Soc. A* **253**, 349–357 (1959).
- 20-10 Jakob J. Stamnes, *Waves in focal regions* (Adam Hilger, Bristol and Boston, 1986).
- 20-11 C. C. Cole, E. Barouch, U. Hollerbach and S. A. Orszag, Derivation and Simulation of Higher Numerical Aperture Scalar Aerial Images *Jpn. J. Appl. Phys.* **31**, 4110 (1992).
- 20-12 G. M. Gallatin, High-numerical-aperture scalar imaging, *Appl. Opt.* **40**, 4958 (2001).
- 20-13 M. Gu, *Advanced Optical Imaging Theory* (Springer, Berlin, 2000).
- 20-14 G.B. Airy, On the diffraction of an object glass with circular aperture, *Transactions of the Cambridge Philosophical Society* **5**, 283–291 (1835).
- 20-15 B.P.Ramsey, E.L. Cleveland, O.T.Koppius, Criteria and the Intensity-Epoch slope, *J. Opt. Soc. Am.* **31**, 26–33 (1941).
- 20-16 H. Osterberg, Microscope Imagery and resolution, *J. Opt. Soc. Am.* **40**, 295–303 (1950).
- 20-17 H. Koehler, On Abbe’s theory of image formation in the microscope, *Optica Acta* **28**, 1691–1701 (1981).
- 20-18 I.P. Agurok, Automation of selection of the elements entering into the assembly of commercially produced objective lenses, *Sov. J. Opt. Technol.* **52**, 400–403 (1985).
- 20-19 I.P. Agurok, Double Expansion of wave-front deformation in Zernike polynomials over pupil and field of view of optical systems, *Proc. SPIE* **3430**, 80–87 (1999).
- 20-20 B. Tatian, Aberration balancing in rotationally symmetric lenses, *J. Opt. Soc. Am.* **64**, 1083–1091 (1974).
- 20-21 V. N. Mahajan, Line of sight of an aberrated optical system, *J. Opt. Soc. Am. A* **2**, 833 (1985).
- 20-22 W. H. Swantner and W. H. Lowrey, Zernike-Tatian polynomials for interferogram reduction, *Appl. Opt.* **19**, 161 (1980).
- 20-23 V. N. Mahajan, Zernike annular polynomials for imaging systems with annular pupils, *J. Opt. Soc. Am.* **71**, 75 (1981).
- 20-24 J. J. Braat, P. Dirksen and A. J. Janssen, Assessment of an extended Nijboer-Zernike approach for the computation of optical point-spread functions, *J. Opt. Soc. Am. A* **19**, 858 (2002).
- 20-25 A. J. Janssen, J. J. Braat and P. Dirksen, On the computation of the Nijboer-Zernike aberration integrals at arbitrary defocus, *Jour. of Mod. Opt.* **51**, 687 (2004).
- 20-26 J. J. Braat, P. Dirksen, A. J. Janssen and A. S. van de Nes, Extended Nijboer-Zernike representation of the vector field in the focal

- region of an aberrated high-aperture optical system, *J. Opt. Soc. Am. A* **20**, 2281 (2003).
- 20-27** A. J. Janssen, Extended Nijboer-Zernike approach for the computation of optical point-spread functions, *J. Opt. Soc. Am. A* **19**, 849 (2002).
- 20-28** H.O. Bartelt, K.-H. Brenner, A.W. Lohmann, The Wigner distribution function and its optical production, *Optics Commun.* **32**, 32–38 (1980).
- 20-29** A. Lohmann, Image rotation, Wigner Rotation, and the fractional Fourier transform, *J. Opt. Soc. Am. A* **10**, 2181–2186 (1993).
- 20-30** A. Gitin, Radiometry, A comprehensive approach, *J. Opt. Technol.* **65**, 132–140 (1998).
- 20-31** R. Straubel, Über einen allgemeinen Satz der geometrischen Optik und einigen Anwendungen, *Physik. Zeitschr.* **4**, 114–117 (1902/03).
- 20-32** W.T. Welford, R. Winston, *High Collection Nonimaging Optics* (Academic Press, 1989).
- 20-33** E. Abbe, Beiträge zur Theorie des Mikroskops und der mikroskopischen Wahrnehmung, *M. Schultze's Archiv für mikroskopische Anatomie*, Vol. **IX**, 413–468 (1873).
- 20-34** M. Berek, Energiebilanz bei Monochromatoren, *Zeitschrift für Instrumentenkunde*, **56**. Jahrgang, 109–115 (1936).
- 20-35** G. Hansen, Die Abbildung von Volumenstrahlern in Spektrographen, *Optik* **6**, 337–347 (1950).
- 20-36** W. Klaunig, Der Lichtleitwert, *Feingeräte-technik* **2**. Jg. Heft 4, 179–182 (1953).
- 20-37** E. Helbig, Grundsätzliches zur Ausleuchtung von optischen Systemen, *Feingeräte-technik* **21**. Jg., Heft 2, 57–60 (1972).
- 20-38** G. Derra, W. Singer, "Collection efficiency of EUV sources", *Proc. SPIE* **5037**, 728–742 (2003).
- 20-39** G. Toraldo di Francia, Resolving power and Information, *J. Opt. Soc. Am.* **45**, 497–501 (1955).

21

The Abbe Theory of Imaging

- 21.1 Introduction 240
- 21.2 Phenomenological Description of Imaging 244
 - 21.2.1 The Explanation of Image Formation According to Abbe and the Abbe Resolution 244
 - 21.2.2 The Information About an Object Contained in an Image 249
 - 21.2.3 Koehler Illumination and the Visibility 252
 - 21.2.4 The Siedentopf Illumination Principle 255
 - 21.2.5 Imaging with Different Colours 259
 - 21.2.6 Aplanatic Correction and Geometrical Optics 260
- 21.3 The Mathematical Description of Fourier Optical Imaging 262
 - 21.3.1 Imaging with Uncorrelated Light Sources 262
 - 21.3.2 Consideration of Magnification 267
- 21.4 Coherence in Imaging 269
 - 21.4.1 The Coherent Image 269
 - 21.4.2 Incoherent Imaging 272
 - 21.4.3 One-Dimensional Incoherent Imaging 273
 - 21.4.4 Systems with Rotational Symmetry 275
 - 21.4.5 Conditions for Incoherent, Partially Coherent and Coherent Imaging 277
 - 21.4.6 Imaging with Correlated Light Sources 280
 - 21.5 Literature 281

21.1

Introduction

The question of the nature of light and image formation has stimulated the thoughts and speculation of many philosophers and scientists for a long time. During the time of Ernst Abbe there were two different theories of imaging. Until the end of the 19th century, image formation was frequently described by central projection, with the Camera Obscura as the most simple optical imaging instrument (figure 21-1).

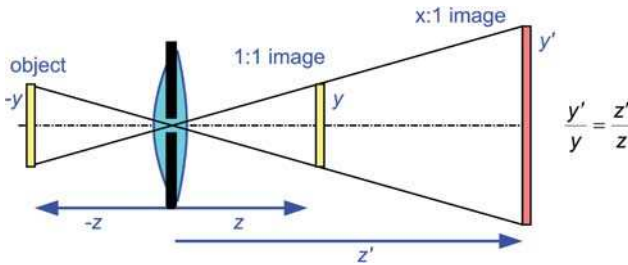


Figure 21-1: The central projection allows every imaging scale.

Within the scope of geometrical optics, both refraction and reflection of light rays at interfaces were considered, and the geometrical ray aberrations were also taken into account. For example, in photographic imaging, optical imaging follows the rules of the Camera Obscura, i.e., the rules of central projection, to a good approximation. This means that any arbitrary magnification of an object should be achievable:

$$\beta = \frac{y'}{y} = \frac{z'}{z} . \quad (21-1)$$

However, this does not mean that an object, however small, can be imaged. As was shown eventually by Abbe in his explanation of optical imaging, the minimum object size is determined by the resolution limit. Although, in principle, the image can be formed with arbitrarily high magnification, e.g., by the appropriate selection of the focal length of the eyepiece, it does not contain any more information about the object. That is why many of the early users of microscopes failed to gain any additional image details simply by increasing the magnification. Resolution and magnification do not have anything in common and the resolution is determined solely by the numerical aperture! Therefore, microscopes have a 'useful magnification', beyond which the imaging of object details cannot be improved.

Diffraction was at first thought to play a role only in astronomical imaging since, in Airy's approach, the diffracting influence of the pupil, i.e., the physical limit of a lens, is considered. The imaging of self-luminous objects like stars was described by Airy who discovered that an object wave is truncated at the aperture of a telescope with diameter D and thus fades to a diffraction image or point-spread function. The point-spread function is equivalent to the Fourier transform of the diffracting aperture of the system and determines the well-known resolution limit with the diameter d [21-1], [21-2]:

$$d = 1.22 \cdot \frac{\lambda}{NA} = 2.44 \frac{\lambda f}{D} . \quad (21-2)$$

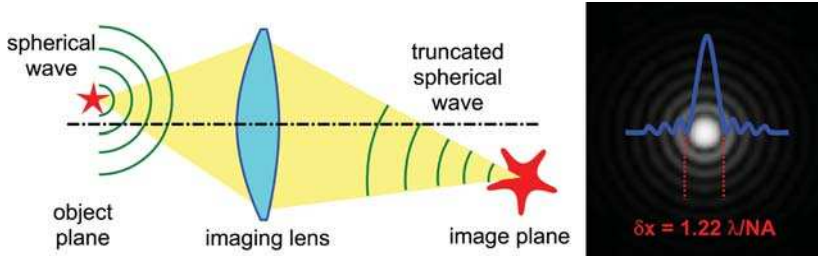


Figure 21-2: Image formation of self-luminous objects.

The geometric-optical image point is consequently additionally blurred by the diffraction spot. The latter is determined by the aperture angle, the wavelength of the light, and the aberrations of the system. So, unlike geometric-optical imaging, each image point is blurred according to the point-spread function of the imaging system. For ideal, perfectly incoherent imaging, Airy's procedure already constitutes an exact calculation of the image intensity. As will be shown in this chapter, for the incoherent case, Airy's approach is indeed equivalent to Abbe's explanation of image formation. The diffraction image, also known as the Airy disc, is described in more detail in chapter 20.

In the 1820s, Joseph Fraunhofer observed that periodic objects, like gratings, diffract light to form separate diffraction orders in the far field, i.e., far from the grating object. Thereby the individual diffraction orders have the same shape and size in accordance with the divergence of the incident light at the grating. Ernst Abbe concluded from Fraunhofer's observation of diffraction, that microscopic imaging does not follow either the rules of central projection, as in photographic imaging, or the rules of astronomical observation. The image formed by interference reproduces the characteristic features of the object according to the diffraction orders passed by the lens.

It was therefore Ernst Abbe, who finally gave the first comprehensive description of image formation in 1873 [21-3], [21-4]. The complete and exact description of partially coherent optical imaging is given by a discussion of the diffraction at the object and image formation by interference (figure 21-3). The exact mathematical description, which Abbe had announced several times, was never published by him. However, once the cornerstone was in place, the whole theory became clear and was worked out in detail by Abbe's co-workers and students, such as Czapski, Koehler and Siedentopf. From this it turned out that the basis of the Abbe theory of image formation was Fourier optics. Fourier optics can in general be described by the application of the expansion of the electromagnetic field into planar waves by Fourier transformation. After Abbe and his co-workers, the description of optical imaging by Fourier methods was then presented by Hopkins in the middle of the last century, following investigations on the coherence of light by Van Cittert and Zernike [21-5], [21-6], [21-7].

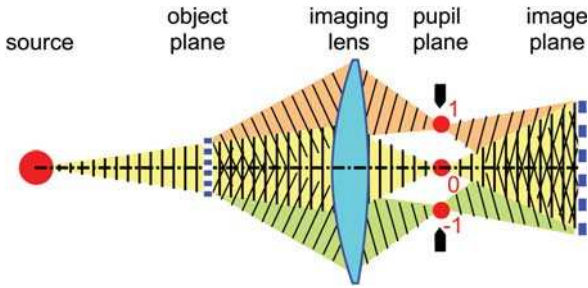


Figure 21-3: Abbe Theory of image formation by diffraction and interference.

After the invention of the laser as a coherent light source, Fourier optics was used more widely and now has new fields of application. For instance, the laser has opened up the possibility of Fourier holography and optical data processing. Several general descriptions of Fourier optics have been published, the first being the book by Goodman 'Introduction to Fourier Optics' [21-8] which has become a standard text book. However, the introduction of Fourier methods is often carried out in the paraxial or far-field approximation. This has often led to the mistaken assumption that the application of Fourier methods in optics is limited to the paraxial field. Here we attempt to cover Fourier optics, and thus the Abbe theory of image formation, more generally.

Ernst Abbe's and August Koehler's work received much attention although many misunderstandings also occurred. For example, Abbe's experimental set-up in the experiments confirming his theory led to the mistaken assumption that the ideal illumination has to be spatially coherent, i.e., one must illuminate with small angular spectra in order for the diffraction orders to be separated in the pupil. The application of oblique illumination with a 'small' divergence of the light source is of course only useful to obtain maximum contrast at the resolution limit. This idealized imaging condition was often found not to be useful in practical applications.

Edward Nelson apparently understood the actual meaning of Abbe's theory and tried to convince the Royal Society of Microscopy of the usefulness of illumination with a large angular spectrum [21-9]. Nelson called this illumination the *critical illumination* when the illumination aperture fills at least 3/4 of the objective pupil ($\sigma > 0.75$). In order to realize this large illumination aperture, Nelson imaged the light source onto the object with a strong diminution. The term critical illumination is today used in this sense, i.e., when an image of the light source is formed on the object plane. Using light bulbs, for example, critical illumination is disadvantageous because the object plane is not illuminated homogeneously. That is why there has been a common misunderstanding that critical illumination in which the light source is imaged onto the object plane is unhelpful and that Koehler illumination would therefore be preferable. However, if the light source is sufficiently large and homogeneous there is no difference between image formation with Koehler or with critical illumination [21-10].

Nelson further argued that his wide-angle cone of illumination – incoherent illumination – is a superior means of suppressing aberrations of microscopic lenses.

He found, as is still valid today, that, in general, the aberration level of imaging instruments depends on the illumination, so that the best focus position for some microscopes depends on the illumination setting and is now automatically corrected.

For quite a long time after his death, the understanding of image formation as an interference phenomenon, as stated in Abbe's theory, remained difficult to accept. Even in 1966, L. C. Martin wrote in his comprehensive textbook '*...however, the interference phenomenon in the image plane is not a true image*' [21-11, p. 262]. In his article series *Coherence and Consonance* in 1926, Max Berek put forward the hypothesis that only self-luminous objects would be imaged 'correctly' according to Airy's theory (primary image), i.e., with the light source in the incident pupil, while the interference pattern would be, at best, similar to the object (secondary image) [21-12]. The distinction between the images of self-luminous and illuminated objects and the corresponding primary and secondary images is now only of historical interest thanks to the discoveries of Mandelstam, Laue, Van Cittert and Zernike but it illustrates how facts which are taken for granted today were regarded as obstacles during the development of the theory [21-13], [21-14], [21-15], [21-5], [21-6].

Berek's refusal to accept Abbe's theory was strengthened by the observation of the afocal nature of the interference pattern, particularly for coherent illumination (i.e., a small light source). In particular, when only two diffraction orders contribute to the image formation, '*there is no preferred image plane*' [21-12]. Even for multiple interferences there are many image planes where a similar pattern is formed, i.e., a '*periodic repetition of equivalent image planes*' – the Talbot planes. Berek therefore concluded that the concept of depth resolution does not exist for the imaging of non-self-luminous objects due to the afocal nature of the secondary interference pattern. Coherent oblique illumination *compromises resolution, degrades depth discrimination and introduces a disturbing apparent movement with change of focus* [21-16]. H.H. Hopkins also noticed the afocal nature of the interference phenomenon, especially for the imaging of fine gratings beyond the coherent resolution limit, with only two diffraction orders contributing to the image formation, when he attempted to determine the transfer function of an optical system with test gratings and different illumination scenarios. He then realized that the imaging of gratings beyond the resolution limit at coherent illumination is independent of the aberrations (like defocusing) of the optical system [21-17]. Martin recognized '*indeed there can be no change with changing focus, so that any focusing of the image is out of the question*'. After a certain grating period it is no longer possible to tune the image plane sharply. Everyone who wishes to observe a grating in a high-aperture microscope knows the phenomenon of the Talbot effect and one can recognize the 'correct' image plane only from scratches, dust particles and edges or other deviations from the periodicity of the object (see figure 21-4) [21-18], [21-19].

It was also the problem of the afocal nature of interference phenomenon that inspired the investigation of three-dimensional image formation. Three-dimensional image formation is of course relevant whenever three-dimensional objects, such as biological objects, are under investigation, or when the three-dimensional image is under question, as in lithography. The expansion of Abbe's description of imaging to three dimensions was finally presented in the 1980s [21-20]. This three-dimen-

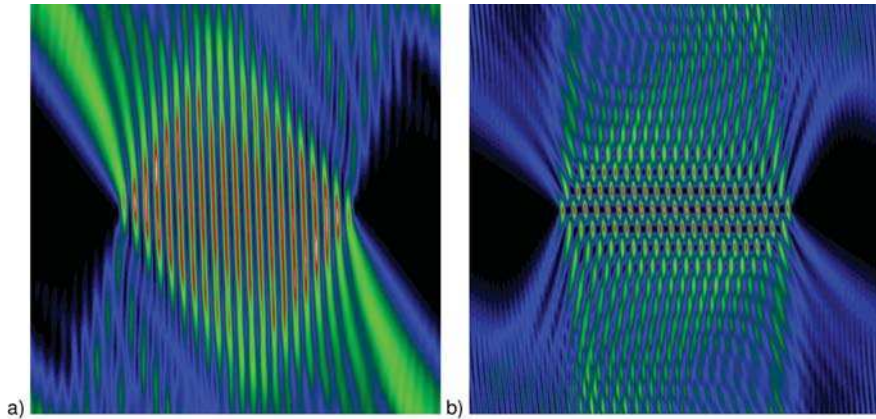


Figure 21-4: View through the focus aerial image plots, z -direction of optical axis is vertical: a) two-beam interference of waves of finite width, producing 21 interference fringes, depth of focus restricted only by the width of the

waves ($NA=1$, $d=0.8\lambda$, $\sigma=0.7$); b) three-beam interference, showing periodic pattern repetition (Talbot effect) in the vertical direction of the optical axis ($NA=1$, $d=1.2\lambda$, $\sigma=0.1$).

sional theory of optical imaging according to diffraction and interference provides the most comprehensive understanding of optical imaging and resolution. However, the first approach of Abbe is still the usual method for calculating the optical image.

The subject of this chapter is the description of optical image formation according to Abbe imaging theory. In section 21.1 it will be shown that the fundamentals of optical imaging can be visually interpreted without a complex mathematical description. In section 21.2 the mathematical description of optical imaging in Fourier steps will be given. Section 21.3 illustrates, for selected examples, the differences between perfectly incoherent and perfectly coherent imaging. In section 21.4, the role of partial coherent illumination is discussed.

The discussion is limited to the imaging of an object in an object plane into an image plane, and is restricted to the scalar treatment of the electromagnetic field. The three-dimensional description of image formation will follow in chapter 23.

21.2

Phenomenological Description of Imaging

21.2.1

The Explanation of Image Formation According to Abbe and the Abbe Resolution

Ernst Abbe found, during the development of his imaging theory, that the spectrum of the object is formed in the rear focal plane of a microscope's objective. In particular, if the object, e.g., a grating, is illuminated by a planar wave the diffraction spectrum of the object can be observed there. The single focus points in the rear focal

plane of the objective are converted by the following optical system, in this case the eyepiece and the eye, into planar waves, which superimpose in the image plane forming a striped interference pattern similar to the object. Therefore optical imaging of non-self-luminous objects is described by the diffraction at the object, cutting back the diffraction orders at the aperture and the image-forming interference of the diffraction orders. By arranging two Fourier transforming systems in a row the so-called $4f$ -system is obtained. Each imaging system can be represented by a $4f$ -system, consisting of a first $2f_1$ and a second $2f_2$ -system (figure 21-5).

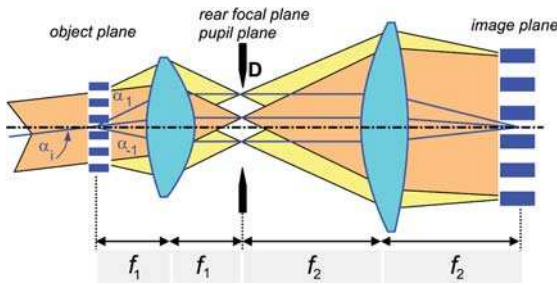


Figure 21-5: Interpretation of optical imaging by a $4f$ -system by diffraction and interference.

Ernst Abbe discovered in 1873 that, in order to obtain the object spectrum in the rear focal plane of an imaging lens, the microscope objective has to be corrected aplanatically. Subsequently this discovery was described in optics text books as the Abbe sine condition. According to the definition given by Abbe the aplanatic correction is achieved when the position of the diffraction orders x_p in the rear objective focal plane is related to the sine of the diffraction angle α (see figure 21-6). Only under this condition is the transverse spatial frequency proportional to the transverse position of the focal point x_p

$$x_p = f_1 \cdot n \sin \alpha = f_1 \cdot \lambda \cdot \nu_x . \quad (21-3)$$

This is the only way in which the proportionality between the pupil coordinate and the transverse spatial frequency is given. The pupil coordinate x_p is then proportional to the canonical optical direction cosine p_x , scaled by the focal length f_1 :

$$\frac{x_p}{f_1} = \lambda \cdot \nu_x = p_x = n \sin \alpha . \quad (21-4)$$

The Fourier plane of an aplanatically corrected optical system set-up is therefore equivalent to the canonical pupil (see chapter 20.2.4). Differently scaled pupil coordinates are used, either – as above – in genuine transverse co-ordinates x_p , as optical direction cosines p_x , p_y or – the most common way – normalized to the maximum aperture NA of the optical system.

In an analogous way a point-like object in the front focal plane – given by the origin of a spherical wave – generates a planar wave behind the lens. The propagation direction γ of the planar wave is given by the transverse coordinate x of the object point (figure 21-6) or

$$x = f_2 \cdot n \sin \gamma . \quad (21-5)$$

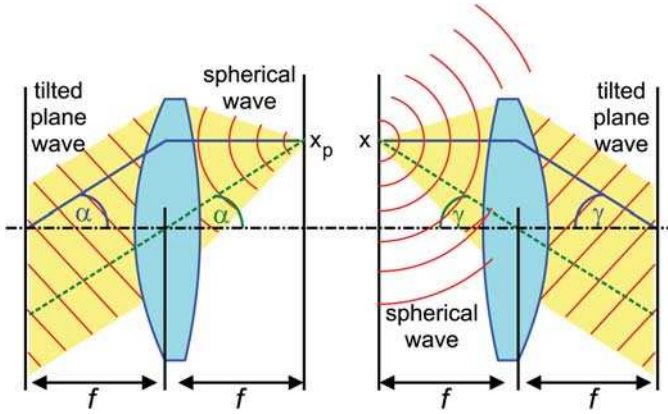


Figure 21-6: An incident plane wave with propagation angle α is focussed at the rear focal plane at x_p , a point object at the front focal plane produces a plane wave with propagation angle γ behind the ideal lens.

Using eq. (21-4) the positions of the m^{th} diffraction orders $x_{p,m}$ in the pupil plane are then given according to the diffraction equation by

$$\frac{x_{p,m}}{f_1} = n \sin \alpha_m = \left(m \frac{\lambda}{d} + n \sin \alpha_i \right) \quad (21-6)$$

with α_m as the diffraction angle of the m^{th} diffraction order and $\alpha_i = \alpha_0$ as the angle of the incident plane wave. In the image plane the planar waves with the angles α'_m or the transverse spatial frequencies ν'_m are given according to eq. (21-5) by

$$n \sin \alpha'_m = \lambda \nu'_m = \frac{x_{p,m}}{f_2} = \frac{f_1}{f_2} \left(m \frac{\lambda}{d} + n \sin \alpha_i \right) . \quad (21-7)$$

The minimum frequency distance $\Delta \nu$, and thus the basic pattern of the diffraction image, is given by interference of, e.g., the zeroth diffraction order with the frequency ν'_0 with one of the first diffraction orders with the frequency $\nu'_{\pm 1}$. For the period d' of the interference pattern it follows with eq. (21-7) that

$$d' = \frac{1}{\nu'_1 - \nu'_0} = \frac{\lambda}{n \sin \alpha'_1 - n \sin \alpha'_0} = \frac{f_2}{f_1} \cdot d = \beta \cdot d . \quad (21-8)$$

The image scale β of the 4f-system is thus given by the ratio of the two focal lengths f_2 and f_1 .

Considering Abbe's way of describing the optical imaging as a two-fold Fourier transformation, it becomes clear, by considering diffraction and interference, why a Fourier transforming optical system has to be corrected to obey the aplanatic condition. After all, the diffraction at the object grating follows the physical laws of diffraction stating that the sine of the diffraction angle is inversely proportional to the grat-

ing period ($\sin \alpha \sim 1/d$). Also, for the interference of two waves, the period of the generated interference pattern is proportional to the difference of the sine values of the propagation angles of the two interfering waves. Therefore, the sine of the diffraction angles rather than the diffraction angles themselves, have to be image linear. Since f_2 is very large in microscopes with large magnification β , the angles in the image space are very small with $\sin \alpha \approx \alpha$, and the aplanatic correction is of special importance for the microscope lens with f_1 . In general, the aplanatic correction is of increasing importance and is achieved with increasing difficulty, the smaller the focal length f of the Fourier-step of an imaging system or the larger the angle α becomes.

The minimum structure size that can still be resolved by optical imaging can be concluded from eq. (21-8). The smaller the object period, the larger is the diffraction angle. Because of the layout of the optical system, with a finite aperture diameter D it is, however:

$$x_p < \frac{D}{2} . \quad (21-9)$$

The maximum diffraction angle relative to the optical axis accepted from an optical imaging system, as a property of the imaging system, is called the numerical aperture NA with

$$NA = n \sin \alpha_{\max} = \frac{D}{2f_1} . \quad (21-10)$$

After inserting and converting eq. (21-8) for the minimum object period d_{\min} , which can still be resolved at illumination under the angle α_i the *Abbe fundamental equation of the optical resolution* follows:

$$d_{\min} = \frac{\lambda}{NA + n \sin \alpha_i} = \frac{\lambda}{NA(1 + \sigma)} . \quad (21-11)$$

To summarize, according to eq. (21-11) the resolution of the optical image is influenced by the four following factors.

- The *aperture angle* α_{\max} of the imaging system: the larger the aperture angle the smaller the object details that can be resolved. The sine of the aperture angle, however, cannot be greater than 1.
- The *wavelength*: for a shorter wavelength there are smaller diffraction angles which can be collected by the imaging system with a finite aperture angle. Already Abbe had suggested therefore to carry out high-resolution imaging with ultraviolet light, for instance, in microscopy [21-3], [21-21].
- The *index of refraction* of the environment: a higher index of refraction n reduces the diffraction angles as well, since the effective wavelength is reduced by n . The numerical aperture NA can – in contrast to the sine of the aperture angle – be greater than 1 with $n > 1$ (liquid or solid immersion [21-22]).
- the *illumination*: the resolution can be increased by oblique illumination with large illumination angles up to the aperture angle α_{\max} (figure 21-7).

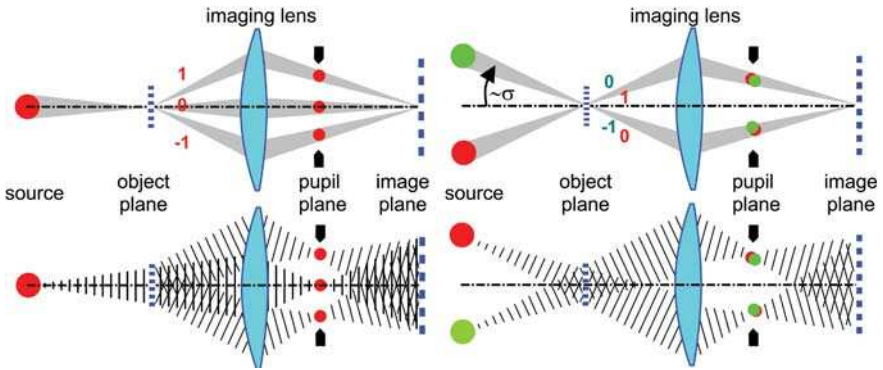


Figure 21-7: The maximum resolution is achieved at oblique illumination when two interfering diffraction orders pass the pupil at opposite edges.

According to Abbe, at the resolution limit for $\sigma = 1$ the normalized intensity is given by the interference pattern of two planar waves with the maximum frequency difference $\Delta\nu = 2NA/\lambda$ (with normalized coordinates $w = NA/\lambda \cdot x$):

$$I_{grid}(w) = 0.5 + 0.5 \cos(2\pi\Delta\nu \cdot x) = 0.5 + 0.5 \cos(4\pi w) . \quad (21-12)$$

Figure 21-8 illustrates the difference between the imaging of a single point and a grating at the resolution limit. The half-width of the ideal point image for rotational-symmetric imaging is approx. $w = 0.516$, while the peak-width of a structure of the interference pattern is only $w = 0.25$. The intensity fringe pattern, however, is not localized, i.e., it is distributed over the complete image plane. Nevertheless, the interpretation of optical imaging by diffraction and interference explains the (more than a factor of two) increased optical resolution compared with the Rayleigh limit.

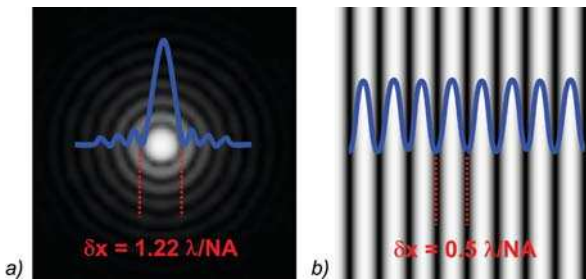


Figure 21-8: a) Ideal Airy Point image; b) 2-beam interference pattern.

In eq. (21-11) the obliquity factor $\sigma = n \sin \alpha_i / NA$ for the description of the maximum illumination angle is introduced [21-12], [21-23]. The resolution limit depends on the illumination and the imaging system, but also on the visibility resolution of the detectors or the light-sensitive medium, respectively. This is expressed in a generalized way by the k -factor of optical imaging introduced by M. Berek [21-12], [21-24], [21-7]. Accord-

ing to this the smallest structure size, in optical lithography called the *critical dimension CD*, and thus the resolution, are determined by the k -factor:

$$CD = k \frac{\lambda}{NA} \quad (21-13)$$

with

$$k = \frac{0.5}{1 + \sigma} . \quad (21-14)$$

The smaller the k -factor, the nearer the CD lies to the resolution limit. For $k < 0.5$ the limit of the coherent resolution is exceeded and single pairs of object points cannot be resolved separately. The minimum k -factor for the imaging of periodical objects lies at the resolution limit at $k = 0.25$.

Even for illumination apertures $n \sin \alpha_i > NA$ the obliquity factor is $\sigma \leq 1$. With illumination set-ups beyond the aperture angle $\sin \alpha_i > NA$ ($\sigma > 1$) no further increase in the resolution can be achieved. After all at least two homologous light-source images have always to pass the pupil in order to form an interference pattern. For $\sigma > 1$ the zeroth-order of the light source is cut back by the aperture of the optical system so that an interference has to be formed by higher diffraction orders, e.g., first and second diffraction orders. This is the case at dark field illumination, e.g., only illumination directions larger than the numerical aperture are selected in order to make small, possibly transparent, objects visible and not to swamp them by large background light, although no higher resolution is achieved by this.

Finally, all constituents for the description of the optical imaging by Fourier steps are available:

- 1) the object distribution is Fourier transformed to obtain the transverse spatial frequency spectrum;
- 2) the transverse spatial frequency spectrum, shifted by the illumination direction, is low-pass filtered by the pupil aperture;
- 3) the image results from the interference of the transmitted diffraction orders, i.e., the Fourier transformation of the frequency-limited object spectrum;
- 4) the process is repeated for all illumination directions.

That is why Abbe remarked in 1873: *it is possible to calculate the image of an arbitrary object with certain illumination in every detail*. The detailed exposition of the mathematical theory of optical imaging follows in section 21.3.

21.2.2

The Information About an Object Contained in an Image

Optical imaging is – in the most simple approximation – a linear operation, and an optical system can be interpreted as a low-pass filter. The intensity distribution generated by optical imaging is the interference pattern consisting of the diffraction

orders, which are able to pass the objective pupil. In this context Abbe had already discussed the question of the similarity of the image and object. In an experiment, Abbe filtered single diffraction orders by means of an aperture in the rear focal plane of the objective and observed the resulting effect on the image. He not only found a complete confirmation of his theory but had to realize again that, using optical imaging, only those details of an object can become visible to which at least two diffraction orders can pass the objective pupil. This means we can only know as much about an object as it has to have those object details, which correspond to the observable interference effects. The real object is, however, not observable.

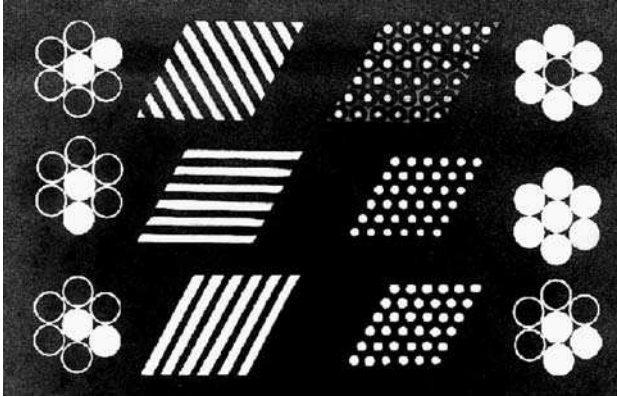


Figure 21-9: A reproduction of the imaging experiment according to Abbe from H. Siedentopf (1915). The circles represent the diffraction orders as they appear in the rear focal plane. Different diffraction orders of the hexagonal object structure contribute to different image details.

The experiment, which today is a popular part of practical training in science studies, was documented by Abbe's student H. Siedentopf for the example of the alga (diatom) *Pleurosigma angulatum* (figure 21-9, [21-25], [21-26]). Until the mid 20th century, diatoms were among the test objects which manufacturers of microscopes used for benchmarking the resolution of their products (see e.g. [21-9], [21-27]). For the experiment, Siedentopf selected a light source with the radius $\sigma \approx 1/3$, so that with the appropriately chosen aperture just three diffraction orders can pass the pupil without overlapping. From figure 21-9 it can be seen how the single diffraction orders in one line contribute to the formation of object details. Two diffraction orders always form only one grating in the direction perpendicular to the diffraction orders, whose Fourier transform is equal to a cosine function (figure 21-9, examples on the left side). The hexagonal structure is not formed until three diffraction orders, which are not arranged in a row, contribute (figure 21-9, example bottom right).

The optical image as an interference phenomenon is always different from the object according to the finite number of diffraction orders transferred by the imaging optical system. It is interesting to illustrate the number of transferred diffraction orders, depending on the size of the structure and the illumination directions. Figure 21-10 depicts imaging with a light source of radius $\sigma = 1$. The thick blue circles

represent the pupil with radius $f \cdot NA = f \cdot n \sin \alpha_{\max}$. The red points represent the diffraction orders corresponding to a certain axial point of the light source. The zeroth diffraction order, equivalent to the axial light source point, passes the pupil in the centre. The overlap regions of the diffraction orders with the pupil are shown in different colours. In the marked areas the number of diffraction orders contributing to the image formation is given. Certain illumination directions are marked in yellow. For each of these, only two diffraction orders consisting of the zeroth diffraction order, i.e., the illumination direction, and one of the first two diffraction orders in each illumination direction can pass the pupil as a border case of image formation. In figure 21-10b a case is shown in which the pupil allows just three diffraction orders to pass, i.e. no four-fold interference is possible. This corresponds to the object period

$$d_3 = \frac{3}{2} \frac{\lambda}{NA} . \quad (21-15)$$

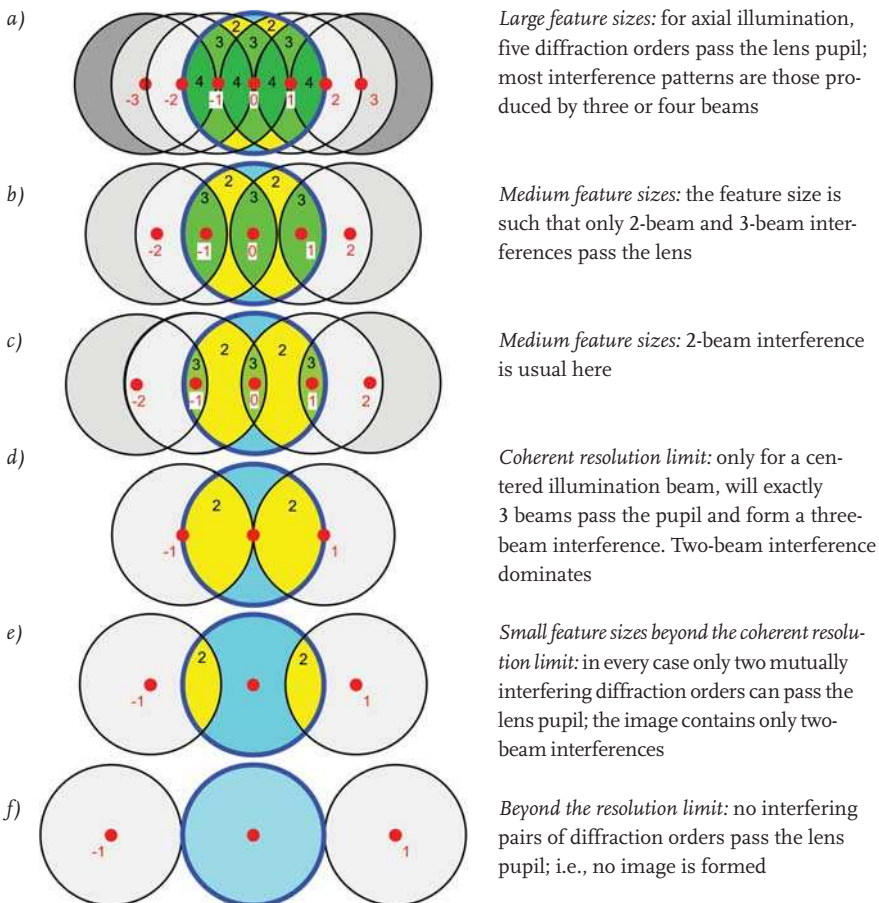


Figure 21-10: Image forming diffraction orders for different feature sizes.

At the *coherent resolution limit* just three diffraction orders can pass the pupil (figure 21-10d). The conditions are then fulfilled for an axial illumination with a perpendicular incident planar wave to just result in an image. For $\sigma = 0$ from eq. (21-11) the coherent resolution limit follows as:

$$d_2 = \frac{\lambda}{NA} . \quad (21-16)$$

For object periods smaller than the coherent resolution limit only two-beam interferences contribute to image formation. An image resolving these details can only be achieved using oblique illumination, i.e., with $\sigma > 0$ (figure 21-10e). Finally, the resolution limit is reached at $\sigma = 1$ when two diffraction orders pass the pupil at opposite edges (figure 21-10f).

It can be seen from figure 21-10, that the formation of image details as an interference phenomenon depends on the illumination direction. Different illumination directions cause interference phenomena with different numbers of interfering diffraction orders. So the resulting image and thus the similarity to the object depend on the choice of illumination! As will be shown below in more detail, the diffraction orders forming the image can be selected by the illumination directions. Therefore it is necessary to give considerable thought to the illumination when the image is investigated [21-26].

21.2.3

Koehler Illumination and the Visibility

The discussion above has shown that optical imaging can be described very easily if the illumination is considered as consisting of plane waves. Points in the Fourier plane and the objective pupil, respectively, correspond to the plane waves in the object plane. August Koehler, student and co-worker of Ernst Abbe, therefore suggested to image the physical light source (sun, light bulb or electrical arc) into the pupil. For this an additional Fourier step, the Abbe condenser, has been added in front of the object plane. Then, in the front focal plane of the condenser, there is usually an image of the light source, the extension and form of which can be influenced by a condenser aperture. Figure 21-11 illustrates the typical layout of a microscope with Koehler illumination.

In modern illumination systems the light source is seldom imaged directly into the condenser pupil. Rather, an effective light source is generated in a Fourier plane relative to the object plane by mixing systems like integrators (e.g. fly's-eye integrators or rod-integrators). Only the image of the effective light source lies in the pupil. So Koehler illumination in reality describes the imaginary reduction of the illumination to planar waves, with the angular spectrum of the illuminating plane waves given by the form and the extension of the effective light source. For this it is assumed that all points in the light source are perfectly incoherent to each other and thus the intensities of the assigned single images can be added to each other. From the effective light source to the image plane, any imaging system can thus be illustrated as a $6f$ -system, with the illumination condenser as the first $2f$ -step (figure 21-12).

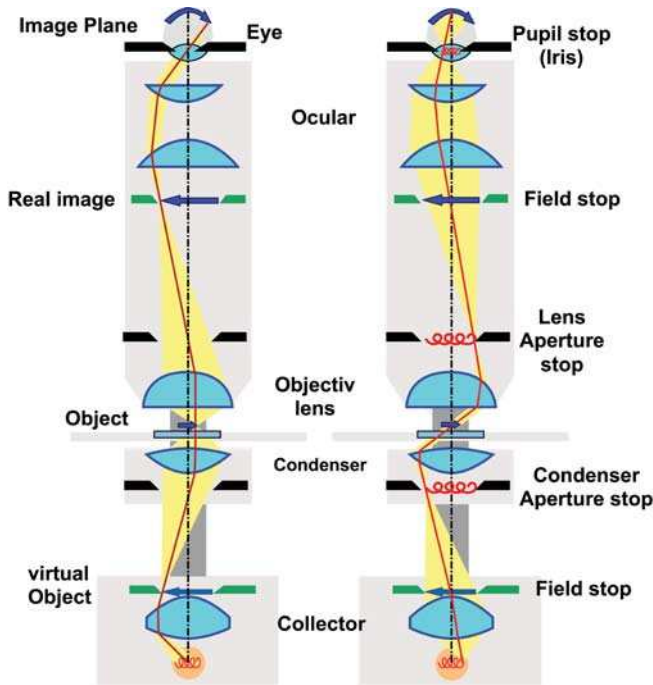


Figure 21-11: Classical layout of a microscope with Koehler illumination; left: the imaging beam path, right: the illumination beam path.

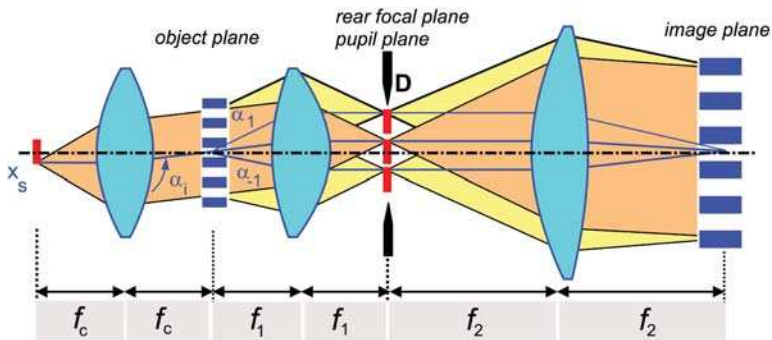


Figure 21-12: Koehler illumination with the Abbe condenser of focal length f_c giving a $6f$ -imaging system.

From the diffraction at the object, homologous source images are generated from an effective source in the pupil of the objective. The position of the source images is determined in accordance with the diffraction orders (figure 21-12). As is already known from Fresnel's interference experiments, light bunches from different light source points of a thermal light source, cannot interfere with each other since they do not have a constant phase relation with each other (see chapter 19). So imaging

interferences are only formed by diffraction orders which are generated by illumination of the object from the same source point. In figure 21-13, following [21-21], three light source images are shown in the pupil in the direction of view perpendicular to the optical axis. The light source images have a radius $\sigma \cdot NA$ and are positioned in such a way that the first diffraction orders $C_{\pm 1}$ for the central light source point C (illumination direction parallel to optical axis, $\sigma=0$) can just pass the pupil in order to form an interference pattern in the image. For each of the source points A and D there are only two diffraction orders in the pupil (A_0 and A_1 , as well as D_0 and D_{-1}). For the light source point B only the zeroth diffraction order B_0 can pass the pupil so that there is no interference partner for this light source point. Light source points for which only one diffraction order can pass the pupil reduce the visibility by superposition of the non-structured intensity (*false light*).

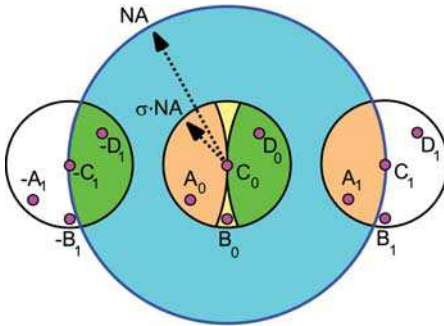


Figure 21-13: Illustration of image formation according to Koehler (figure 942 of [21-21]).

The visibility is of course also reduced by diffraction and by aberrations. More about this will follow. According to Michelson the visibility V is defined by

$$V = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} . \tag{21-17}$$

Neglecting the three-beam interference of figure 21-13 due to the diffraction orders C_{-1} , C_1 and C_0 , the image is composed only of two-beam interferences, and the image intensity can be described by

$$I(x) = a + b \cdot \cos\left(\frac{2\pi}{d} x\right) . \tag{21-18}$$

The visibility V or contrast C according to eq. (21-17) is thus given by

$$V = \frac{b}{a} . \tag{21-19}$$

The contrast loss, on the other hand, is given by $\Delta V = 1 - V$, i.e.

$$\Delta V = 1 - V = \frac{a - b}{a} . \tag{21-20}$$

Figure 21-14 gives some schematic examples with increasing amounts of offset light $I_{\min} = a - b$.

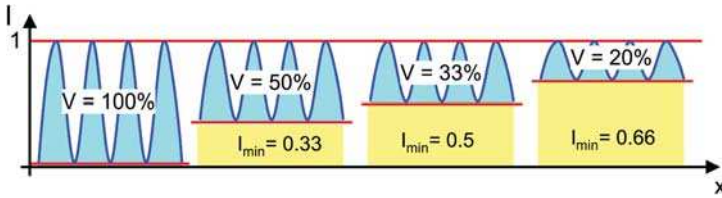


Figure 21-14: Contrast loss according to the amount of offset light I_{\min} .

In figure 21-15 the visibility is shown for a light source with radius $\sigma = 1$ for different grating periods. This description of the imaging system is also called the modulation transfer function MTF. The diffraction angle, and thus the distance of the light source images in the pupil increases with decreasing object detail size or object periods, leading to an increase in the share of the false light and, as a result of this, the visibility is reduced. So the visibility is mainly determined by the share of the light which can pass the objective pupil without an interference partner. Unfortunately, this simple interpretation has often been missing in the Fourier-theoretical description of imaging.

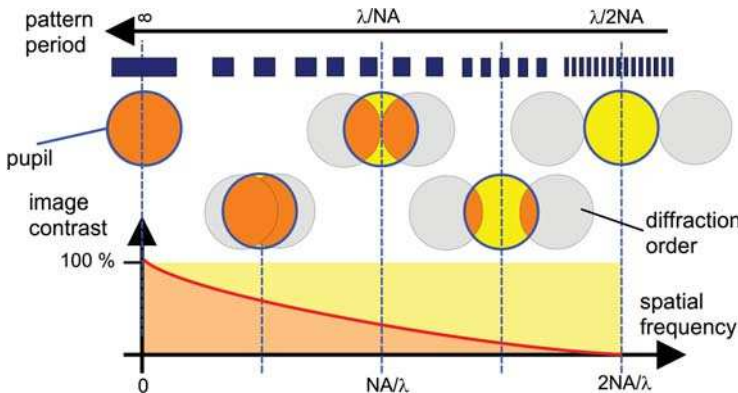


Figure 21-15: Modulation transfer function MTF: Contrast loss is proportional to the zero-order light passing the lens aperture without an interference partner, which only contributes to a constant offset.

21.2.4

The Siedentopf Illumination Principle

This insight into the reason for the reduced visibility for finer structures, i.e., a recognition of the relevance of the false light share, led H.Siedentopf to a simple general rule for ideal illumination [21-25]. He varied the objective aperture at the constant light source diameter and marked the homologous, i.e., interferable shares of

the diffraction orders of the hexagonal structure of *Pleurosigma angulatum* in different colours (see figure 21-16 from the original work by H. Siedentopf). In figure 21-16, c,d and e there are six groups of four homologous light source images, each of which form a four-fold interference (coloured blue; one of the groups is marked in dark blue). For a small aperture, i.e., at the resolution limit, however, there is a central share of the light source for which there is no interference partner:

In the Figures 1a and 1b no homologous parts of the side images (of the light source) exist in the central part of the aperture image of the lens. These blank parts therefore cannot produce an image, and act like false light, which shines upon the images produced by the groups of two and three orders. In such cases it would therefore be very advantageous to apply a central obscuration to the condenser aperture to make the central part inoperative [21-25].

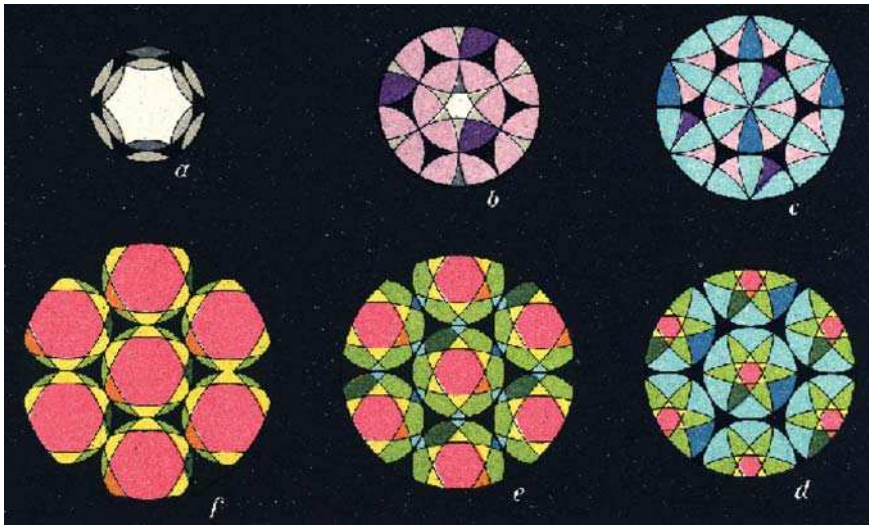


Figure 21-16: Reproduction of figure 1 from Siedentopfs original paper from 1915.

In a generalized manner, the Siedentopf principle of ideal illumination can be formulated as follows: Illuminate only from those directions which contribute to the image formation of the desired image detail. This principle has been used frequently in microscopy by using the exchangeable Abbe condenser aperture or other optical means to produce structured illumination, e.g., annular illumination, off-axis, monopole and dipole illumination [21-11], [21-26], [21-28]. However, since in many cases microscopic objects are not known in advance, often the selection has to be done by trial-and-error until the optimum visibility has been found. In general, one sorts out the overlap regions of the diffraction orders with the pupil and adjusts the illumination to those diffraction orders whose assigned image details are to be imaged with a high contrast, and suppresses the illumination from those directions which would lead to contrast loss or formation of undesired image details.

On the other hand, for the imaging of known periodical objects like gratings, the use of the Siedentopf principle works particularly well. Considering figure 21-16, it can be seen that the overlap regions of the diffraction orders and thus the areas for a structured illumination to emphasize image details, are particularly simple for gratings. The one-side off-axis illumination and dipole illumination were also suggested in order to emphasize certain structure orientations of an image using a higher imaging contrast [21-29], [21-30]. The Siedentopf principle has also been used by H. H. Hopkins for the imaging of gratings beyond the coherent resolution limit with periods of $d < \lambda/NA$, when he realized that only a part of the selected annular light source contributes to the image formation and thus this ought to be cut back by a double-side pair of condenser apertures for ideal visibility (figure 21-17; see also [21-17], Figure 3 and eq. (50)). For a certain object period, both plus and minus first diffraction orders only partially pass the pupil, such that there is an increasing share of false light from the zeroth diffraction with decreasing object period. This amount of false light is reduced by a condenser stop in order to shape the light source to form two segments of a circle.

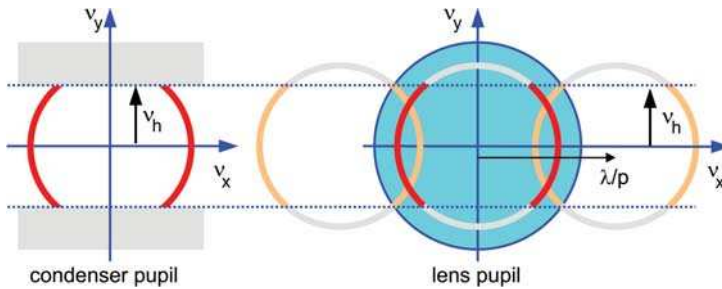


Figure 21-17a: Hopkins-Dipol illumination with $\nu_y < \nu_h$ as an application of the Siedentopf principle to the imaging of a grating with period d (after [21-17]).

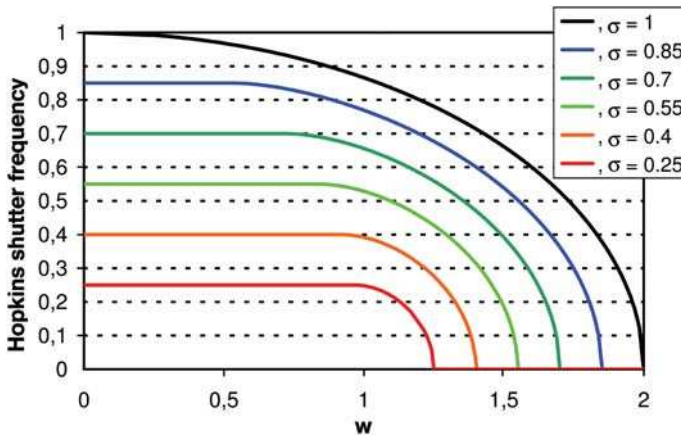


Figure 21-17b: Hopkins shutter frequency ν_h as a function of w for different illumination apertures σNA .

In figure 21-17b the Hopkins shutter frequency ν_h is shown in units of the numerical aperture NA over the normalized object frequency w with $w = \lambda/(d \cdot NA)$ for different σ . The frequency ν_h is obtained by a simple calculation of the intercept of both circles shown in figure 21-17a. The Hopkins aperture, however, can no longer be used below the normalized object period of $w = 1/\sqrt{1 - \sigma^2}$. From figure 21-17b it can be seen that for large σ -values the Hopkins dipole illumination offers an advantage to imaging even at frequencies below the coherent resolution limit ($w < 1$). With this effective dipole illumination the visibility is increased resulting at the same time in the formation of an afocal interference pattern by the two-beam interference [21-17]. For fine gratings beyond the coherent resolution limit, for which only two-beam interferences contribute to image formation, the dipole illumination is thus the direct consequence of applying the Siedentopf principle (figure 21-18).

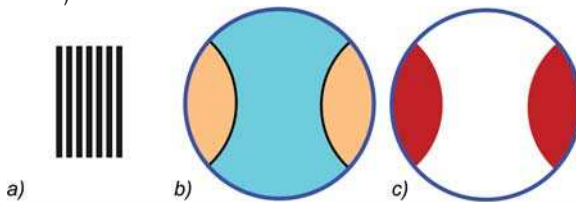


Figure 21-18: a) Linear grating; b) maximum possible overlap of the diffraction orders for grating periods beyond the coherent resolution limit; c) ideal illumination directions according to Siedentopf.

For objects consisting of gratings with different orientations, the Siedentopf principle can also be applied if the diffraction orders of the differently oriented grating structures overlap. An example is shown in figure 21-19. However, the Siedentopf principle cannot always be fulfilled for all existing object orientations. In figure 21-20 the case is shown where the diffraction orders for different object details do not overlap. Thus no common illumination directions can be found. Nevertheless, even for this case, the structured illumination according to Siedentopf is the best compromise for a high contrast image since the false light share is reduced to a minimum.

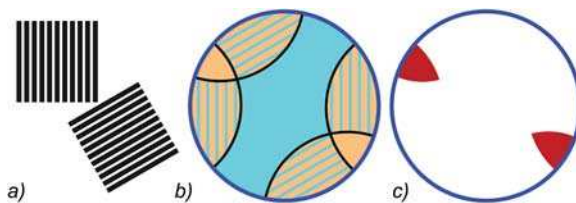


Figure 21-19: a) Differently oriented structures with overlapping diffraction orders (b); and c) illumination directions according to the Siedentopf principle. Maximum contrast is achieved when there is illumination only from the overlap regions of the diffraction orders for different object details and orientation, respectively.

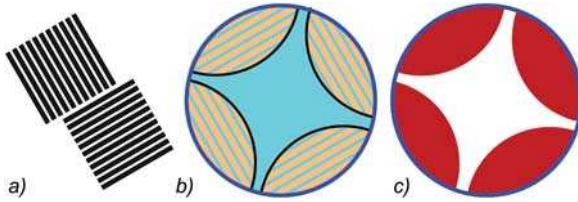


Figure 21-20: a) Differently oriented structures whose diffraction orders do not overlap (b); and c) illumination directions according to the Siedentopf principle. Maximum contrast cannot be achieved since the diffraction orders for different object details and orientations do not overlap.

The Siedentopf principle for structured illumination is applicable to all types of imaging, from microscopy of biological samples to microscopic inspection and micro-lithographic manufacturing of semiconductor structures. For the latter applications, it can be used in a particularly simple way since the objects to be imaged and the desired images are known in advance [21-31], [21-32], [21-33], [21-34]. Semiconductor structures comprise gratings in both horizontal and vertical directions. Figures 21-20 and 21-21 show examples of the illumination of objects consisting only of horizontal and vertical structures. In the case of no overlap of the diffraction orders, as shown in figure 21-20, ideal contrast cannot be achieved for both orientations, therefore a double exposure technique has been proposed [21-35].

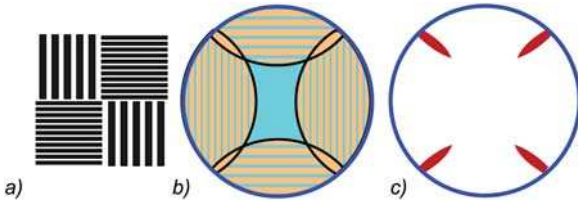


Figure 21-21: a) Horizontal and vertical structure of a semiconductor mask; b) overlapping diffraction orders; and c) ideal illumination directions according to the Siedentopf principle.

21.2.5

Imaging with Different Colours

The angles of the diffraction orders are given by the diffraction equation with $\sin\gamma = \lambda/d$. In the rear focal plane of a lens with focal length f the diffraction orders are at the coordinates x_p according to eq. (21-6)

$$\frac{x_{p,m}}{f_1} = n \sin \alpha_m = m \frac{\lambda}{d} + n \sin \alpha_i \quad (21-21)$$

With perpendicular incidence ($\alpha_i = 0$) the diffraction order position $x_{p,m}$ is simply proportional to the wavelength λ :

$$\frac{x_{p,m}}{f_1} = \lambda \cdot \frac{m}{d} \quad (21-22)$$

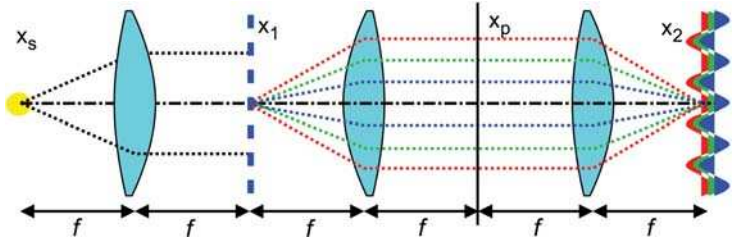


Figure 21-22: Generation of white interference fringes using a diffraction grating.

As illustrated in figure 21-22 and figure 21-23, with increasing wavelength, the diffraction order coordinate at the pupil plane increases. Therefore the resolution decreases with increasing wavelength. However, the image period d' according eq. (21-8) is independent of the wavelength. With decreasing object period, the larger wavelengths cannot contribute to the interference pattern of the image and the ultimate image is obtained from the shortest wavelength. For conventional illumination with white light, the finest image details are blue in color.

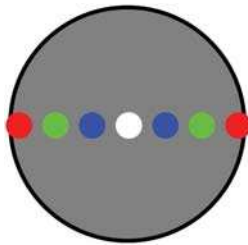


Figure 21-23: Diffraction order position in the pupil plane depends on wavelength.

21.2.6

Aplanatic Correction and Geometrical Optics

Assuming a lens in the sense of mathematical imaging, according to axial-symmetric collineation (see figure 21-24) with planar principal planes, the following relation is obtained

$$x'_p = f \cdot \tan \alpha \quad (21-23)$$

The paraxial correlation eq. (21-23) is given in several text books without detailed explanation, which can lead to misunderstanding. It means that a lens with a planar principal or unit planes carries out a Fourier transformation only in the paraxial regime where $\tan \alpha \approx \sin \alpha$ (figure 21-24). Furthermore, it is frequently shown in an arithmetic way, with the help of the (paraxial) Fresnel transformation, that there is a Fourier correlation between both focal planes of a single thin lens (see, e.g., chapter 12, Vol. 1 or [21-8], [21-36]).

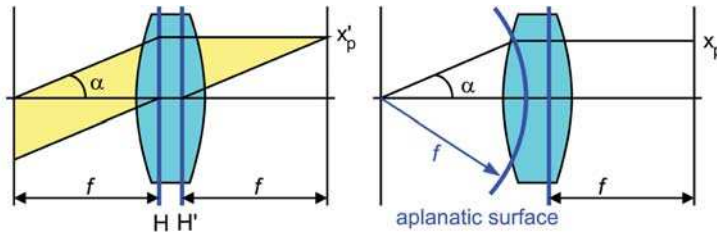


Figure 21-24: Mathematical mapping and ideal optical imaging with spherical principal plane: the aplanatic plane.

However, the aplanatic correction is equivalent to a system with a spherical principal plane, the aplanatic surface. The intersecting point of a light ray going out from an axial point with the aplanatic plane is thus equivalent to the pupil coordinate x_p according to eq. (21-3), which differs from the Ewald-sphere coordinate in the spatial frequency space only by the scaling factor $f \cdot \lambda$. The principal plane as a system feature depends on the correction of the optical system. That is why the aplanatic correction can be exactly fulfilled just for one point – the axial point; in real systems with finite object field it can be fulfilled only approximately.

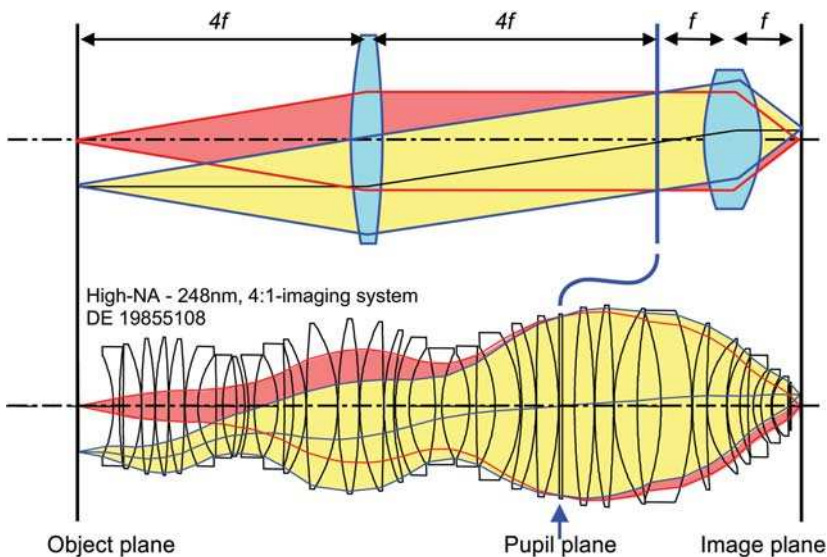


Figure 21-25: Example of a lithography lens with a rear aperture $NA=0.7$; at the top, a schematic description as a $4f$ -system is given.

The description of optical systems for carrying out a Fourier transformation with just one lens of focal length f is an extreme simplification. For real systems there has to be an arrangement of multiple lenses to achieve a sufficient correction, particularly the aplanatic correction. The example of a projection lens for optical lithography in figure 21-25 illustrates the complexity of an almost aberration-free optical system when large

fields and high apertures are involved [21-37]. Due to the position of the principal planes of the two Fourier steps, the distance ratios do not match those of the simplified $4f$ set-up.

In addition to the aplanatic correction, it is also necessary for the application of the linear system theory (that will be described in the following chapters) that the imaging conditions are invariant at least over small regions of the object or field plane (isoplanatic condition). The isoplanatic condition is usually fulfilled for microscope objectives with a higher aperture for a small image field only, but – as shown in figure 21-25 – it can be achieved even for large image fields with more than 30 mm diameter by using many lenses and, for example, aspherical surfaces.

The inequality of equations (21-3) and (21-23) is based on the fundamental differences and the incompatibility of mathematical mapping and physical imaging. In practice it can be accomplished by using several lenses to achieve a sufficiently aplanatic correction for a finite object field even with aspherical surfaces or diffractive optical elements, for example. The terms which are valid in the paraxial area of the mathematical mapping, as well as Listing's construction, however, lose their validity in this case. For Listing's construction and the terms of geometric optics the reader is referred to the text books about geometric optics, e.g. [21-38] or volumes 1 and 3 of this series.

Finally we comment on the spherical principal plane, the aplanatic surface of a Fourier optical system. At the beginning of the 19th century H. Siedentopf had already used the aplanatic surface as a spherical surface to illustrate the case of imaging with oblique illumination [21-39]. Later the model of the Ewald sphere, which had been developed for the description of crystal diffraction, had been transferred to optics and so the analogy between an aplanatic surface and the Ewald sphere was established for infinity corrected systems.

21.3

The Mathematical Description of Fourier Optical Imaging

21.3.1

Imaging with Uncorrelated Light Sources

In this section, optical imaging is described by Fourier steps: the theory of image formation of a planar object in an image plane is described by Fourier-transformations and a linear filter operation in the pupil plane. The result of the incoherent image formation is in general complex, non-linear dependency of the image intensity on the object and light source, which is to be solved numerically. Meanwhile, due to the advance in computers, the numerical solution of the partial coherent image formation is standard for many commercial optics designs and analysis software, and also is an easy task to implement. Before computers were available, a formal mathematical theory of partial coherent image formation was developed, which is interesting for the comprehension and interpretation of optical imaging. This mathematical theory of optical imaging is described in chapter 22.

Here the conventional Fourier-optical description of optical imaging will be illustrated for the Fourier-transforming property of a lens. Optical imaging systems are frequently described by $4f$ -systems, comprising first a $2f$ -system from object plane to pupil plane, and second a $2f$ -system from pupil plane to image plane. Some textbooks derive the Fourier-transforming property of a lens within the scope of the paraxial approximation, but optical imaging is not limited to the paraxial region. The description of optical imaging by the $4f$ -system is only a simplified symbolic illustration and should not be overemphasized. Optical imaging is a linear operation, which is – as for many operations in optics – better described in the frequency domain. Therefore, the field distribution in the object space is expanded in its spatial frequency spectrum. The spatial frequency spectrum in image space is given by multiplication of the frequency spectrum in object space by a complex transmission function – which is, in the most simple case, given by a low-pass filter. This generalized description of optical imaging as a filter operation in frequency space will be discussed further in chapter 23.

The description here is given for telecentric imaging. For non-telecentric imaging, an equivalent method can be found. In the following, the formal description of the imaging is further limited to a one-dimensional representation with the imaging scale 1:1. An extension to two-dimensional objects and arbitrary magnifications is straightforward.

In partial coherent imaging, it is necessary to give a higher importance to the light source because the form and size of the light source have a significant influence on the imaging. For a consequent Fourier-optical description it is advantageous to assume an effective light source, which is in a Fourier plane in reference to the object plane (Köhler illumination). In addition to the $4f$ set-up for the imaging, a Fourier step is achieved again and thus a $6f$ set-up (figure 21-26). This approach of the so-called Köhler illumination can be applied to almost all relevant cases, even if the physical light source does not necessarily lay in a Fourier-inverse plane in front of the field plane. Because of this the light source is also called the effective light source $s(\nu_s)$. Only for special cases of coherent illumination is the approach of the effective light source no longer an appropriate description.

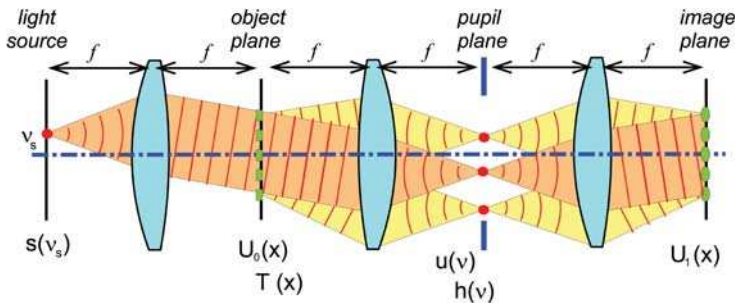


Figure 21-26: Description of optical imaging in Fourier steps.

Starting from the effective light source $s(\nu_s)$ in the object plane, an illumination wave is obtained as the Fourier transform of the light source

$$S(x) = \int s(\nu_s) \cdot e^{2\pi i \nu_s x} d\nu_s . \quad (21-24)$$

The object arranged in the object plane is described for the most simple case of the approximation of a thin element (thin element approximation TEA) by its complex transfer function

$$T(x) = A(x) \cdot e^{i\varphi(x)} \quad (21-25)$$

and multiplied with the incident illumination field distribution $S(x)$ in order to obtain the object wave or the field distribution directly behind the object:

$$U_0(x) = T(x) \cdot S(x) . \quad (21-26)$$

In the pupil plane there is then the Fourier transform of $U_0(x)$. Since the Fourier transform of a product of two functions is given by the convolution of the spectra of the functions, the field distribution u_0 in the pupil plane is given by a convolution of the effective light source with the object spectrum:

$$u_0(\nu) = F\{U_0(x)\} = t(\nu) \otimes s(\nu) . \quad (21-27)$$

Now, in the pupil plane the filtering effect of the aperture stop and the optical system aberrations, generally described by a complex transmission function, takes place. This is called the *coherent transfer function CTF*:

$$h(\nu) = P(\nu) \cdot e^{i\frac{2\pi}{\lambda}W(\nu)} \quad (21-28)$$

with the *pupil function* $P(\nu)$ and the *wave front aberration* $W(\nu)$. Corresponding to the finite wavelength and the limited aperture of an optical system, the optical imaging is always low-pass filtering. This is expressed by the pupil function P which differs from zero only in a finite range. In the most simple case the pupil function is equal to a rect- or circ-function, respectively (the definition of these functions is given in the mathematical appendix). In addition to this, imaging aberrations may influence the optical imaging. This is expressed by the wave front aberration W . As shown in chapter 20, the wave front error $W(\nu)$ in eq. (21-28) corresponds to Hamilton's mixed characteristic function $W(x_0, p_1)$ and can be determined, e.g., by ray-tracing.

Thus the field distribution behind the aperture is the product of the field distribution according to eq. (21-27) times the filter function $h(\nu)$:

$$u_1(\nu) = h(\nu) \cdot u_0(\nu) = h(\nu) \cdot [t(\nu) \otimes s(\nu)] . \quad (21-29)$$

The field distribution in the image plane is obtained by a repeated Fourier transformation:

$$U_1(x) = F\{h(\nu)\} \otimes [T(x) \cdot S(x)] = H(x) \otimes U_0(x) . \quad (21-30)$$

As can be seen, now the object wave U_0 is convoluted with the Fourier transform of the filter function $h(\nu)$. The Fourier transform of the filter function is also called the *amplitude distribution function* $H(x)$:

$$H(x) = \int h(\nu) \cdot e^{2\pi i \nu x} d\nu = \int P(\nu) \cdot e^{\frac{i2\pi}{\lambda} W(\nu)} e^{2\pi i \nu x} d\nu . \quad (21-31)$$

With the coherent optical imaging the complex amplitude is therefore linearly transferred. The effect of the filter $h(\nu)$ in the frequency space is expressed in the image plane by a convolution of the object wave $U_0(x)$ with the amplitude distribution function $H(x)$, which can be written as

$$U_1(x) = \int U_0(x') \cdot H(x - x') dx' . \quad (21-32)$$

However, the detection process can only measure the image intensity. In the most simple approximation this is given by the squared value of the amplitude:

$$I_1(x) = |U_1(x)|^2 = |H(x) \otimes U_0(x)|^2 . \quad (21-33)$$

For coherent imaging there is no linear correlation between the image intensity I_1 and the object intensity $|U_0(x)|^2$!

In order to further investigate the influence of the light source on the image intensity, a single delta-like light source point p at coordinate ν_p is considered, and $S(x)$ in eq. (21-26) is given by a plane wave. After insertion into (21-32) we obtain

$$U_p(x, \nu_p) = \int T(x') \cdot H(x - x') \cdot e^{i2\pi \nu_p x'} dx' \quad (21-34)$$

for the normalized image amplitude U_p for a light source point p . The image intensity is given by the squared value of the integral over all image amplitudes to all light source points:

$$I(x) = \left| \int s(\nu_p) \cdot U_p(x, \nu_p) d\nu_p \right|^2 . \quad (21-35)$$

After insertion of eq. (21-34) one formally gets a quadruple integral for the image intensity consisting of a two-fold convolution and the double integral over all light source points:

$$I(x) = \int s(\nu_p) \cdot \int T(x_1) \cdot H(x - x_1) \cdot e^{i2\pi \nu_p x_1} dx_1 d\nu_p \cdot \int s^*(\nu_q) \cdot \int T^*(x_2) \cdot H^*(x - x_2) \cdot e^{-i2\pi \nu_q x_2} dx_2 d\nu_q . \quad (21-36)$$

In the double light source integral in eq. (21-36) the mutual *coherence function of the light source* Γ_S at the position of the object can be found, which is also called the mutual intensity of the illumination distribution:

$$\Gamma_S(x_1, x_2) = \iint s(\nu_p) \cdot s^*(\nu_q) e^{i2\pi(\nu_p x_1 - \nu_q x_2)} d\nu_p d\nu_q = S(x_1) \cdot S^*(x_2) . \quad (21-37)$$

After insertion of the mutual coherence function of the light source one obtains for the image intensity from eq. (21-36):

$$I_1(x) = \iint \Gamma_S(x_1, x_2) \cdot T(x_1) \cdot T^*(x_2) \cdot H(x - x_1)H^*(x - x_2)dx_1dx_2 . \tag{21-38}$$

In the normal case, effective light sources are considered which are completely uncorrelated, similar to thermal light sources. For uncorrelated light sources interference contributions of different light source points vanish due to the averaging in time of the intensity formation. Only those interference terms that come from one and the same homologous source point, and thus have a rigid phase correlation to each other, will remain. Under this assumption it follows for eq. (21-37) that the coherence function is given by the Fourier transform of the intensity distribution of the effective light source:

$$\begin{aligned} \Gamma_S(x_1, x_2) &= \iint s(\nu_p) \cdot s^*(\nu_q) \cdot \delta(\nu_p - \nu_q) \cdot e^{i2\pi(\nu_p x_1 - \nu_q x_2)} d\nu_p d\nu_q \\ &= \int |s(\nu_q)|^2 \cdot e^{i2\pi\nu_q(x_1 - x_2)} d\nu_q . \end{aligned} \tag{21-39}$$

The calculation of the intensity consequently becomes more simple and is described now by a double integral only:

$$I(x) = \int |s(\nu_q)|^2 \cdot \left| \int T(x') \cdot H(x - x') \cdot e^{i2\pi\nu_q x'} dx' \right|^2 d\nu_q . \tag{21-40}$$

The image formation for the optical imaging with Köhler illumination can be visualized as an incoherent superposition of all intensities to the coherent partial images to all light source points. Figure 21-27 illustrates this approach.

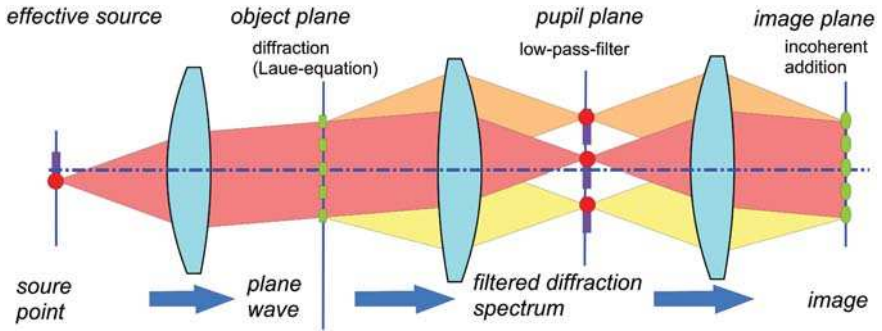


Figure 21-27: Image formation in Fourier steps of a light source point for a grating.

Depending on the extension of the light source, the imaging is called incoherent, coherent or partially coherent. If the direct image of the light source in the pupil plane is larger than the diameter of the aperture, it is always an incoherent image. For a point-like light source, the object is illuminated only by a plane wave. It is then always called a coherent image. For light source dimensions in between, it is called a partially coherent image. The two special cases of completely coherent imaging and completely incoherent imaging, as well as the general case of partially coherent imaging, are discussed in the next sections and chapter 24. Partial coherent imag-

ing, according to eq. (21-40), is frequently solved numerically by a sum over a finite number of light source points. Simulation results will be shown in chapter 24. The scalar description of the image formation derived so far contains some approximations, which in a numerical simulation can be replaced by exact calculations:

- The approximate consideration of the object by a transfer function can be replaced by so-called rigorous methods. For this the spectrum $t(\nu)$ has to be replaced by a diffraction spectrum $t'(\nu, \nu_s)$ correspondingly calculated, which can be a function of the illumination angle and the polarization.
- Besides wave front errors, in $h(\nu)$ apodization effects and deviations of the pupil function $P(\nu)$ from the ideal transmission caused by, e.g., anti-reflection coatings and absorption in bulk material can also be taken into account. The filter function can depend on the polarization.
- Some energy normalization factors can be taken into account in the pupil function.
- Polarization can be considered. Dealing with the change in polarization is more difficult. It is discussed in chapter 26 and the following section about polarization-optical imaging.
- Finally, modern software for the simulation of image formation also takes into account the effects of the light-sensitive medium, e.g., for lithography, the multi-reflections at the resist layers (e.g. [21-40]).

Finally, the reason for the choice of the effective light source in a conjugated plane to the pupil plane now becomes clear: because of the description of the illumination by plane waves incoherent to each other eq. (21-40) takes the simple form. For illumination with a multi-mode laser, however, a significantly more complex expansion in eq. (21-40) has to be accepted, as will be shown in chapter 21.4.6.

21.3.2

Consideration of Magnification

So far only imaging with a 1:1 magnification has been discussed. Magnification means only a linear scaling in the space domain from the object to the image plane as well as in the frequency domain, i.e., from the entrance to the exit pupil. In most cases magnification is thus considered in a proper image coordinate scaling. It is common to carry out all calculations only in, e.g., the image space, so that everything is scaled in image coordinates.

For consideration of the magnification, the field distribution behind the object is described by the object amplitude function $U_0(x_0)$ in coordinates of the object space x_0 and the field distribution of the image by the image amplitude function $U_1(x_1)$ in image space coordinates x_1 . The image function $U_1(x_1)$ should emerge from the object function $U_0(x_0)$ using a linear mapping, it is:

$$x_1 = \beta x_0 . \quad (21-41)$$

From the scaling properties of the Fourier transformation, the linear image of both spectra follows automatically. These are given by

$$u_0(\nu_0) = \int_{-\infty}^{\infty} U_0(x_0) e^{-2\pi i x_0 \nu_0} dx_0 \tag{21-42}$$

and

$$u_1(\nu_1) = \int_{-\infty}^{\infty} U_1(x_1) e^{-2\pi i x_1 \nu_1} dx_1 . \tag{21-43}$$

For the scaling of the spectra, the following relationship therefore has to be valid:

$$\nu_1 = \frac{1}{\beta} \nu_0 . \tag{21-44}$$

After a conversion from eq. (21-44) the *Abbe sine condition* is obtained:

$$\beta = \frac{\nu_0}{\nu_1} = \frac{\sin \alpha_0}{\sin \alpha_1} . \tag{21-45}$$

Equations (21-41) and (21-45) now offer different possibilities of taking the magnification into account for the Fourier-optical description of the image (figure 21-28).

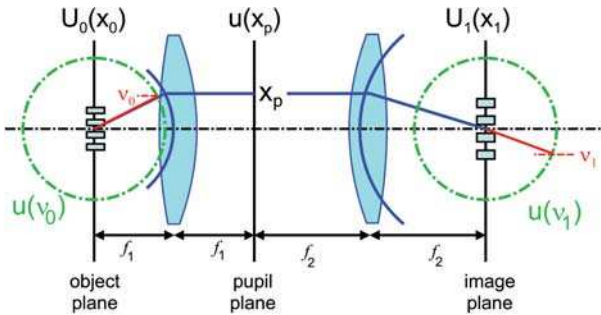


Figure 21-28: Two possibilities for taking the magnification into account: using the focal lengths for the unit planes or using a linear mapping in the frequency domain.

In the first case the pupil coordinate x_p is set proportionally to the focal length f_1 with $x_p = f_1 \sin \alpha_0$. The angle α_1 in the image space then follows from $\sin \alpha_1 = x_p / f_2$. For the mathematical treatment the second method is preferable. Here the magnification is taken into account in the transverse frequency domain by a linear mapping according to the Abbe sine condition (aplanatic correction). The magnification then causes a re-scaling during the inverse transformation of the filtered spectrum from the frequency domain. Here it should be mentioned that performing the Fourier transformation twice delivers an inverse image from the beginning, according to the negative magnification for common imaging without intermediate image.

In addition to a proper scaling of the pupil coordinates, energy apodization factors as derived in chapter 20.3.2 may be considered for imaging with $\beta \neq 1$. Throughout the scalar treatment of chapters 21–25, however, energy factors are neglected.

21.4

Coherence in Imaging

21.4.1

The Coherent Image

For illumination with a planar wave, e.g., by a collimated laser beam, the effective light source can be simplified by a delta-like point:

$$s(\nu) = S_0 \cdot \delta(\nu - \nu_i) . \quad (21-46)$$

It follows for the coherence function (with $S_0 = 1$):

$$\Gamma_S(x_1, x_2) = e^{i2\pi(\nu_i(x_1 - x_2))} . \quad (21-47)$$

Especially under perpendicular illumination, i.e., $\nu_i = 0$ the coherence function with $\Gamma_L(x_1, x_2) = 1$ degenerates to a constant. With (21-46), the amplitude distribution behind the object is given by

$$U_0(x) = T(x) \cdot S(x) = A(x) \cdot e^{i\phi(x)} e^{i2\pi\nu_i x} . \quad (21-48)$$

In the pupil plane one obtains the frequency spectrum of the object shifted in accordance with the illumination direction and multiplied by the transfer function $h(\nu)$:

$$u_1(\nu) = h(\nu) \cdot t(\nu - \nu_i) . \quad (21-49)$$

After Fourier transformation it follows that the intensity in the image plane is:

$$I(x) = \left| \int T(x') \cdot H(x - x') \cdot e^{i2\pi\nu_i x'} dx' \right|^2 . \quad (21-50)$$

For coherent illumination there is thus a linear transfer of the amplitude, described by the transfer function $T(x)$. The coherent transfer function $h(\nu)$ as a low-pass filter, determines the maximum transferred object frequency. In figure 21-29, coherent image formation is illustrated for the example of a linear grating. During image formation the so-called Gibbs phenomenon occurs and the image amplitude is more or less similar according to the incomplete spectrum of the object amplitude. The object transmission function $T(x)$ of an amplitude grating is described as the binary Ronchi Grating (see chapter 25) with

$$T(x_0) = \text{ronchi}\left(\frac{x_0}{d}\right) . \quad (21-51)$$

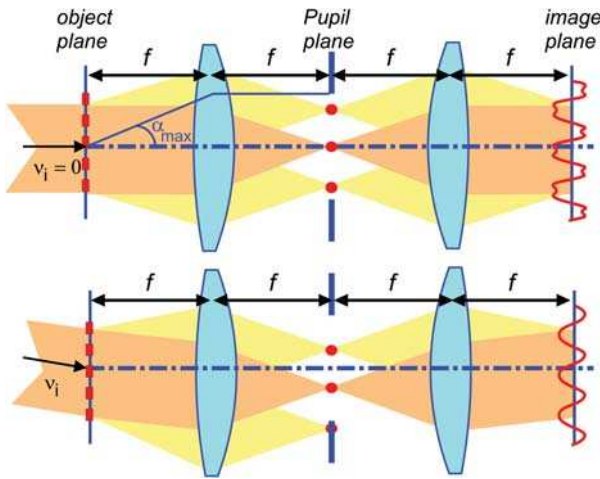


Figure 21-29: Low-pass filtering for optical imaging by the pupil function.

After illumination by a plane wave with the transverse frequency ν_i and after multiplying it by the aperture function $P(\nu)$ in the pupil plane, the spectrum in the exit pupil is obtained:

$$u_1(\nu) = \frac{d^2}{2} \left[\text{sinc}\left(\frac{d}{2}(\nu - \nu_i)\right) \cdot \text{comb}(d \cdot (\nu - \nu_i)) \right] \cdot \text{circ}\left(\frac{\nu}{NA/\lambda}\right). \quad (21-52)$$

Of the diffraction orders only those, which pass the pupil function determined by the radius NA/λ contribute to the image formation:

$$\sqrt{\left(\frac{m}{d} - \nu_{ix}\right)^2 + \nu_{iy}^2} \leq \frac{NA}{\lambda} \quad (21-53)$$

with the illumination direction ν_{ix}, ν_{iy} . Figure 21-30a illustrated the position of the diffraction order in the pupil for $\nu_{iy} = 0$. With eq. (21-53) the resolution limit is given in accordance. In figure 21-30b it is shown how a y -component of the illumination direction shifts the diffraction orders by ν_{iy} [21-21]. The number of diffraction orders passing the pupil is now less than that for $\nu_{iy} = 0$. The number of the diffraction orders forming the image grating is limited by the aperture and depends generally on the illumination direction as shown in figure 21-30a and b. The ultimate resolution is obtained when only two diffraction orders pass the lens pupil, as shown at an example in figure 21-31.

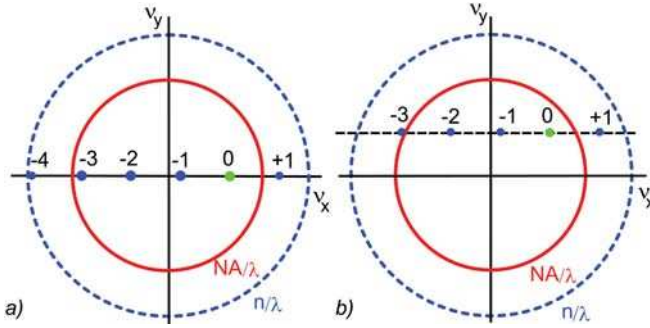


Figure 21-30: a) Diffraction orders in the pupil at oblique coherent illumination; b) for the two-dimensional discussion of generally oblique illumination, the obliquity in the y -direction has to be taken into account.

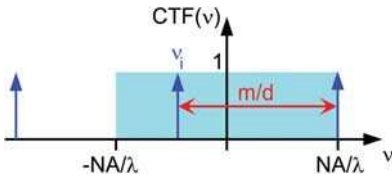


Figure 21-31: Transfer function at the resolution limit for coherent imaging with oblique illumination ν_i .

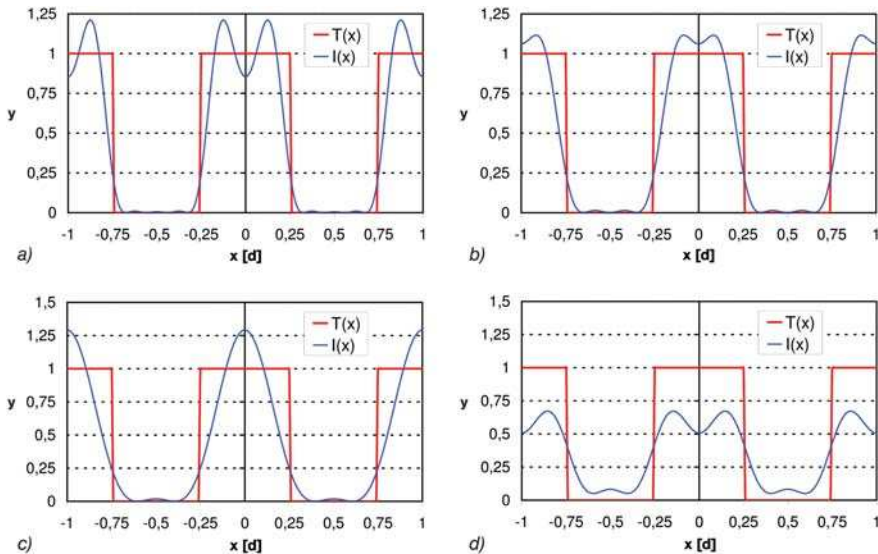


Figure 21-32: Coherent grating images formed by different diffraction orders (numbers for diffraction orders given in brackets) a) $(-3, -1, 0, 1, 3)$; b) $(-1, 0, 1, 3)$; c) $(-1, 0, 1)$; d) $(0, 1, 3)$.

Figure 21-32 compares the coherent grating images for different numbers of diffraction orders contributing to the image according to eq. (21-53). The different grating image intensities deviate significantly from the object transmission function $T(x)$ and from each other. Occasionally, the similarity between object and image is expressed by the so-called *fidelity* which is given by the mean-squared deviation of the image from the object with [21-41]

$$F = \sqrt{\int [I(x) - |T(x)|^2]^2 dx} . \quad (21-54)$$

Besides the Gaussian norm in eq. (21-50) other norms and weighted norms, respectively, are common [21-42]. The fidelity according to (21-54) of the examples in figure 21-32 is $F_a=8.9\%$, $F_b=10\%$, $F_c=16.4\%$ and $F_d=9.5\%$. From a comparison of figure 21-32 c and d it follows that, for similarity of image to object, a higher diffraction order gathered by oblique illumination is preferable to symmetrical image conditions. However, this is at the expense of intensity.

21.4.2

Incoherent Imaging

For an infinitely extended constant effective light source the coherence function is given by a δ -function:

$$\Gamma_S(x_1, x_2) = \int C \cdot e^{i2\pi\nu_q(x_1-x_2)} d\nu_q = C \cdot \delta(x_1 - x_2) . \quad (21-55)$$

After inserting into (21-37) one obtains (with $C = 1$):

$$\begin{aligned} I_1(x) &= \iint \delta(x_1 - x_2) \cdot T(x_1) T^*(x_2) \cdot H(x' - x_1) H^*(x' - x_2) dx_1 dx_2 \\ &= |T(x)|^2 \otimes |H(x)|^2 . \end{aligned} \quad (21-56)$$

The last term of eq. (21-56) can be identified with the point-spread function PSF of the optical system, given by the square value of the amplitude distribution function $H(x)$:

$$I_{\text{PSF}}(x) = |H(x)|^2 . \quad (21-57)$$

In the case of an incoherent image, the intensity is transferred linearly, i.e., the intensity in the image plane results from a convolution of the intensity in the object plane $I_0 = |S(x) \cdot T(x)|^2$ with the point-spread function of the optical system:

$$I_1(x) = I_0(x) \otimes I_{\text{PSF}}(x) . \quad (21-58)$$

Incoherent imaging is therefore very similar to geometrical-optical image formation as discussed in chapter 20.

Equation (21-55), however, is a simplification. Since for monochromatic imaging there are no optical frequencies larger than n/λ , an infinitely extended light source does not exist. However, it can be shown that the result of eq. (21-56) is a very good approximation, if the image of the effective light source fills the pupil of the system completely ($\sigma \geq 1$), with the coherence function given by the point-spread function according to (21-57), in the case of an optical system with rotational symmetry, by the airy pattern:

$$\Gamma_S(x_1, x_2) = \int_{|v_q| < NA/\lambda} e^{i2\pi v_q(x_1 - x_2)} dv_q = \left(\frac{J_1[2\pi(w_1 - w_2)]}{\pi(w_1 - w_2)} \right)^2 \quad (21-59)$$

with normalized coordinates $w = NA/\lambda \cdot r$.

According to linear system theory, the Fourier transform of the point-spread function or point response corresponds to a transfer function of the linear system. From this it follows that the *incoherent optical transfer function* – OTF – $g(v)$ of the optical system is given by the autocorrelation of the coherent transfer function – CTF – $h(v)$:

$$g(v) = F[I_{PSF}(x)] = h(v) \circ h^*(v) = \int h(v') \cdot h^*(v' + v) dv' . \quad (21-60)$$

Equation (21-60) is sometimes called *Duffieux integral*. In frequency domain, the autocorrelation of the spectrum of the transfer function is multiplied to the optical transfer function $g(v)$. In the image plane the image filtered by the OTF is generated:

$$i_1(v) = F[|T(x)|^2] \cdot g(v) = [t(v) \circ t^*(v)] \cdot g(v) . \quad (21-61)$$

It is finally obtained:

$$I_1(x) = \int |T(x')|^2 \cdot G(x - x') dx' . \quad (21-62)$$

In the next paragraphs incoherent imaging is further discussed at the examples of one-dimensional image formation and systems with rotational symmetry.

21.4.3

One-Dimensional Incoherent Imaging

As an example first the point response of a one-dimensional ideal image ($W(v) = 0$) is discussed. In correspondence to the finite wavelength and the limited aperture of an optical system the optical imaging is a low-pass filtering expressed by the pupil function P which only in a finite range is different from zero. The pupil function may thus be described by the rect-function:

$$h(v) = P(v) = \text{rect}\left(\frac{v}{NA/\lambda}\right) . \quad (21-63)$$

Accordinging eq. (21-60) the optical transfer function OTF is given by:

$$\begin{aligned}
 g(\nu) &= \text{rect}\left(\frac{\nu}{NA/\lambda}\right) \circ \text{rect}^*\left(\frac{\nu}{NA/\lambda}\right) \\
 &= \text{rect}\left(\frac{\nu}{2NA/\lambda}\right) \cdot \left[1 - \frac{\lambda}{NA} \text{sign}(\nu) \cdot \nu\right].
 \end{aligned}
 \tag{21-64}$$

The normalized point-spread function (PSF) is given by the squared value of the amplitude distribution function with

$$I_{PSF}^{1D}(x) = |H(x)|^2 = \text{sinc}^2\left(\frac{NA}{\lambda} x\right).
 \tag{21-65}$$

CTF and OTF for a one-dimensional system are illustrated in Figure 21-33. The OTF as the transfer function, describes the amplitude a certain object frequency with which ν is transferred and thus the contrast of the object details formed by that frequency ν . The resolution limit for the incoherent imaging is given by the maximum frequency of the OTF, i.e., by $2NA/\lambda$, according to the maximum resolution of the coherent imaging with oblique illumination, but with vanishing contrast in the incoherent image. The amplitude distribution function and the one-dimensional line image intensity are illustrated in figure 21-34 in normalized coordinates $w = x \cdot NA/\lambda$. The intensity vanishes at distances $m \cdot \lambda/NA$.

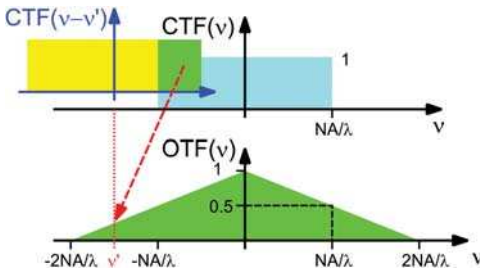


Figure 21-33: Coherent Transfer function CTF and optical transfer function OTF.

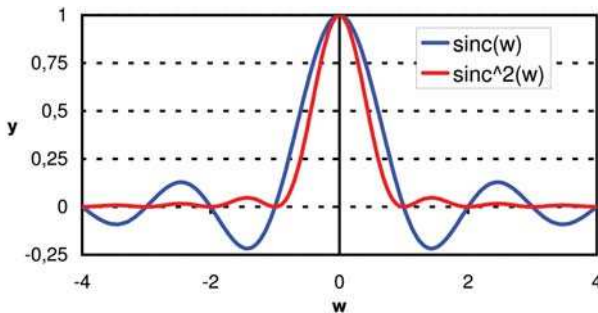


Figure 21-34: Amplitude distribution function $U(w)$ (blue) and point-spread function $I_{PSF}(w)$ (red).

21.4.4

Systems with Rotational Symmetry

For rotation-symmetric imaging systems the amplitude distribution can be written in cylindrical coordinates:

$$H(r, \theta) = \int_0^{2\pi} \int_{\rho} h_{\rho}(\rho, \vartheta) \cdot e^{i2\pi r \rho \cos(\theta - \vartheta)} \rho d\rho d\vartheta . \quad (21-66)$$

For rotation-symmetric systems the ϑ -integration can be carried out and it is obtained with $\theta = 0$

$$H(r) = 2\pi \int_{\rho} h_{\rho}(\rho) \cdot J_0(2\pi r \rho) \rho d\rho . \quad (21-67)$$

For ideal systems the CTF is now described by the circ-function:

$$h(\rho) = P(\rho) = \text{circ}\left(\frac{\rho}{NA/\lambda}\right) . \quad (21-68)$$

For a constant function $h_{\rho}(\rho) = 1$ it follows for the amplitude distribution function $H(r)$

$$\begin{aligned} H(r) &= 2\pi \int \text{circ}\left(\frac{\rho}{NA/\lambda}\right) \cdot J_0(2\pi r \rho) \rho d\rho = 2\pi \int_{\rho < NA/\lambda} J_0(2\pi r \rho) \rho d\rho \\ &= \frac{NA}{r\lambda} J_1\left(2\pi r \cdot \frac{NA}{\lambda}\right) . \end{aligned} \quad (21-69)$$

It should be mentioned here that it is not trivial to obtain a homogenous illumination in the exit pupil! In particular, a homogeneous illumination is not achieved by an ideal spherical wave, rather a Lambertian radiator is necessary, because the canonical pupil coordinate is linear to the sine of the angle, not to the angle itself. In Fourier theory, however, a delta-like spot with constant spectrum exhibits a Lambertian characteristic.

The intensity formation of eq. (21-69) gives the *Airy disc* (see chapter 20)

$$I_{PSF}^{2D}(r) = |H(r)|^2 = \left(\frac{\pi \cdot NA}{\lambda}\right)^2 \left(\frac{J_1(2\pi w)}{\pi w}\right)^2 . \quad (21-70)$$

with normalized coordinates $w = NA/\lambda \cdot r$. Figure 21-35 shows the normalized amplitude $H(w)$ and the intensity distribution $I(w)$ of the Airy disc with the first zeros and side maxima. The first minimum of the diffraction image or Airy disk is at $r_0 = 0.611 \cdot \lambda/NA$.

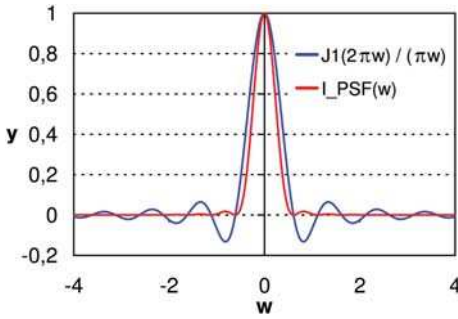


Figure 21-35: Cross-section through Airy disc.

The optical transfer function OTF can be determined from the (two-dimensional) autocorrelation of the circ-function. From a geometrical construction we obtain

$$g(\rho) = \frac{2}{\pi} \left[\arccos\left(\frac{\lambda \cdot \rho}{2NA}\right) - \rho \cdot \sqrt{\left(\frac{2NA}{\lambda}\right)^2 - \rho^2} \right]. \quad (21-71)$$

Figure 21-36 illustrates the geometrical construction and the optical transmission function OTF of a circular aperture in comparison with the one-dimensional OTF. As can be seen, the optical transmission function and thus the image contrast for a circular aperture is significantly lower compared with the linear response of a one-dimensional system (or square aperture). At the coherent resolution limit with $\rho = NA/\lambda$ the difference is 10.9%, corresponding to a relative contrast loss of 21.8%. The maximum difference is at $\rho = 1.24 \cdot NA/\lambda$ with 11.5% difference in OTF, and 30.1% relative contrast loss. The maximum relative contrast loss is in the limit at $\rho = 2NA/\lambda$.

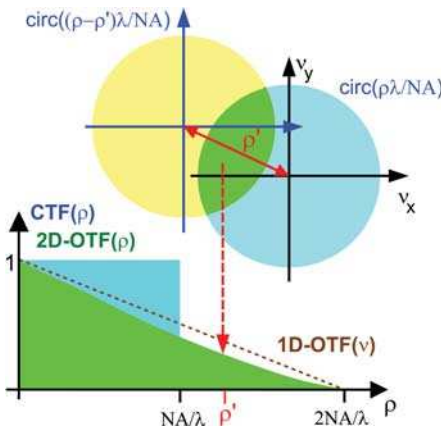


Figure 21-36a: Construction of the transfer function of a rotation-symmetric system: CTF in the coherent case and OTF in the incoherent case.

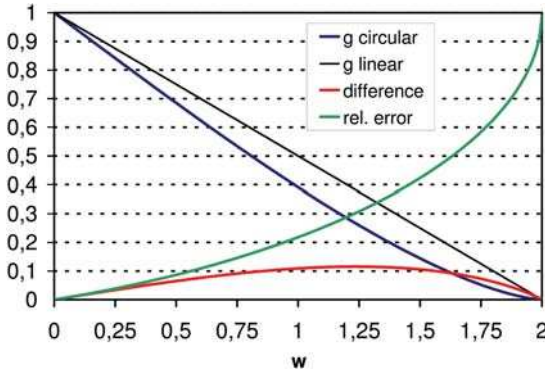


Figure 21-36b: Transfer function of a rotation-symmetric system in comparison with a one-dimensional system (linear g) in normalized coordinates $w = \rho \cdot \lambda/NA$.

For the rotation-symmetric incoherent image, the contrast of object details with high frequencies vanishes even faster than for the one-dimensional image. The limit frequency, however, is again at $\nu_{\max} = 2NA/\lambda$. Thus, theoretically, object details with a period of $0.5\lambda/NA$ can be transferred. Two examples of image simulation results for incoherent images of a grating are illustrated in figure 21-37.

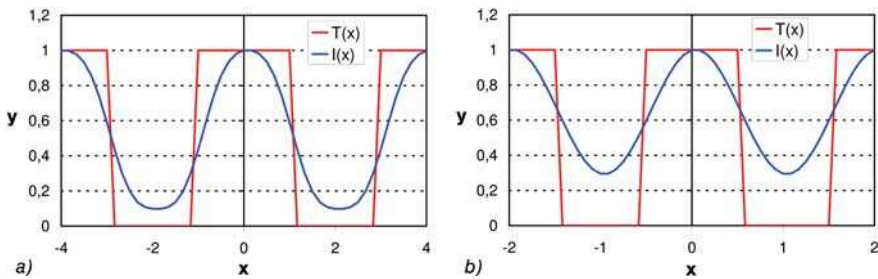


Figure 21-37: Incoherent grating images with $NA=0.5$, $\lambda=1$; a) with grating period $d=4\lambda$; b) with grating period $d=2\lambda$.

21.4.5

Conditions for Incoherent, Partially Coherent and Coherent Imaging

Only incoherent optical imaging can be considered as a linear system for the transfer of the intensity. Every object point $\delta(x - x_0)$ is imaged in an image point. Due to the wave-optical diffraction, the delta-like object point is “blurred”. Each object point $\delta(x - x_0)$ is transferred by the optical system characterized by a transfer function into the point response, given by the Airy disk. In ideal systems, the condition that the diffraction image given by the Airy disk does not depend on the position of the object point and shift invariance is strictly fulfilled. The image $I_1(x)$ of an object distribution $I_0(x)$ is therefore given by the linear superposition of the single point responses according to the object points in $I_0(x)$, i.e., by a convolution of the object

$I_0(x)$ with the point response. Only for the case of incoherent imaging does the optical image follow the Raleigh description of image formation. Incoherent imaging is always applicable when an effective light source is used for illumination whose image fills completely the pupil of the imaging optics or if self-luminous objects like stars are considered. However, as will be shown in this section, the coherence properties of optical imaging also depend on the object.

First the imaging of a *pinhole diaphragm* is considered. The imaging of a pinhole diaphragm is always coherent. The intensity is the squared value of the amplitude distribution function or point-spread function. It does not matter which light source is used. The pupil or filter plane is completely illuminated by the delta peak since the Fourier transform of the object wave is a plane wave (figure 21-38). After illumination with a light source point ν_q the object spectrum is shifted by ν_q . However, since the object spectrum (plane wave) is infinitely extended, the convolution with the light source has no significant effect apart from an unimportant constant phase. For all illumination angles, the pupil illumination is identical and all images to all light source points are equal. Approximately this is of course valid as well for small slits and openings whose diffraction spectrum is much larger than the opening angle of the optical set-up.

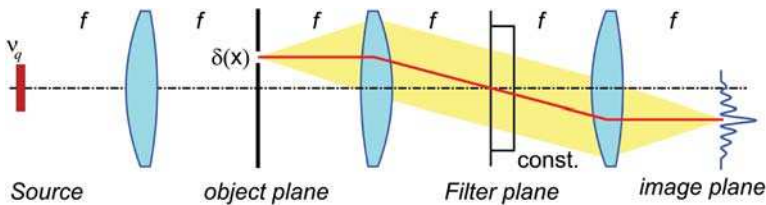


Figure 21-38: Imaging of a delta point.

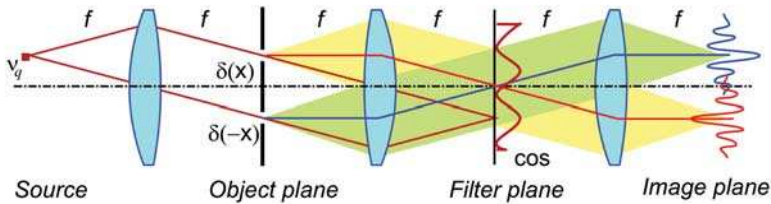


Figure 21-39: Coherent imaging of a double slit.

If a second slit is added, i.e., for a double slit, interference effects can occur depending on the illumination. Considering first a single source point, both slits become secondary sources coherent to each other, which are emitting cylindrical waves with a locked-phase correlation (figure 21-39). Therefore in the pupil plane the interference pattern of the double slit diffraction in the far-field is formed. For slit openings assumed small, this is given by cosine-shaped interference fringes with a phase position according to the illumination direction. In the image plane the two cylindrical wave amplitudes are added – filtered by a low pass – and interfere with each other. Here the phase position of the two interfering line images is given

by the position of the source point and the corresponding phase shift of the interference fringe pattern in the filter plane. Coherent imaging means therefore that the phase differences of the wave trains are rigidly locked. At an image position in the limiting case either destructive or constructive interference can occur depending on the illumination angle.

For complete incoherence (incoherent imaging), as illustrated in figure 21-40, the discussion is again very simple: it is achieved when all relative phase differences and thus all possible phase positions of the interference patterns occur. In the filter plane there are no longer any interference fringes visible according to the arbitrary phase position of the cosine fringes pattern. In the image this shows itself in adjacent point responses of object points not interfering with each other, since the mean value over time is equal to the sum of the intensities of the single images. The coherence function in this case degenerates to a delta function, an infinitely extended source is assumed. In general, a source whose angle extension is equal to or larger than the NA of the imaging optical system is sufficient.

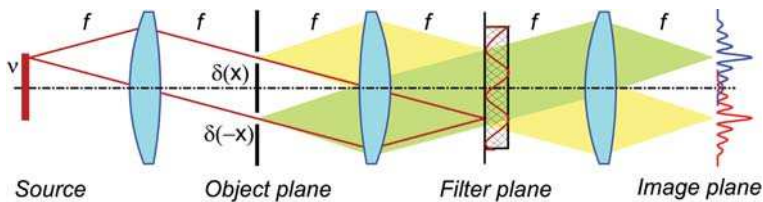


Figure 21-40: Incoherent image of a double slit.

For *partially incoherent imaging* with a light source having a radius proportional to $\sigma \cdot NA$ the spectra of the object are incoherent to each other are shifted just by the angle divergence of the light source in the filter plane. The averaging effect now depends on how far the spectra are shifted. The contrast of the “interference fringes” in the pupil grows correspondingly worse (figure 21-41).

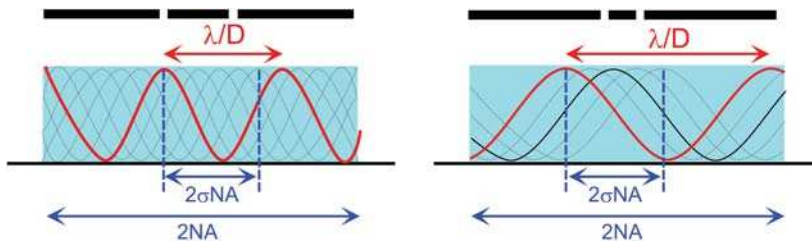


Figure 21-41: Different phase positions of the Fourier spectrum of the object depending on the illumination angle for the same light source size but different slit distances.

For closely adjacent slits, the modulation period in the frequency domain is large. The averaging by the angle divergence of the light source thus becomes smaller with

decreasing slit distance. With a slit distance D it follows for a period p of the interference fringes in the filter plane (in units $f = 1$)

$$p = \frac{\lambda}{D} . \quad (21-72)$$

The slit distance D at which the interference effects just vanish, it is given by

$$D = \frac{\lambda}{2\sigma \cdot NA} . \quad (21-73)$$

Inverting the derivation above as a condition for an incoherent imaging, it follows for the minimum illumination divergence in dependence on the slit with (or object period) D that:

$$\sigma \cdot NA > \frac{\lambda}{2D} . \quad (21-74)$$

From this it can be concluded that an image has in general increasing incoherent properties with increasing slit distances and with increasing angular divergence $\sigma \cdot NA$ of the source.

In the case of coherent and incoherent imaging, for selected examples, analytical solutions exist, while for partial coherent imaging usually numerical simulation is required.

21.4.6

Imaging with Correlated Light Sources

Lamps like discharge sources or plasma sources usually have a large volume in the order of a few millimeters and emit incoherent light in all directions. The product of a projected cross-section of the source volume and collected aperture angle defines the étendue or geometrical flux of a light source. The étendue or geometrical flux of lamps or discharge sources is thus very large, sometimes even larger than the geometrical flux for the largest imaging setting of the imaging optical system. Since the invention of lasers in the 1960s, there has been a widespread application of laser to microscopic imaging, including optical filtering techniques like phase contrast imaging, in holography, in scanning systems and in various others. In applications with laser light as the illumination source, however, deleterious effects in the image can be observed, such as the presence of speckle noise and interference fringes.

For the use of lasers as light sources the distinction between Köhler and critical illumination is invalid since, for laser radiation a sensible definition of the object and image plane cannot be given, and incoherent “light source points” do not exist. In other words, in imaging systems with laser radiation the single light source points assumed in the condenser pupil are no longer uncorrelated with each other. In the general case, the effective light source consists of a mixture of M modes

which are incoherent to each other. The image amplitude for each mode m can thus be calculated according to eq. (21-34)

$$U_m(x) = \iint s_m(\nu_q) T(x') \cdot H(x - x') \cdot e^{i2\pi\nu_p x'} dx' \nu_q. \quad (21-75)$$

Then the image intensity is given by the sum of the intensities over all modes m incoherent to each other:

$$I(x) = \sum_{m=0}^M |U_m(x)|^2. \quad (21-76)$$

The calculation of the image intensity according to eq. (21-75)–(21-76) is much more complex for correlated light sources and therefore often avoided. The correlation of the light source is often destroyed by e.g. rotating diffusing disks or a spinning wedge, both in order to change the phase of the illumination light, which destroys the coherence and thus reduces the speckle pattern [21-43], [21-44], [21-45], [21-46], [21-47]. In the remainder of this book the discussion will be restricted to uncorrelated light sources.

21.5 Literature

- 21-1 G.B. Airy, On the diffraction of an object glass with circular aperture, *Transactions of the Cambridge Philosophical Society* 5, 283–291 (1835).
- 21-2 Lord Rayleigh, On the theory of optical images, with special reference to the microscope, *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science* 42, Part XV, 167–195 (1896).
- 21-3 E. Abbe, Beiträge zur Theorie des Mikroskops und der mikroskopischen Wahrnehmung, *M. Schultze's Archiv für mikroskopische Anatomie*, Vol. IX, 413–468 (1873).
- 21-4 H. Volkmann, 'Ernst Abbe and his work', *Appl. Opt.* Vol. 5, p 1720–1731 (1966).
- 21-5 P. H. Van Cittert, Die wahrscheinliche Schwingungsverteilung in einer von einer Lichtquelle direkt oder mittels einer Linse beleuchteten Ebene, *Physica I*, 201–210 (1934).
- 21-6 F. Zernike, The concepts of degree of coherence and its application to optical problems, *Physica V*, 785–795 (1938).
- 21-7 H.H. Hopkins, The influence of the condenser on microscopic resolution, *Proc. Phys. Soc. (B)* 63, 737–744 (1950).
- 21-8 J.W. Goodman, *Introduction to Fourier Optics* (1968).
- 21-9 E. M. Nelson, *Critical microscopy*, J. R. Microscop. Soc. 282–289 (1910).
- 21-10 M. Pluta, *Advanced Light Microscopy*, Vol. 1, Principles and Basic Properties, Elsevier (1988).
- 21-11 L.C. Martin, *The Theory of the Microscope* (Blackie, London, 1966).
- 21-12 M. Berek Über Kohärenz und Konsonanz des Lichtes, IV. Die optische Abbildung nichtselbstleuchtender Objekte, *Z. f. Physik* 37, 420–449 (1926).
- 21-13 A. B. Porter, On the diffraction theory of microscopic vision, *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science*, Vol. 11, Part XII, 154–166 (1906).
- 21-14 L. Mandelstam, *Ann. d. Phys.* 35, 881, (1911).
- 21-15 M. v. Laue, Zur Theorie der optischen Abbildung, *Ann. d. Physik* 43, 165–168 (1914).
- 21-16 B. Kachar, Asymmetric illumination contrast: A Method of image formation for video light microscopy, *Science* 227, 766–768 (1985).

- 21-17 H.H. Hopkins, On the diffraction theory of optical imaging, Proc. Royal Soc. of London 217A, 408–432 (1953).
- 21-18 Talbot, F. Facts relating to optical science. IV, Philos. Mag. 9, 401–407 (1836).
- 21-19 Patorski, The self- imaging phenomenon and its applications, Prog. Optics 27, 1–108 (1989).
- 21-20 N. Streibl, Three-dimensional imaging by a microscope, J. Opt. Soc. Am. A Vol. 2, 121–127 (1985).
- 21-21 A. Koehler Die Mikrophotographie bei ultravioletem Licht (1926).
- 21-22 C. Troester, Ein Vorschlag zur Steigerung der Leistung des Mikroskopes, Centralblatt f. Bakt. Etc. I. Abt. Originale Bd. 95, 94–96 (1925).
- 21-23 H. Osterberg, Microscope imagery and resolution, J. Opt. Soc. Am. 40, 295–303 (1950).
- 21-24 G.H. Needham, The Practical Use of the Microscope (Thomas Publ., 1958).
- 21-25 H. Siedentopf, Über das Auflösungsvermögen der Mikroskope bei Hellfeld- und Dunkelfeldbeleuchtung, Z. Wiss. Mikroskopie 32, 1–42 (1915).
- 21-26 K. Michel, Die Grundzüge der Theorie des Mikroskops, Wissenschaftliche Verlagsgesellschaft M.B.H., Stuttgart (1981).
- 21-27 L.V. Foster, Microscope Optics, J. Opt. Soc. Am. 40, 275–282 (1950).
- 21-28 J. Rienitz, Two-point resolution with partially coherent illumination, and the relation between self-luminous and illuminated objects, Microscopica Acta 87, 311–328 (1983).
- 21-29 H. Siedentopf, Die Sichtbarmachung von Kanten im Mikroskopischen Bilde Z. Wiss. Mikroskopie 25, 424–431 (1908).
- 21-30 A. Szegvari, Über die Anwendung einer Beleuchtungsazimutblende bei koaxialer Dunkelfeldbeleuchtung, Physik. Zeitschr. 24, 91–95 (1923).
- 21-31 D. L. Fehrs, Lovering, Scruton, Illuminator modification of an optical aligner, Proc. KTI Microelectronics Seminar; KTI Chemicals, Inc. 217–230 (1989).
- 21-32 K. Kamon et al., Photolithography system using annular illumination, Jap. J. Appl. Phys. 30 (11B), 3021–3029 (1991).
- 21-33 A. Kwok-Kit Wong, Resolution enhancement techniques in optical lithography, SPIE TT47 (2001).
- 21-34 F.M. Schellenberg, Resolution enhancement techniques in optical Lithography, SPIE Milestone Series Vol. MS 178 (2004).
- 21-35 M. Eurlings, E. van Setten, J.A. Torres, M. Dusa, R. Socha, L. Capodiecici and J. Finders, 0.11 μm Imaging in KrF lithography using dipole illumination, Proc. SPIE 4404, 266–278 (2001).
- 21-36 W.T. Cathey Optical Information Processing and Holography John Wiley and Sons, New York (1974).
- 21-37 W. Ulrich, S. Beiersdörfer, H.-J. Mann, Trends in optical design of projection lenses for UV- and EUV-Lithography, SPIE 4146, 13–24 (2000).
- 21-38 Ch. Hoffmann Die optische Abbildung (Akademische Verlagsgesellschaft, Leipzig, 1980).
- 21-39 H. Siedentopf, Über die ultramikroskopische Abbildung linearer Objekte, Z. Wiss. Mikroskopie 29, 1–47 (1912).
- 21-40 D.G. Flagello, T. Milster, A.E. Rosenbluth, Theory of high-NA imaging in homogeneous thin films, J. Opt. Soc. Am. A 13, 53–64.
- 21-41 E.H. Linfoot, Qualitätsbewertung optischer Bilder (Vieweg, Braunschweig, 1960).
- 21-42 Ch. Hofmann Die Bewertung optischer Systeme, Fortsch. Physik 27, 595–639 (1979).
- 21-43 M. J. Bowman Two new methods of improving optical image quality, Appl. Opt. 7, 2280–2311 (1968).
- 21-44 H.J. Gerritsen, W.J. Hannan, E.G. Ramberg, Elimination of speckle noise in holograms with redundancy, Appl. Opt. 7, 2301–2311 (1968).
- 21-45 S. Lowenthal, D. Joyeux, Speckle removal by a slowly moving diffuser associated with a motionless diffuser, J. Opt. Soc. Am. 61, 847–851 (1971).
- 21-46 van Ligten, Speckle reduction by simulation of partially coherent object illumination in holography, Appl. Opt. 12, 255–265 (1973).
- 21-47 C. S. Ih, L.A. Baxter, Improved random spatial phase modulation for speckle elimination, Appl. Opt. 17, 1447–1454 (1978).

22

Coherence Theory of Optical Imaging

- 22.1 Introduction 284
- 22.2 Theoretical Description of Partially Coherent Image Formation 284
 - 22.2.1 Hopkins Transmission Cross Coefficient 284
 - 22.2.2 Image Fidelity 287
 - 22.2.3 Hopkins Formalism for Periodic Objects 288
 - 22.2.4 Aberrations in the Linear Grating Image 293
- 22.3 The Coherence Function and the Coherence Transfer Function 296
- 22.4 The Phase Space Description 300
 - 22.4.1 Transformation of Coherence and Wigner Distribution Function 300
 - 22.4.2 Propagation of the Wigner Distribution Function in Free Space 303
 - 22.4.3 Compilation of the Transformations 307
- 22.5 Optical Imaging in the Presence of Aberrations 309
 - 22.5.1 Linear Systems and Classification of Aberrations 309
 - 22.5.2 Random Non-stationary Aberrations: Stray Light and Flare 314
- 22.6 Literature 317

22.1

Introduction

A phenomenological description of image formation according to Abbe already provides a complete interpretation of the aerial images. According to Abbe, every point of the effective light source is to be regarded as independently forming coherent images of the object with varying obliquity of the illumination and the intensity of all the coherent images is summed in the image plane. For the layout of an imaging system and the correct selection of the illumination distribution, the phenomenological approach according to the Siedentopf principle is sufficient in most cases. For confirmation, the partially coherent image can be calculated by simulation routines now available in simulation software. Several commercial optical design and special analysis software packages include appropriate routines for this. The commercial software used to calculate the partially coherent image, enables the computer to take over the complex theoretical description of the partially coherent image formation from the user.

In order to comprehend the events which take place during optical imaging, an advanced discussion of partially coherent image formation is very useful. In this chapter, the Hopkins formalism, the coherence function and the Wigner function are explained in order to achieve this. The role of partially coherent, oblique and structured illumination will be visualized for several examples. In section 22.2 the analytical description of Fourier optics is described, and in particular the Hopkins transmission cross coefficient, TCC , is derived. Section 22.3 summarizes the coherence theory of optical imaging and the transfer of the coherence function. Section 22.4 describes the interpretation of optical imaging by use of the Wigner function. Section 22.5 considers imaging aberrations and gives a classification of these.

22.2

Theoretical Description of Partially Coherent Image Formation

22.2.1

Hopkins Transmission Cross Coefficient

In chapter 21, partial coherent image formation was derived as a superposition of all the individual intensity patterns for each effective light source point. This interpretation of partially coherent optical image formation is easy to be implemented in computer software routines (e.g., [22-1], [22-2]). As well as this method of computing the optical image it is, however, possible to give a formal description even for partially coherent image formation and to calculate a linear transfer function, at least approximately.

Starting with eq. (21-36) the intensity in the image space is described by a Fourier integral

$$I(x) = \iiint |s(\nu_q)|^2 \cdot t(\nu_1) \cdot t^*(\nu_2) \cdot h(\nu_1 + \nu_q) \cdot h^*(\nu_2 + \nu_q) \cdot e^{i2\pi x \cdot (\nu_1 - \nu_2)} d\nu_1 d\nu_2 d\nu_q. \quad (22-1)$$

Introducing the Hopkins transmission cross coefficient $TCC(\nu_1, \nu_2)$ [22-3]:

$$TCC(\nu_1, \nu_2) = \int |s(\nu_q)|^2 \cdot h(\nu_1 + \nu_q) \cdot h^*(\nu_2 + \nu_q) \cdot d\nu_q, \quad (22-2)$$

we now obtain the transfer function for partially coherent image formation:

$$I(x) = \iint TCC(\nu_1, \nu_2) \cdot t(\nu_1) \cdot t^*(\nu_2) \cdot e^{i2\pi x \cdot (\nu_1 - \nu_2)} d\nu_1 d\nu_2. \quad (22-3)$$

The image intensity is therefore computed by the linear superposition of the interference patterns of plane waves with spatial frequency ν_1 and ν_2 and amplitude given by the Hopkins transmission cross coefficient $TCC(\nu_1, \nu_2)$ and the spatial object spectrum contribution at ν_1 and ν_2 . As a consequence, the image intensity in eq. (22-3) depends bilinearly on the object transmission spectrum. Since the transmission cross coefficient depends according eq. (22-2) on the optical system as well as on the illumination, it is again a quite unwieldy measure.

In the following the normalized TCC will be used, i.e., the maximum value of the TCC is normalized to 1:

$$TCC(\nu_1, \nu_2) = \frac{1}{TCC_{\max}} \int |s(\nu_q)|^2 h(\nu_1 + \nu_q) \cdot h^*(\nu_2 + \nu_q) d\nu_q. \quad (22-4)$$

For conventional illumination by circular effective sources centred on the optical axis, the maximum value TCC_{\max} is equivalent to $TCC(0,0)$ and is given by the intensity of the source integrated over the pupil. As a consequence, TCC_{\max} is equivalent to the total source power I_S , if the effective source is smaller or of identical size when compared with the pupil. For structured effective light sources larger compared to the pupil or darkfield illumination, TCC_{\max} is more difficult to be obtained.

Frequently the use of the transmission cross coefficient is limited to the so-called linear part with $\nu_2 = 0$ [22-4]:

$$TCC(\nu, 0) = \frac{1}{TCC_{\max}} \int |s(\nu_q)|^2 h(\nu + \nu_q) \cdot h^*(\nu_q) d\nu_q. \quad (22-5)$$

The linear part of the transmission cross coefficient TCC is equal to 1 if the source is totally enclosed in the region where the pupil, shifted by the frequency ν , overlaps with the unshifted pupil. The transmission cross coefficient goes over into the incoherent contrast transfer function $g(\nu)$ if the effective light source has an equal or larger extension compared with the pupil:

$$g(\nu) = TCC(\nu, 0) = \int h(\nu + \nu) \cdot h^*(\nu) d\nu. \quad (22-6)$$

For ideal imaging without aberrations the integration over the light source region can be converted into:

$$TCC(\nu, 0) = \int P(\nu + \nu_q) \cdot |s^*(\nu_q)| d\nu_q. \quad (22-7)$$

The linear TCC according eq. (22-7) is interpreted as the integral over the image of the effective light source, projected into the pupil. In figure 22-1, the linear normalized transmission cross coefficients $|TCC(\nu_1, 0)/TCC(0, 0)|$ for three different cases for ideal systems are visualized: 1) for incoherent imaging (light source spectrum larger than or equal to the pupil), 2) for partially coherent imaging (light source spectrum smaller than the pupil) and 3) partially coherent oblique illumination. Although the perfectly incoherent image has a higher limit frequency and a higher resolution limit is obtained for it than for the partially incoherent image, the image details at a higher frequency are imaged with low contrast. On the other hand, for partially incoherent imaging, a higher image contrast is achieved due to the better transfer at the low frequencies. For normal imaging, it has transpired to be a good compromise when the light source image fills approximately half of the pupil ($\sigma \sim 0.7$; not shown). For an annular pupil or annular illumination, the transmission cross coefficient can be derived in a similar manner [22-5], [22-6].

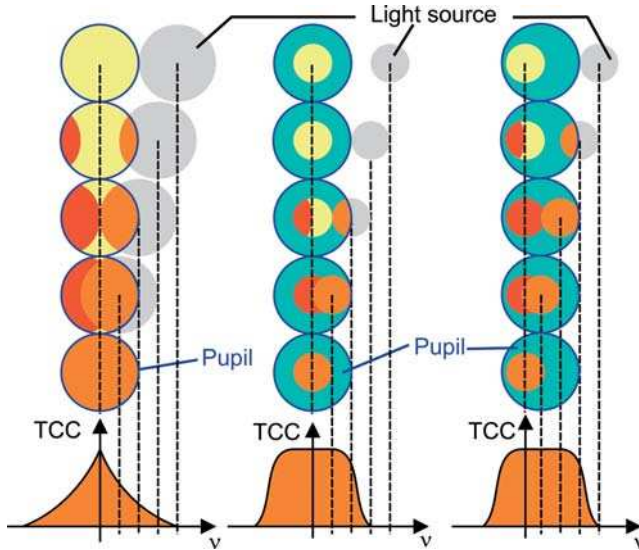


Figure 22-1: Normalized linear part of the Hopkins transmission cross coefficient.

The linear part of the transmission cross coefficient illustrates how a frequency is dampened by the low-pass filter effect of the optical imaging and thus the limiting case for the imaging of infinitely extended sine gratings. For real objects with finite extension and a form deviating from the sine grating, the image contrast deviates significantly from the linear part due to the contributions of the higher diffraction orders and the envelope (see section 5 of chapter 24). Figure 22-2 compares the linear normalized transmission cross coefficient TCC for different light source diameters σ , with the image contrast V , for an amplitude grating with aspect ratio 1:1 with seven periods (simulation results).

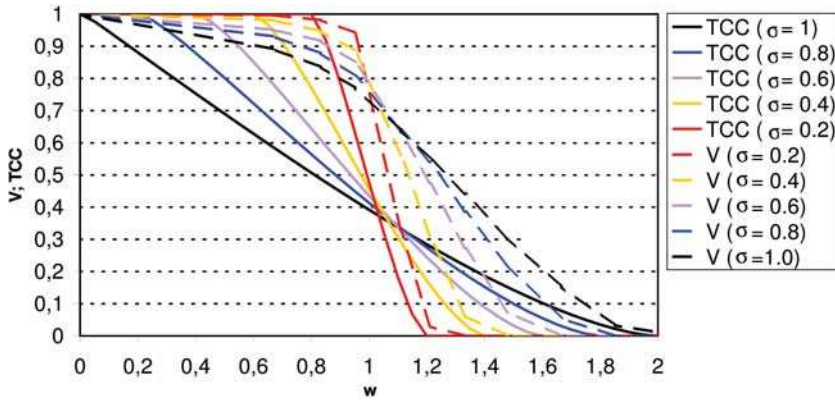


Figure 22-2: Linear part of the Hopkins transmission cross coefficients for different light source radius σ in comparison with simulation result.

The interpretation of the image contrast by the overlap region of the diffraction orders with the pupil, according to the transmission coefficient, might be misleading. As emphasized in chapter 21, the loss in contrast depends rather on that share of the light which is not contributing to the image formation so, e.g., that part of the zero order without interference partner (false light). Due to the smaller overlap, under appropriate illumination the image grows darker at most without the false light share because a larger share of the diffracted light is lost at the pupil. Theoretically, however, the contrast can in principle remain unchanged at $V = 100\%$ up to the incoherent resolution limit. In order to increase the contrast at high object frequencies, oblique illumination must be used. As shown in figure 22-1 the transfer function is then shifted by the incident angle of the main beam of the illumination. In this way a grating is no longer transferred under axial illumination (only the zero order passes the pupil).

22.2.2

Image Fidelity

Since the object spectrum is low-pass filtered by the TCC , the image intensity deviates significantly from the intensity distribution in the object plane. The similarity between the object and the image is expressed by the so-called *fidelity* which is given, e.g., by the mean-squared deviation of the image from the object [22-7] with

$$F = \sqrt{\int [I(x) - |T(x)|^2]^2 dx} . \quad (22-8)$$

Besides the Gaussian norm in eq. (22-8) other norms and also weighted norms, are common [22-8]. The Hopkins formalism allows for an elegant fidelity analysis of the optical image. The ideal image is represented by the square of the object transmission function:

$$I_{\text{ideal}}(\mathbf{x}) = |T(\mathbf{x})|^2, \quad (22-9)$$

or, the ideal image spatial frequency spectrum is given by

$$i_{\text{ideal}}(\nu) = t(\nu) \otimes t^*(\nu) = \int t(\nu') \cdot t^*(\nu - \nu') d\nu'. \quad (22-10)$$

The real image according to eq. (22-3) is, however, given by

$$i(\nu) = \int TCC(\nu', \nu - \nu') \cdot t(\nu') \cdot t^*(\nu - \nu') d\nu'. \quad (22-11)$$

The image fidelity F can thus be written as

$$\begin{aligned} F &= \sqrt{\int [I_{\text{ideal}}(\mathbf{x}) - I(\mathbf{x})]^2 d\mathbf{x}} \\ &= \sqrt{\int [[1 - TCC(\nu', \nu - \nu')] \cdot t(\nu') \cdot t^*(\nu - \nu')]^2 d\nu'}. \end{aligned} \quad (22-12)$$

Due to the low-pass filtering of optical systems and the wave nature of light, the optical image is formed only by a band-limited spatial frequency spectrum, where the damping function of the image frequency spectrum is described by the difference of the normalized TCC 's from one.

22.2.3

Hopkins Formalism for Periodic Objects

The description of optical image formation by the transmission cross coefficient is suited in particular for strictly periodical objects with discrete spectra. If the periodical objects with periods d_x and d_y are illuminated by a plane wave of frequency ν_q , then the object wave is described by a Fourier series expansion with

$$U_0(x, y; \nu_q^x, \nu_q^y) = \sum_{m,n} g_{m,n} e^{i2\pi \left[\left(\frac{m}{d_x} - \nu_q^x \right) x + \left(\frac{n}{d_y} - \nu_q^y \right) y \right]}. \quad (22-13)$$

In the pupil the shifted discrete grating spectrum is obtained and multiplied by the coherent transfer function:

$$u_1(m, n; \nu_q^x, \nu_q^y) = g_{mn} \cdot h\left(\frac{m}{d_x} - \nu_q^x, \frac{n}{d_y} - \nu_q^y\right). \quad (22-14)$$

Considering different light source points, which are not interfering, one obtains for the transmission cross coefficient of the grating frequencies m, n, t and s with intensity $I(\nu_q^x, \nu_q^y)$ of the effective light source distribution (see figure 22-3):

$$TCC(m, n, t, s) = \int I(\nu_q^x, \nu_q^y) h\left(\frac{m}{d_x} - \nu_q^x, \frac{n}{d_y} - \nu_q^y\right) h^*\left(\frac{t}{d_x} - \nu_q^x, \frac{s}{d_y} - \nu_q^y\right) d\nu_q^x d\nu_q^y. \quad (22-15)$$

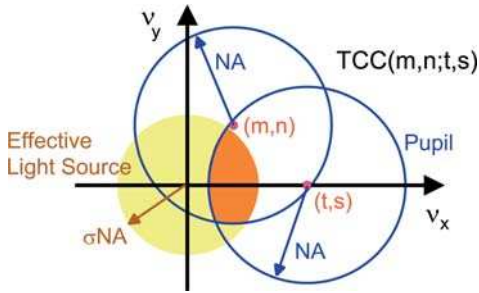


Figure 22-3: A contribution to the transmission cross coefficient of a periodical object.

Thus the intensity in the image plane is given by:

$$I(x, y) = \sum_{m,n} \sum_{t,s} g_{m,n} g_{t,s}^* \cdot TCC(m, n, t, s) \cdot e^{i2\pi \left[\frac{(m-t)}{d_x} x + \frac{(n-s)}{d_y} y \right]}. \quad (22-16)$$

Every multi-beam interference pattern is considered as composed of several two-beam interferences. This special interpretation of partial coherent image formation according to Hopkins becomes more apparent for periodic objects and the discrete formulation of eq. (22-16), compared with the continuous formulation for general objects according to eq. (22-3).

Figure 22-4 illustrates the different shapes of transmission cross coefficients TCC of order (m,n,t,s) for a two-dimensional grating for different light source sizes. For partial coherent illumination with $\sigma=0.7$, the TCC 's are limited by the red circle (which is the effective light source with $\sigma=0.7$). For $\sigma=1$, the limit of the effective light source is identical to the pupil, indicated by the blue circle. In the case of an effective light source image which is larger than the entrance pupil, the light source in integral eq. (22-15) may not be the limiting factor, as the case for TCC 's with either $m \neq 0$ or $n \neq 0$. In the case of dark-field illumination with effective light-source contributions with $\sigma > 1$, the image is formed only by interference between higher diffraction orders such as, e.g., the $TCC(1,1,1,2)$. Care has to be taken, however, that the numerical aperture of the light source does not extend the refractive index into the object space therefore limiting the effective light source size.

In figure 22-5 the TCC of lowest order $m,n < 1$ and $t,s = -3, \dots, 3$ for a two-dimensional grating with $d_x/d_y = 4/5$ and an effective light source diameter $\sigma = 0.85$ are compiled. The example illustrates that, due to the large number of TCC 's, even in the case of periodical objects the solution of eq. (22-16) is in general rather impractical and time consuming for the simulation of partial coherent image formation. In particular, when optical systems with aberrations are considered, symmetry arguments may not be applied and each of the numerous TCC 's has to be computed separately.

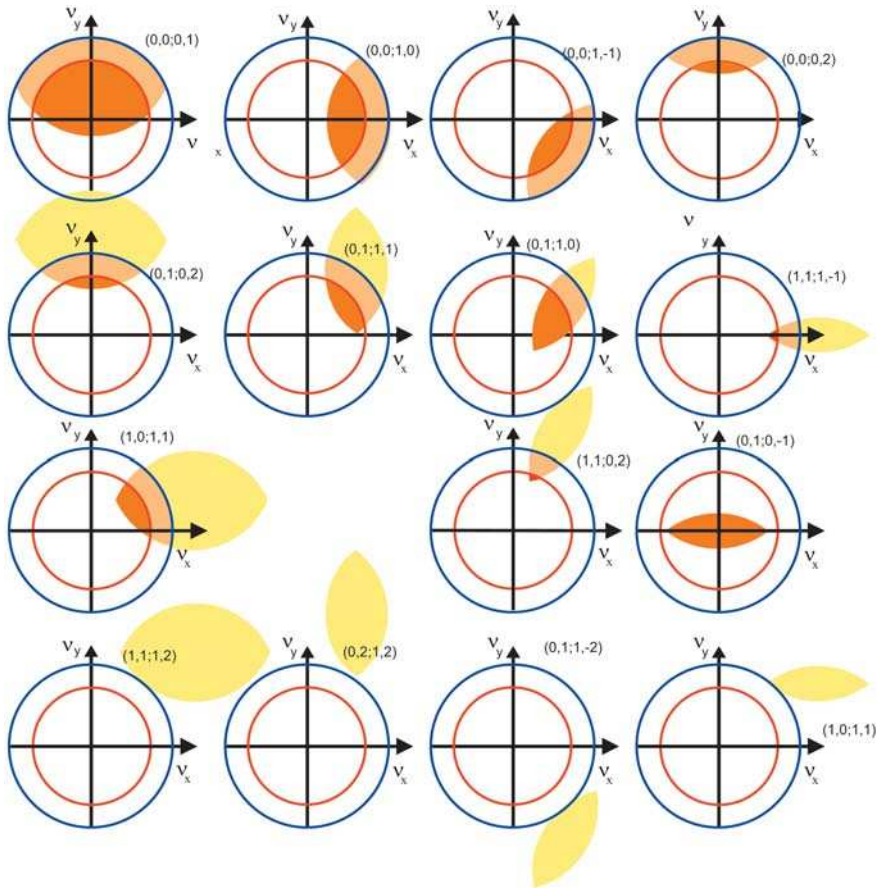


Figure 22-4: Graphical illustration of transmission cross coefficients of order (m, n, t, s) for an infinitely extended light source, incoherent illumination (inside blue circle: $\sigma=1$) and partial coherent illumination (inside red circle: $\sigma=0.7$).

Despite the computational effort required, the partial coherent image formation according to Hopkins, eq. (22-16), nevertheless offers a convenient way of optimizing the ideal effective light source intensity distribution according to the Siedentopf principle. As has been shown in chapter 21, section 2.4, by using a suitable shape for the light source, the diffraction orders contributing to the image can be selected. Certain sections of the light source allow for the transmission of a different number of diffraction orders, and therefore, for different light source sections, different interference patterns will contribute to the image. The different sections of the light

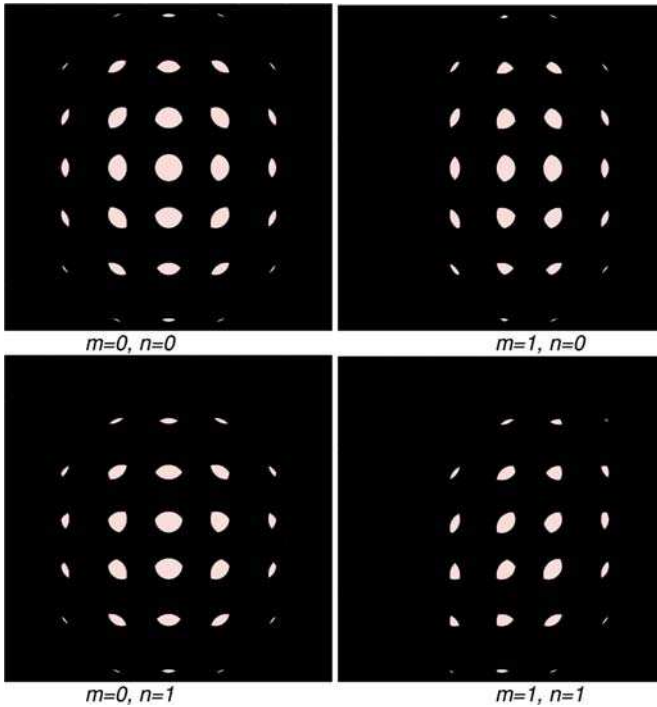


Figure 22-5: $TCC(m, n, t = -3 \dots 3, s = -3 \dots 3)$ with $d_x/d_y = 4/5$ and $\sigma = 0.85$.

source can be identified from the disjoint overlapping regions of the diffraction orders at the entrance pupil. The image intensity can thus be written as a sum over the disjoint light source sections of intensity I_h by

$$I(x, y) = \sum_h I_h \cdot TCC_h \cdot \left\{ \sum_{m,n} \sum_{t,s} g_{m,n} g_{t,s}^* \cdot e^{i2\pi \left[\frac{(m-t)x}{d_x} + \frac{(n-s)y}{d_y} \right]} \right\}_{TCC_h \subset TCC(m,n,t,s)}. \quad (22-17)$$

By means of tailored illumination, emphasis can be added to certain object details, e.g., by an effective light source comprising only selected sections h contributing to the desired image [22-9], [22-10], [22-11], [22-12]. For inspection, the amplitude of a certain interference pattern can be tuned by the appropriate selection of I_h for the different light source sections h . In optical lithography, the amplitudes of certain interference patterns contributing to the image are additionally influenced by the mask with assist features, phase shifters or serifs [22-13].

In the following, the image formation according to eq. (22-17) will be described in more detail using the example of an incoherent light source with $\sigma = 1$ as a starting point. In figure 22-6 the diffraction orders for a two-dimensional grating with incoherent illumination with $\sigma = 1$ are illustrated, where – reduced by symmetry arguments – 11 different sections of the effective light source can be separated. Each section of the effective light source contributes a different set of interference patterns to

the image, formed by the diffraction orders (m,n) and (t,s) (Table 22-1). If, for example, only the effective light source section $h=2$ is selected, a vertical linear grating image consisting of two-beam interferences is formed. The source section $h=8$, on the other hand, forms a horizontal linear grating image comprising only two-beam interferences. A source shape with $I_h \neq 0$ only for $h=2$ and $h=8$, produces therefore a two-dimensional grating image with high contrast and large depth of focus. If two-beam interferences between $m=n=0$ and $s=0, t=1$ should be avoided, permitted illumination directions are thus $h=8, 10$ and 11 . For a correct balance of horizontal and vertical interference patterns, I_h has to be selected in accordance with TCC_h and the contributing diffraction orders $g_{m,n}$ and $g_{t,s}$.

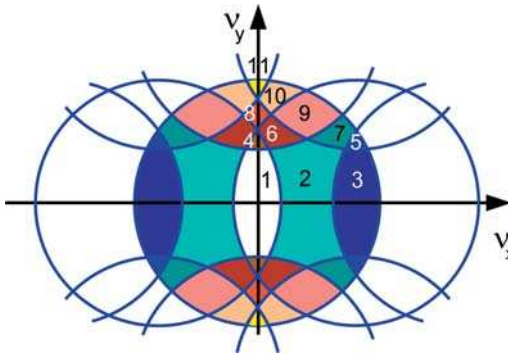


Figure 22-6: Illustration of the source sections $h=0 \dots 11$ forming the transmission cross coefficient TCC_h for a periodical object with incoherent illumination ($\sigma=1$).

Table 22-1: Two-beam interference patterns contributing to source section h of figure 22-6.

m,n	0,0	0,0	0,0	0,0	0,0	0,0	1,0	1,0	1,0	1,0	-1,0	2,0	0,1	0,1	
t,s	1,0	0,1	1,1	2,0	-1,0	-1,1	0,1	-1,0	1,1	2,0	0,1	1,1	1,1	-1,1	
1	x				x			x							
2	x														
3	x			x						x					
4	x	x			x		x	x			x				
5	x		x	x					x	x		x			
6	x	x					x								
7	x		x						x						
8		x													
9	x	x	x				x		x				x		
10		x	x										x		
11		x	x			x							x	x	

22.2.4

Aberrations in the Linear Grating Image

For the example of a linear grating, the role of the TCC is further discussed. For ease of illustration, the TCC is again derived on the basis of Fourier theory in this section. The set-up under consideration is illustrated in figure 22-7. Directly behind the grating in the object plane one obtains:

$$U_0(x, \gamma) = U_L(x, \gamma) \cdot \sum_m g_m \cdot e^{i2\pi\frac{m}{d}x}. \quad (22-18)$$

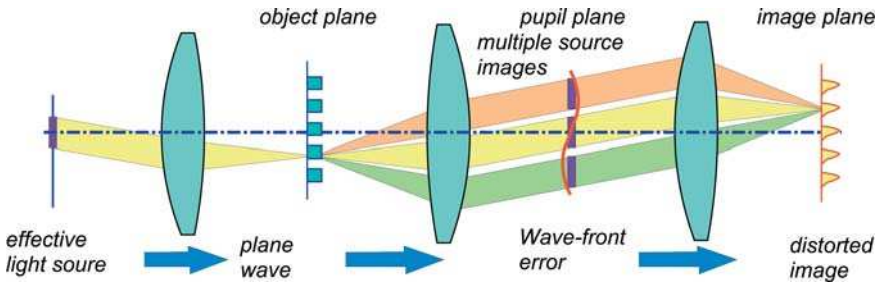


Figure 22-7: Partial coherent imaging of a linear grating with coma.

The transmitted spectrum in the exit pupil is multiplied by the transmission function $h(\nu_x, \nu_y)$:

$$u(\nu_x, \nu_y) = \sum_m g_m \cdot u_L\left(\nu_x - \frac{m}{d}, \nu_y\right) \cdot h(\nu_x, \nu_y). \quad (22-19)$$

For the amplitude in the image plane it follows that:

$$U_1(x, \gamma) = \sum_m g_m \cdot B_m \cdot e^{i2\pi\frac{m}{d}x}. \quad (22-20)$$

with the complex factor B_m defined by

$$B_m = \int h\left(\nu_x + \frac{m}{d}, \nu_y\right) \cdot u_L(\nu_x, \nu_y) e^{i2\pi(\nu_x x + \nu_y \gamma)} d\nu_x d\nu_y. \quad (22-21)$$

Each frequency of the amplitude distribution in image space is thus modulated by an additional complex factor B_m . The factors B_m depend on (22-21) from the light source and the transmission function. The intensity can now be written by a series expansion:

$$I(x) = \sum_m \sum_t g_m g_t^* \cdot B_m \cdot B_t^* \cdot e^{i2\pi\frac{m-t}{d}x} = \sum_n i_n \cdot e^{i2\pi\frac{n}{d}x} \quad (22-22)$$

with the complex coefficients i_n given by

$$i_n = \sum_m g_m g_{m-n}^* \cdot B_m \cdot B_{m-n}^* . \quad (22-23)$$

The image is thus again a periodic object of period d , but with coefficients modulated and dampened by $B_m \cdot B_{m-k}^*$. It has to be noted that description of eq. (22-22) is complex. Since the intensity is real, the imaginary parts of the individual series expansion coefficients must cancel after addition. With an uncorrelated light source, the product $B_m \cdot B_t^*$ can easily be identified as the transmission cross coefficient $TCC(m,t)$, which reduces for one-dimensional objects to a two-dimensional function on the remaining frequency coordinate:

$$TCC(m, t) = B_m \cdot B_t^* = \int I(\nu_q^x, \nu_q^y) h\left(\frac{m}{d} - \nu_q^x, \nu_q^y\right) h^*\left(\frac{t}{d} - \nu_q^x, \nu_q^y\right) d\nu_q^x d\nu_q^y . \quad (22-24)$$

After insertion we obtain from eq. (22-22):

$$I(x, \gamma) = \sum_m \sum_n g_m g_{m-n}^* \cdot TCC(m, m-n) \cdot e^{i2\pi \frac{m}{d} x} . \quad (22-25)$$

It can be recognized, that the $TCC(t,m)$ is equal to the complex conjugate of $TCC(m,t)$:

$$TCC(t, m) = B_t \cdot B_m^* = TCC^*(m, t) . \quad (22-26)$$

For a symmetrical grating with respect to the origin, the expansion coefficients are real, i.e., $g_m^* = g_m$. Thus the summation of eq. (22-25) can be reduced to a single series with

$$b_n = 2 \sum_m g_m g_{m-n} \cdot \text{Re}[TCC(m, m-n)] . \quad (22-27a)$$

$$c_n = 2 \sum_m g_m g_{m-n} \cdot \text{Im}[TCC(m, m-n)] . \quad (22-27b)$$

and for the intensity we obtain:

$$I(x, \gamma) = \frac{b_0}{2} + \sum_{n>0} \left\{ b_n \cdot \cos\left(2\pi \frac{n}{d} x\right) - c_n \cdot \sin\left(2\pi \frac{n}{d} x\right) \right\} . \quad (22-28)$$

The image intensity is thus – like the object – composed by symmetrical cosine-interference patterns, but additionally superposed by a sinusoidal interference pattern due to the imaginary part of the TCC . Imaginary parts of the TCC arise especially from asymmetrical aberrations, as illustrated in figure 22-7. In general, the coherent TCC_{coh} with wave-front aberrations and illumination direction ν_x, ν_y is given by

$$TCC_{\text{coh}}(m, t; \nu_x, \nu_y) = P\left(\nu_x + \frac{m}{d}, \nu_y\right)P\left(\nu_x + \frac{t}{d}, \nu_y\right)e^{\frac{j2\pi}{\lambda}[W(\nu_x + \frac{m}{d}, \nu_y) - W(\nu_x + \frac{t}{d}, \nu_y)]}. \quad (22-29)$$

The $TCC_{\text{coh}}(m, n; \nu_x, \nu_y)$ depends thus on the difference between the wave-front aberrations for the two frequencies m/d and t/d . For partial coherent imaging conditions, the wave-front error differences are averaged over the effective light source intensity distribution:

$$TCC(m, n) = C \iint |u_L(\nu_x, \nu_y)|^2 P\left(\nu_x + \frac{m}{d}, \nu_y\right)P\left(\nu_x + \frac{n}{d}, \nu_y\right) \cdot e^{\frac{j2\pi}{\lambda}[W(\nu_x + \frac{m}{d}, \nu_y) - W(\nu_x + \frac{n}{d}, \nu_y)]} d\nu_x d\nu_y. \quad (22-30)$$

Figure 22-8 compares the effect of aberrations on the grating image with the ideal image for the example of symmetrical and asymmetrical aberrations. All grating images with period $d = 3.6\lambda/NA$ show deformed image periods. In the case of aberration-free imaging, the imaginary part of the TCC_{coh} vanishes and the image is given by a cosine-series expansion like the object, but with dampened series expansion coefficients. For symmetrical image conditions, i.e., symmetrical aberrations and symmetrical diffraction pattern, the complex part of the TCC cancels out to a good approximation and the image is given by a superposition of cosine functions (figure 22-8a). In the case of asymmetrical aberrations like coma, individual interference patterns of frequency k/d in general suffer from phase shifts. The image intensity is again given by a periodic pattern, but with the asymmetrical shape of each period due to imaginary parts of the TCC (figure 22-8b). The phase shifts for the different series terms for Z_7 cause a shift in the centre of gravity of the expected period. In the example, the centre of gravity is approximately at $\sim 0.11/d$. Since this effect is observed particularly by asymmetric wave-front errors like coma, the effect is called coma-induced distortion. Coma-induced distortion depends on imaging conditions such as illumination, aberration level, object period and orientation.

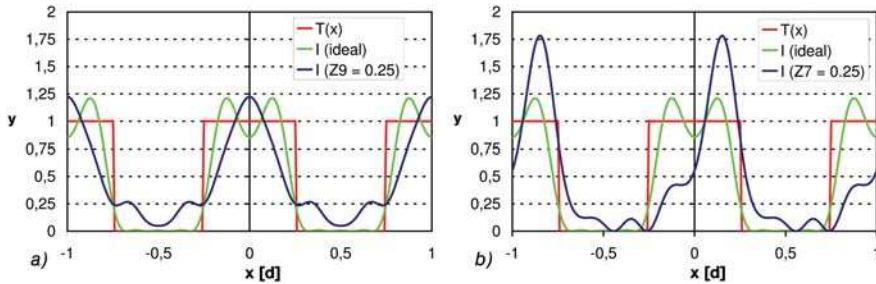


Figure 22-8: a) Effect of spherical aberration and b) coma-induced distortion for the example of a linear grating with $d = 3.6\lambda/NA$.

22.3

The Coherence Function and the Coherence Transfer Function

The partially coherent image in eq. (22-3) or eq. (22-16) turns out to be the sum over the transmission cross coefficients, which are quite awkward to calculate. The calculation of the transmission cross coefficients is particularly inefficient for finite gratings of irregular objects whose spectrum does not consist of single points. Therefore, in general, the simulation of the image formation using transmission coefficients has not caught on and instead the partial coherent image formation using integration over coherent images to source points according to eq. (21-40) is preferred.

The reason for the difficulty with partial coherent image formation is that, for partially coherent imaging, neither the intensity nor the amplitude is transferred linearly. For partial image formation, however, the coherence function is linearly transferred but that is usually four-dimensional. Although modern computers allow the calculation of large amounts of data in a short time, the visualization of four-dimensional functions is not possible. So in the following, the illustration for the discussion is limited to the x dimension.

With the coherence function Γ_S of the light source at the position of the object the intensity in the image plane is given by (see eq. (21-38))

$$I_1(x) = \iint \Gamma_S(x_1, x_2) \cdot T(x_1) T^*(x_2) \cdot H(x - x_1) H^*(x - x_2) dx_1 dx_2 . \quad (22-31)$$

For the (usually considered) uncorrelated light source, the coherence function is given by the Fourier transform of the intensity distribution for the effective light source (figure 22-9a):

$$\Gamma_S(x_1, x_2) = \Gamma_S(|x_2 - x_1|) = \int |s(\nu_q)|^2 \cdot e^{i2\pi\nu_q(x_1 - x_2)} d\nu_q . \quad (22-32)$$

Considering the object as a new source, the coherence function at the position of the object must first be multiplied twice by the object transfer function:

$$\Gamma_{OS}(x_1, x_2) = \Gamma_S(x_1, x_2) \cdot T(x_1) \cdot T^*(x_2) . \quad (22-33)$$

The coherence function now depends on the object transmission $T(x_1)$ and $T(x_2)$ and the coherence function of the illumination. The intensity $I(x)$ behind the object is given by the values “on the diagonal” of the coherence function with $I(x) = \Gamma_{OS}(x, x)$. After the Fourier transformation, the coherence function in the entrance pupil is obtained and this is finally multiplied twice by the transfer function $h(\nu)$. In the image space it follows by inverse transformation for the coherence function Γ_{IOS} in image space

$$\Gamma_{IOS}(x_1, x_2) = \iint \tilde{\Gamma}_{OS}(\nu_1, \nu_2) \cdot k(\nu_1, \nu_2) \cdot e^{2\pi i(\nu_1 x_1 + \nu_2 x_2)} d\nu_1 d\nu_2 \quad (22-34)$$

with the coherence transfer function $k(\nu_1, \nu_2)$ given by

$$k(\nu_1, \nu_2) = h(\nu_1) \cdot h^*(\nu_2) . \quad (22-35)$$

The image intensity again corresponds to the coherence function in the image space “on the diagonal” with $I(x) = \Gamma_{\text{IOS}}(x, x)$. Thus, for the partially coherent imaging the coherence function Γ is transferred linearly with the coherence transfer function $k(\nu_1, \nu_2)$.

The transfer of the coherence function through an imaging optical system is illustrated in figure 22-9, starting with the coherence function $\Gamma_S(x_1, x_2)$ of the light source in object plane in figure 22-9a. In the case of incoherent illumination with an uncorrelated light source, following eq. (22-31), modulations can only be observed perpendicular to the diagonal $\Gamma_S(x, x)$, which corresponds to the coherence function $\Gamma(x = x_2 - x_1)$ as the Fourier transform of a light source. In figure 22-9b, the coherence function $\Gamma_{\text{OS}}(x_1, x_2)$ in the object plane behind of the grating can be seen. All grating openings are homogeneously illuminated (grating with a period $d = 1.2\lambda/\text{NA}$, $\text{NA} = 1$). Figure 22-9c and e illustrates the coherence function in the entrance and exit pupil, respectively. For an ideal imaging system, the coherence transfer function $k(\nu_1, \nu_2)$ is simply given by a two-dimensional rect-function with

$$k(\nu_1, \nu_2) = \text{rect}(\lambda \cdot \nu_1 / \text{NA}) \cdot \text{rect}(\lambda \cdot \nu_2 / \text{NA}). \tag{22-36}$$

The coherence transfer function is illustrated in figure 22-9d. Figure 22-9f finally shows the coherence function in the image plane.

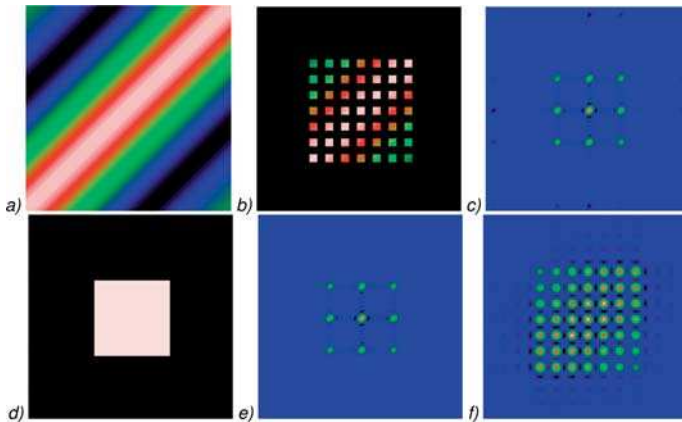


Figure 22-9: Cross-sections through the coherence function (x_1 -axis to the right, x_2 -axis upwards) a) of a conventional light source with $\sigma = 0.05$; b) of a grating illuminated by it; c) after Fourier-transformation in the entrance pupil; d) coherence transfer function k of the ideal image; e) after filtering in the exit pupil and f) in the image plane.

From eq. (22-34) the coherence function in the image space Γ_{IOS} can be written as a two-fold convolution of the complex two-dimensional object coherence function Γ_{OS} with the Fourier transform of the coherence transfer function, $K(x_1, x_2)$:

$$\Gamma_{\text{IOS}}(x_1, x_2) = \iint \Gamma_{\text{OS}}(x'_1, x'_2) \cdot K(x_1 - x'_1, x_2 - x'_2) dx'_1 dx'_2 \tag{22-37}$$

with $K(x_1, x_2)$ given by

$$K(x_1, x_2) = H(x_1) \cdot H^*(x_2) . \quad (22-38)$$

The coherence function in image space $\Gamma_{\text{IOS}}(x_1, x_2)$ is thus blurred by the amplitude distribution function $H(x)$ in both coordinates x_1 and x_2 .

In addition to the coherence properties of the source, the correlation function of the amplitude distribution $K(x_1, x_2)$ also influences the minimum region of coherence in an optical image. With the coherence function of the source and no object (i.e., $T(x) = 1$), the coherence function Γ_{IS} in the image plane is given by

$$\Gamma_{\text{IS}}(x_1, x_2) = \iint \Gamma_S(x'_1, x'_2) \cdot K(x_1 - x'_1, x_2 - x'_2) dx'_1 dx'_2 . \quad (22-39)$$

The coherence function $\Gamma_{\text{IS}}(x_1, x_2)$ determines those regions in which interference effects can occur during the imaging. Two object points, x_1 and x_2 , can be imaged free from disturbing interferences only when Γ_{IS} vanishes for the point distance $x_2 - x_1$. If the image of the effective light source in the entrance pupil is larger than the entrance pupil, then it is limited by the aperture. The coherence properties are thus invariant for any effective light source diameter with $\sigma > 1$. For $\sigma \geq 1$, the coherence properties in the image plane are dominated by the aperture of the optical system. In any incoherent imaging, for illuminated objects with $\sigma \geq 1$ as well as for self-luminous objects, the minimum region of coherence in the image plane is given by the point-spread function $I_{\text{PSF}}(r) = |H(r)|^2$. In ideal systems of rotational symmetry the point-spread function is given by the Airy disc for the ideal rotation-symmetric imaging system, with its first minimum at $r_1 = 0.611 \lambda/\text{NA}$. If the light source is smaller however, the coherence radius accordingly increases to

$$r_c = 0.611 \cdot \frac{\lambda}{\sigma \cdot \text{NA}} = \frac{r_1}{\sigma} . \quad (22-40)$$

As a consequence, the coherence properties of imaging, i.e. the image point distances, for which coherence effects vanish, can be influenced with special structures, sizes and shapes of the effective light source.

In the presence of aberrations, the mutual coherence function in the image plane is additionally changed in accordance with the aberrated point-spread function. Thus the range of the coherence function might be significantly increased, diminishing the imaging quality even for separated image points. Figures 22-10 and 22-11 compare the coherence transfer functions $k(\nu_1, \nu_2)$ (real parts) and the correlation functions of the amplitude distribution function $K(x_1, x_2)$ obtained by Fourier transformation for the ideal system and for different aberrations. As for the coherence functions, the point-spread function, as the image intensity of a single object point, is given by the diagonal of K with $I_{\text{PSF}}(x) = |H(x)|^2 = K(x, x)$. An anti-symmetric wavefront aberration coma leads to an asymmetric amplitude distribution function. Therefore, coherence effects with neighbouring points appear only in one direction.

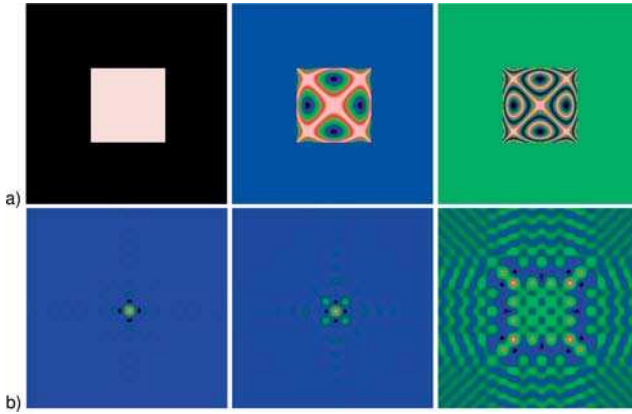


Figure 22-10: a) Real part of the coherence transfer function $k(\nu_1, \nu_2)$ and b) the corresponding correlation functions of the amplitude distribution function $K(x_1, x_2)$ for ideal imaging (left), with spherical aberration $Z_9 = \lambda/5$ (centre), and with spherical aberration $Z_9 = \lambda$ (right).

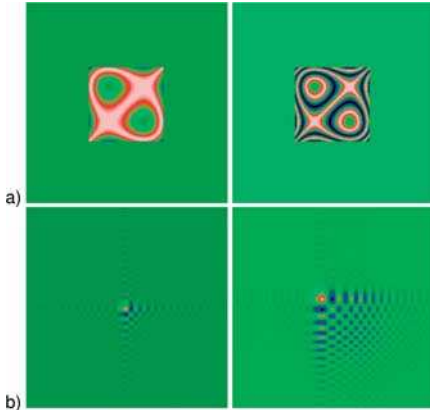


Figure 22-11: a) Real part of the coherence transfer function $k(\nu_1, \nu_2)$ and b) the corresponding correlation functions of the amplitude distribution function $K(x_1, x_2)$ with Coma $Z_7 = \lambda/5$ (left) and $Z_7 = \lambda$ (right).

Figure 22-12 compares the coherence function in the image plane for the imaging of a wide slit for a coherent (22-12a) and incoherent (22-12b) illumination. The diagonals of the real part of the coherence functions represent the intensities in the image plane (figure 22-13). While for incoherent imaging a rounding of the edge occurs, for coherent imaging, increasing modulations are generated towards the edge, according to the Gibbs phenomenon for low-pass filtering. For an asymmetric wave-front error such as, e.g., coma, the coherence function is faded only on one side, according to the point-spread function of a coma. In the coherence function this is expressed by secondary maxima to the diagonal amplified only on one side (figure 22-12c).

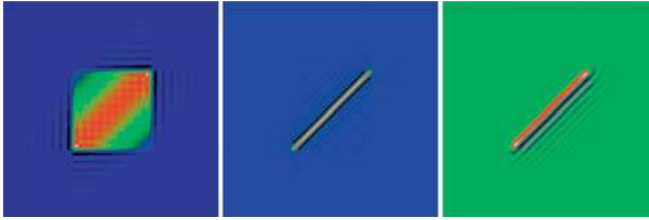


Figure 22-12: Coherence function $\Gamma(x_1, x_2)$ in the image plane for the imaging of a slit of 10λ width; a) coherent imaging with $\sigma = 0.05$; b) incoherent imaging with $\sigma = 1$ and c) in coherent imaging with coma $Z_7 \sim \lambda/5$.

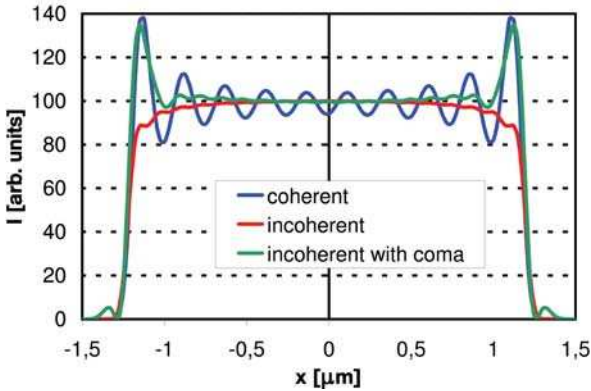


Figure 22-13: The diagonal of the coherence function delivers the image intensity $I(x_1) = \Gamma(x_1, x_1)$.

The calculation of the partial coherent image by the coherence transfer function offers the capability of considering subsequent imaging, also with subsequent filter operations for linear system theory [22-14]. Due to the large numerical effort required it is, however, impractical and therefore not widely used, although it may be used for illustration. Some examples of imaging calculations with the coherence function can be found in [22-15], [22-16], [22-17], [22-18].

22.4

The Phase Space Description

22.4.1

Transformation of Coherence and Wigner Distribution Function

The coherence function as a two-dimensional function is defined by two position or frequency coordinates, respectively. Coherence functions in the position space and in the frequency space are correlated by a two-fold Fourier transformation. As introduced in chapter 19, phase space representations are mixed representations in space

and the spatial frequency domain. Phase space functions are, e.g., the Wigner distribution function, the luminance (in physical units: radiance) or the geometrical flux (see chapter 20.5). In the following, the Wigner distribution function WDF is applied to partially coherent imaging [22-19], [22-20]. The Wigner distribution function is given by

$$W(x, \nu) = \int U\left(x + \frac{\Delta x}{2}\right) U^*\left(x - \frac{\Delta x}{2}\right) \cdot e^{-2\pi i \Delta x \cdot \nu} d\Delta x . \quad (22-41)$$

The projection of the Wigner distribution function in the direction of the frequency axis gives the intensity:

$$I(\nu) = \int W(x, \nu) d\nu . \quad (22-42)$$

As shown in chapter 19, the Wigner distribution function $W(x, \nu)$ follows from the coherence function $\Gamma(x_1, x_2)$ after coordinate transformation to $J(x, \Delta x)$ with the difference coordinate Δx , the centre-of-mass coordinate x , and a one-dimensional Fourier transformation with respect to the difference coordinates [22-21]:

$$W(x, \nu) = \int J(x, \Delta x) \cdot e^{-2\pi i \Delta x \cdot \nu} d\Delta x . \quad (22-43)$$

Unlike the geometrical-optical phase space distribution functions, such as luminance or geometrical flux, the Wigner distribution function can assume negative values since, as a wave-optical density function, it takes the influence of diffraction into account. Figure 22-14 shows the geometrical-optical phase space density and the Wigner distribution function in the object and image plane for the imaging of a 10λ -slit under coherent and incoherent illumination. The x -axis is pointing to the right, the frequency axis upwards. The geometrical-optical phase space densities in figure 22-14a are limited in the direction of the x -axis by the width of the slit, and in the direction of the ν -axis by the illumination aperture with $\nu = \sigma NA/\lambda$. For incoherent illumination (figure 22-14c and e) with $\sigma = 1$ the frequency spectrum is limited at the end by the aperture $NA = 1$. The Wigner distribution function, however, shows diffraction effects in the object plane due to the diffraction at the edge of the slit opening which is more distinct with coherent illumination (figure 22-14b and d) [22-22]. The effect of low-pass filtering of the optical system becomes noticeable when cutting back the higher propagation angles. At the same time, this leads to fading at the edge of the diffracting structures in the x -direction (22-14d and e).

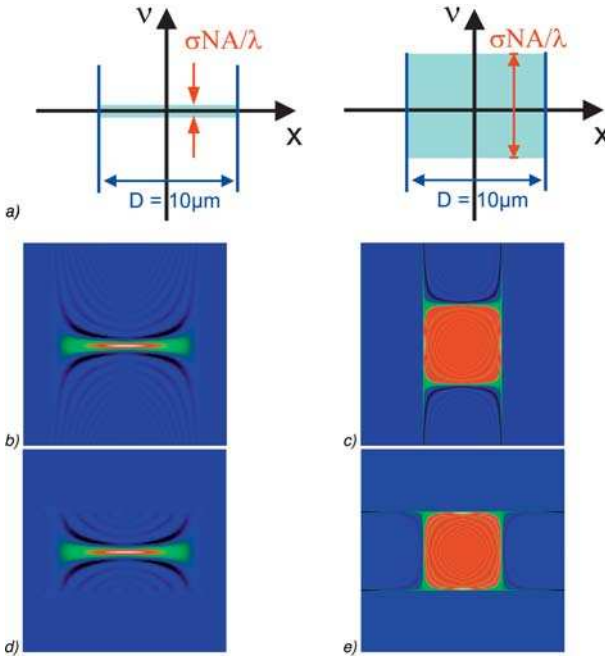


Figure 22-14: a) Illustration of the phase space coordinates for the example of a 10λ -slit under coherent illumination with $\sigma = 0.05$ and under incoherent illumination with $\sigma = 1$. Wigner distribution function of a 10λ -slit for b) coherent illumination and c) incoherent illumination; d) and e) WDF in the image plane.

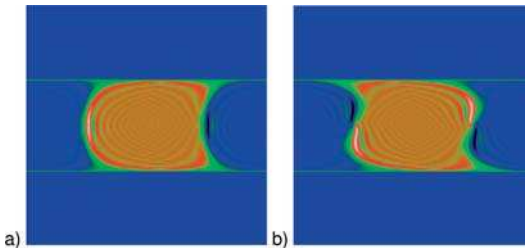


Figure 22-15: Wigner distribution function of a 10λ -slit in the image plane for incoherent illumination a) with coma of $Z_7 = \lambda/5$ and b) spherical aberration $Z_9 = \lambda/5$.

Figure 22-15 illustrates the influence of the aberrations coma (Z_7) and spherical aberration (Z_9) on the Wigner distribution function. The role of aberration becomes clearer when considering the Wigner distribution function of the point image [22-23], [22-24]. Figure 22-16 shows this for an imaging system with the aperture $NA=1$. The low-pass filtering of the spectrum in the vertical frequency direction causes the formation of the secondary maxima next to the central maximum.

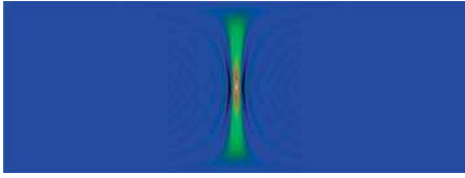


Figure 22-16: Wigner function $W(x, \nu)$ of the ideal point image.

The Wigner function of a single image point can be determined either from the amplitude distribution function or from the coherence transfer function $k(\nu_1, \nu_2)$ after coordinate transformation, $\kappa(\Delta\nu, \nu)$. As illustrated in figure 22-17, Fourier transformation of $\kappa(\Delta\nu, \nu)$ in the direction of $\Delta\nu$ gives the Wigner function of the point image $W(x, \nu)$. The Fourier inverse function to the Wigner function is the ambiguity function, thus the ambiguity function of the point image is given by a Fourier transformation of the coherence transfer function in the direction of the centre frequency ν :

$$A(\Delta\nu, \Delta x) = \int \kappa(\Delta\nu, \nu) \cdot e^{2\pi i \Delta x \cdot \nu} d\nu. \tag{22-44}$$

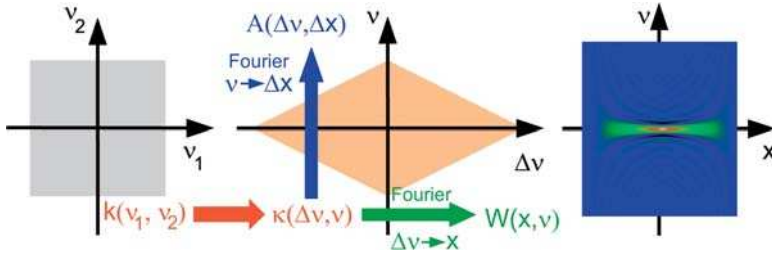


Figure 22-17: Ambiguity function $A(\Delta\nu, \Delta x)$ of an imaging system from the coherence transfer function after coordinate transformation ($\Delta\nu$ -axis pointing to the right, Δx -axis upwards).

The ambiguity function for ideal imaging thus corresponds to the Wigner function of the point image rotated through 90° and re-scaled. Figure 22-17 illustrates the coordinate transformation to centre-of-mass coordinates ν and difference coordinates $\Delta\nu$ applied to the coherence transfer function $k(\nu_1, \nu_2)$. The Fourier transformation along the vertical ν -axis gives the ambiguity function $A(\Delta x, \Delta\nu)$, which is equal to the two-fold Fourier transformation of the Wigner distribution function for the point image.

22.4.2

Propagation of the Wigner Distribution Function in Free Space

Using the Wigner distribution function, the effects of propagation in free space can be illustrated in a simple way. Figure 22-18 illustrates the free-space propagation of the Wigner distribution function for the point image in the ideal image plane and in different planes defocused to it.

In the two-dimensional discussion, the next higher aberration after defocusing is the spherical aberration. The longitudinal spherical aberration as a deviation of the ideal plane of reference over the aperture angle, describes a so-called zonal aberration of the lowest order. Therefore, the Wigner distribution function of the point image with spherical aberration looks similar to the one at defocusing but cannot be brought over the centre of the point image either for small or large angles by propagating the distribution function (figure 22-19). The S-shape of the Wigner distribution function for the projection leads always to a broadening of the intensity distribution of the point image function.

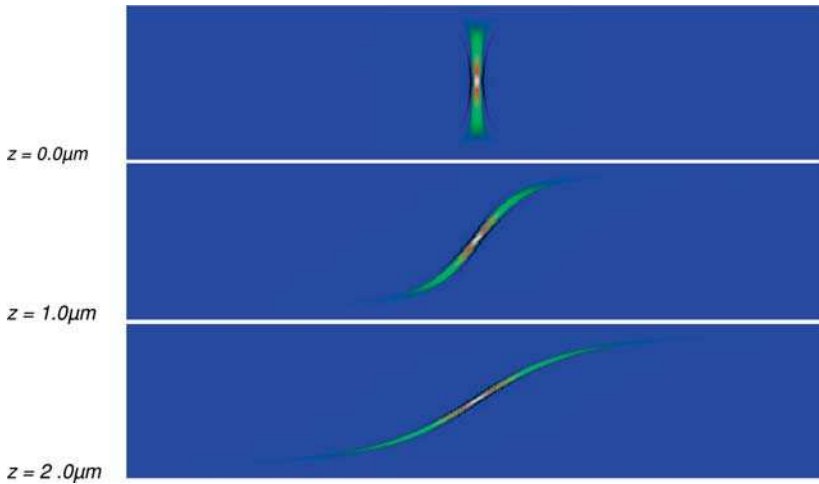


Figure 22-18: Wigner distribution functions of a slit image by free-space propagation from $z = 0 \mu\text{m}$ to $2 \mu\text{m}$.

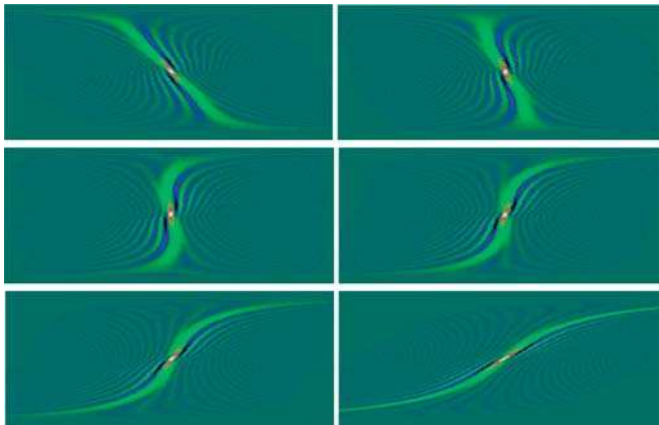


Figure 22-19: Free-space propagated Wigner distribution functions of the point image with spherical aberration $Z_9 = \lambda/5$ of $z = -1, -0.5, 0, 0.5, 1$ and $z = 2$.

Each projection of the Wigner distribution function delivers the intensity distribution in its observation plane. Some of them are shown in figure 2-20 as examples. Because of the free-space propagation, the point image is blurred.

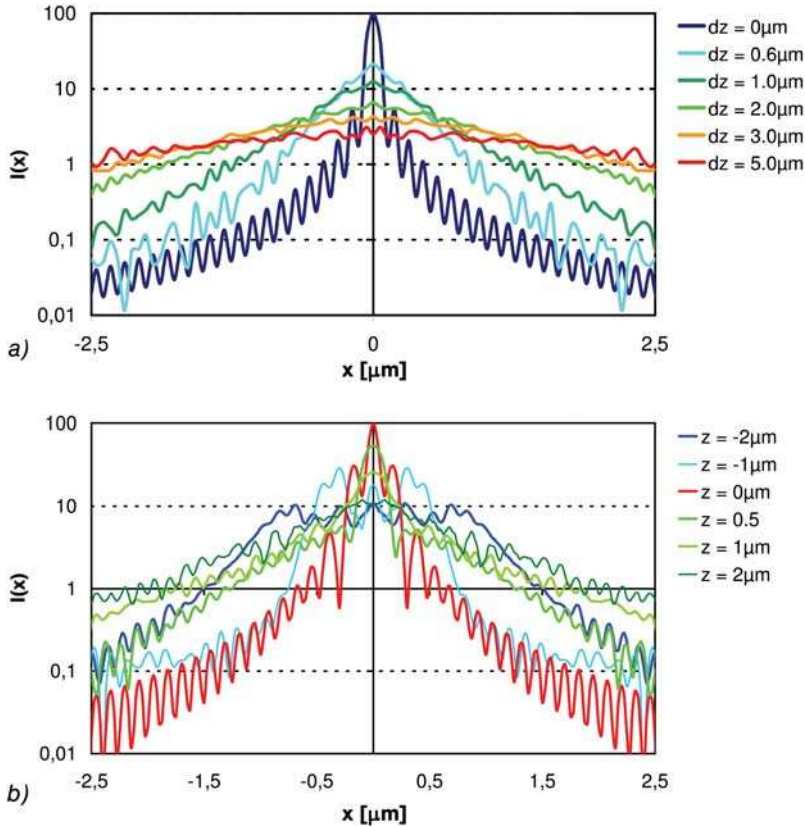


Figure 22-20: Point spread functions at a) ideal imaging and b) spherical aberration for different defocus positions.

For non-rotational-symmetric wave-front errors, the coherence transfer function is no longer rotation-invariant (see figure 22-11), therefore, the Wigner distribution function and ambiguity function of the point image are different (figure 22-21).

Figure 22-22 illustrates some Wigner distribution functions for coma of $Z_7 = \lambda/5$ and different amounts of defocus. Since the Wigner distribution function of the point image with reference to the x -axis is mirror-symmetric, with $\nu=0$, the point image reacts symmetrically to defocusing, i.e., for negative or positive defocusing by a certain amount there are identical defocused point-spread functions for coma in contrast to the behavior for spherical aberration (figure 22-23).

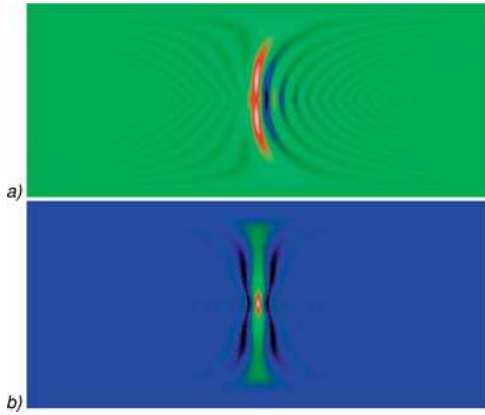


Figure 22-21: a) Wigner distribution function $W(x, \nu)$ and b) ambiguity function $A(\Delta \nu, \Delta x)$ of the point image at coma of $Z_7 = \lambda/5$.

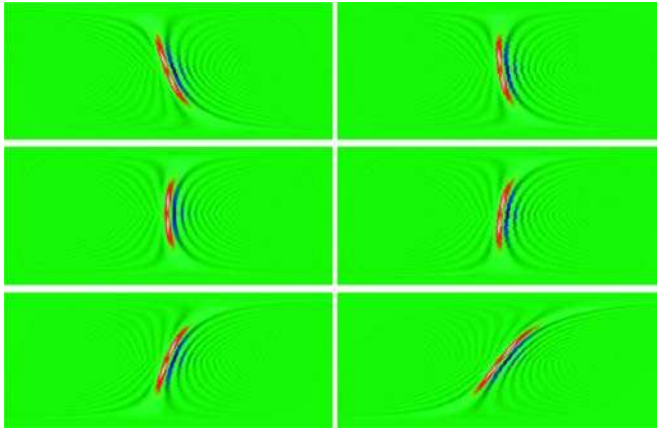


Figure 22-22: Free-space propagated Wigner distribution functions of the point image for coma of $Z_7 = \lambda/5$ from $z = -0.5, -0.2, 0, 0.2, 0.5$ and $z = 1 \mu\text{m}$.

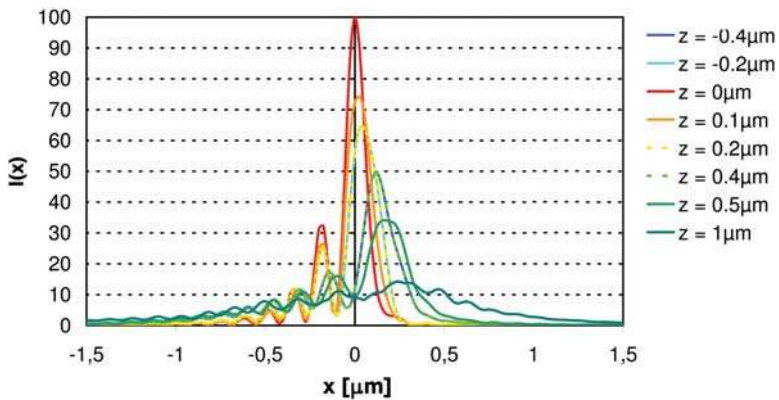


Figure 22-23: Point spread functions for coma of $Z_7 = \lambda/5$ for different defocus positions.

22.4.3

Compilation of the Transformations

Figure 22-24 shows the coherence transfer function, followed by the coherence transfer function after coordinate transformation, and the ambiguity function for a conventional imaging system and for a system with central vignetting. Figure 22-25 shows the coherence function and the Wigner distribution function for optical imaging of a grating with coherent illumination of $\sigma = 0.2$ ($NA = 1, d = 1.2\lambda$). The correlation of the functions is shown in table 22-2.

Table 22-2: Functions and their correlation in figure 22-25

Position space – Frequency space	Center-of-mass coordinates	Wigner distribution function
$(x_1, x_2) - (\nu_1, \nu_2)$	$(x, \Delta x) - (\nu, \Delta \nu)$	$(x, \nu) - (\Delta \nu, \Delta x)$
Coherence function of the light source in the object plane $\Gamma_S(x_1, x_2) = S(x_1) \oplus S^*(x_2)$	$j_S(x, \Delta x)$	$W_S(x, \nu)$
Coherence function of the illuminated objects $\Gamma_{OS}(x_1, x_2) = \Gamma_S(x_1, x_2) \cdot T^*(x_2)$	$j_{OS}(x, \Delta x)$	$W_{OS}(x, \nu)$
⇒ Fourier transformation		
Coherence function of the illuminated objects in the incident pupil $\tilde{\Gamma}_{OS}(\nu_1, \nu_2) = \tilde{\Gamma}_S(\nu_1, \nu_2) \otimes t(\nu_1) \otimes t^*(\nu_2)$	$j_{OS}(\Delta \nu, \nu)$	$A_{OS}(\Delta \nu, \Delta x)$
⇒ Multiplication with the coherence transfer function		
Coherence function in the exit pupil $\tilde{\Gamma}_B(\nu_1, \nu_2) = \tilde{\Gamma}_{OS}(\nu_1, \nu_2) \cdot k(\nu_1, \nu_2)$	$j_{k OS}(\Delta \nu, \nu)$	$A_{k OS}(\Delta \nu, \Delta x)$
⇒ Fourier transformation		
Coherence function in the image plane $\Gamma_B(x_1, x_2)$	$j_B(x, \Delta x)$	$W_B(x, \nu)$

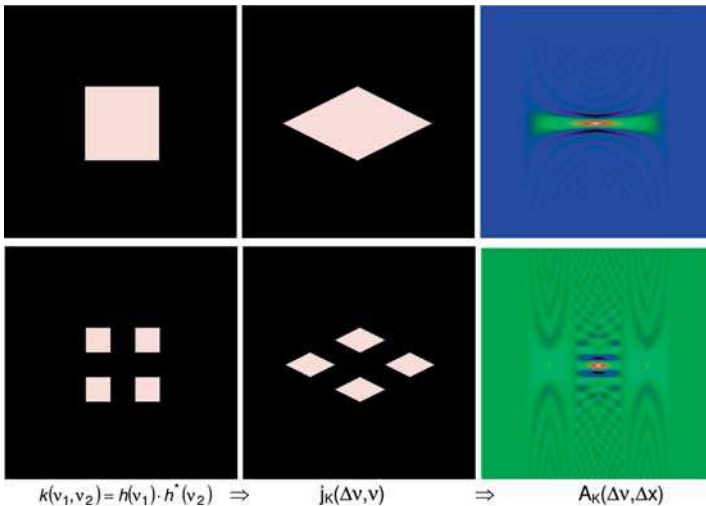


Figure 22-24: Coherence transfer function, after coordinate transformation and ambiguity function for a conventional system (top) and a system with central vignetting of 1/3 NA (bottom).

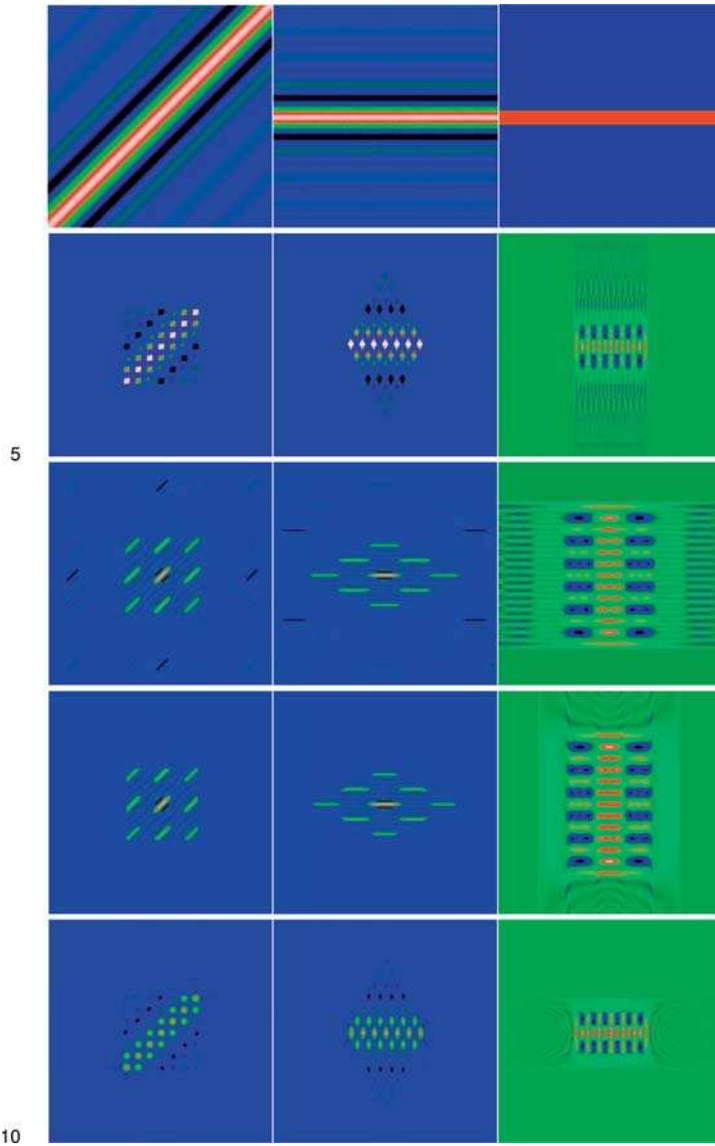


Fig 22-25: Compilation for the functions from the object plane to the image plane for the imaging of a grating with 7 periods and $\sigma=0.2$ ($d=1.2\lambda$, $NA=1$).

22.5

Optical Imaging in the Presence of Aberrations

22.5.1

Linear Systems and Classification of Aberrations

As introduced in chapter 20, the wave-front error of an optical system can be expanded into a power series according to

$$W_1(x_0, y_0; p_1, q_1) = \sum_{n,m,k,l} A_{nmkl} x_0^n y_0^m p_1^k q_1^l . \quad (22-45)$$

The expansion covers the systematic aberrations of an optical system due to design and adjustment errors. As well as these systematic errors, the optical performance is diminished by fabrication imperfections such as surface errors or inhomogeneities of the refractive index. Furthermore, in addition to the static aberrations, temporal effects such as thermal drifts, aging and fluctuations might alter the performance of optical systems over different time-scales.

While the systematic design and adjustment aberrations of optical systems are of low spatial frequency, the isoplanatic state of a position-independent transmission function is frequently affected by rapidly-varying errors such as roughness errors. The following paragraph discusses the difference between the aberrations and the different methods of consideration in the modelling and simulation of image formation.

Starting with eq. (21-38), the intensity of the optical image is given by the two-fold convolution integral:

$$I_1(x) = \iint T(x_1) \cdot H(x - x_1) \cdot \Gamma(x_1, x_2) \cdot H^*(x - x_2) \cdot T^*(x_2) dx_1 dx_2 . \quad (22-46)$$

In the special case of an infinite incoherent light source, the mutual coherence function in the object plane degenerates to a delta-function with $\Gamma(x_1, x_2) = \delta(x_1 - x_2)$, and the image intensity is given by a linear transfer of the intensities. Equation (22-46) can thus be written as a convolution between the intensity in object space and the intensity of the point-spread function:

$$I_1(x) = \int |T(x_1)|^2 |H(x - x_1)|^2 dx_1 = I(x) \otimes |H(x)|^2 . \quad (22-47)$$

For perfectly coherent illumination, the mutual coherence function is constant, $\Gamma(x_1, x_2) = 1$, for all pairs of object points x_1 and x_2 , and the image intensity is simply given by the square of the convolution of the amplitude in the object plane with the transmission function:

$$I_1(x) = |T(x) \otimes H(x)|^2 . \quad (22-48)$$

The amplitude distribution $H(x)$ is given by the Fourier transform of the complex transfer function $h(\nu)$, with the pupil function $P(\nu)$ as a real part and the wave-front aberration $W(\nu)$ as the phase error:

$$H(x) = \int h(\nu) \cdot e^{2\pi i \nu x} d\nu = \int P(\nu) \cdot e^{\frac{j2\pi}{\lambda} W(\nu)} e^{2\pi i \nu x} d\nu . \quad (22-49)$$

A necessary condition for the computation of the optical image according to eq. (22-46), (22-47) or (22-48) is the invariance of the transfer function $h(\nu)$ over a finite area. With a varying transfer function the necessary shift invariance of the linear system theory is violated and its exact calculation is no longer possible. However, aberrations can be separated and slowly varying aberrations of low frequency, which are constant at least over small regions of the field, can be treated by partially coherent imaging theory, according eq. (22-48).

Since the transfer of the amplitudes is linear, the transfer function $h(\nu)$ can be separated into several transfer functions, e.g., corresponding to different origins. A simple example is an imaging system comprising several sub-systems with intermediate images. The total system may be free from aberrations, while sub-systems may suffer from severe imaging aberrations, which cancel each other out (figure 22-26).

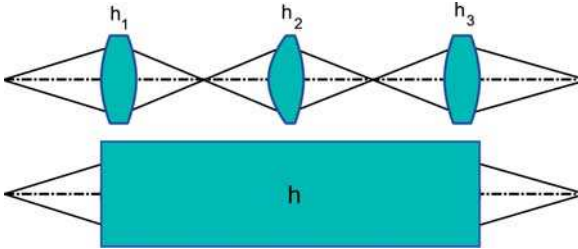


Figure 22-26: Linearity of the transfer of amplitudes, visualized for the example of a system which is composed of several sub-systems.

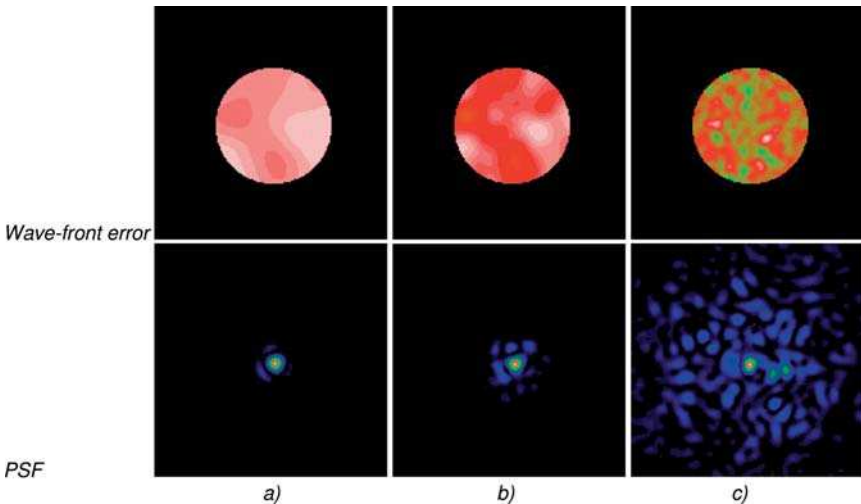


Figure 22-27: Different types of aberration and corresponding point-spread functions. a) local stationary aberrations of low frequency due to design and assembly, b) mid-spatial frequent aberrations due to, e.g., atmospheric turbulence, c) random highly frequent aberrations due to, e.g., polishing errors.

Another example is the separation of aberrations of different kinds, e.g., aberrations due to design errors and aberrations due to fabrication and assembly of the imaging system (see figure 22-27). The amplitude transfer function is in general written as a product of the transfer functions of the sub-systems:

$$h(\nu) = P(\nu) \cdot e^{i\frac{2\pi}{\lambda}W(\nu)} = \prod_i h_i(\nu) = \prod_i P_i(\nu) e^{i\frac{2\pi}{\lambda}W_i(\nu)} \quad (22-50)$$

with

$$P(\nu) = \prod_i P_i(\nu), \quad (22-51a)$$

$$W(\nu) = \sum_i W_i(\nu). \quad (22-51b)$$

The amplitudes are transferred linearly and therefore the amplitude transfer can be split into several transfer functions. The intensity transfer, however, is non-linear and the transfer function in general can no longer be separated. For example, for incoherent imaging according to eq. (22-47), the image quality may only be diminished by subsequent convolutions of the image intensity with point-spread functions of parts of optical systems or detectors. However, as will be shown below, some types of aberration may be adequately considered by the convolution of the intensities according eq. (22-47).

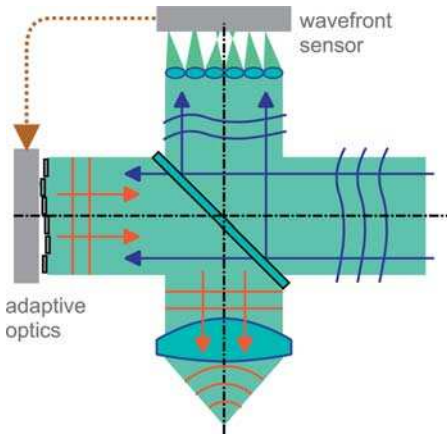


Figure 22-28: Wave-front correction with adaptive optics.

Figure 22-28 illustrates schematically an application of the linearity of amplitudes for the example of adaptive wave-front corrections, e.g., for terrestrial telescopes. The average value of the turbulence of the atmosphere leads to a randomly distributed point image. Terrestrial astronomy therefore suffered for a long time from the limited resolution. However, since the aberrations $W_{\text{atmosphere}}$ are almost indepen-

dent of the field, they can be compensated for in the pupil plane by a single inverse filter function with $W_{\text{filter}} = -W_{\text{atmosphere}}$. The effect is an almost perfect point-spread function. Due to the temporal drift of the ‘aberrations’, however, it is necessary that the correction filter is temporally variable, i.e., by means of an active device, such as an adaptive mirror.

The theoretical limit of the aberration compensation is given, finally, by the diffraction effects due to the free-space propagation between the source of the aberration (i.e., the turbulence) and the compensation filter. Although practically more difficult, the principle of compensation for aberrations is extendable to complex systems with field-dependent aberrations, as long as the aberrations may be compensated unequivocally. Complex optical systems such as, e.g., projection imaging systems or microscopes, comprise several lenses and optical surfaces. The ray bundle corresponding to different field points intersects the optical surfaces at different positions and – depending on the distance to the pupil position – with different diameter (figure 22-29). The intersection of a surface with the ray bundle to a field point is here called the effective pupil to a field point. Every surface error, caused by deformation or fabrication, in the beam path between the object and the field point will add to the resulting wave-front aberration.

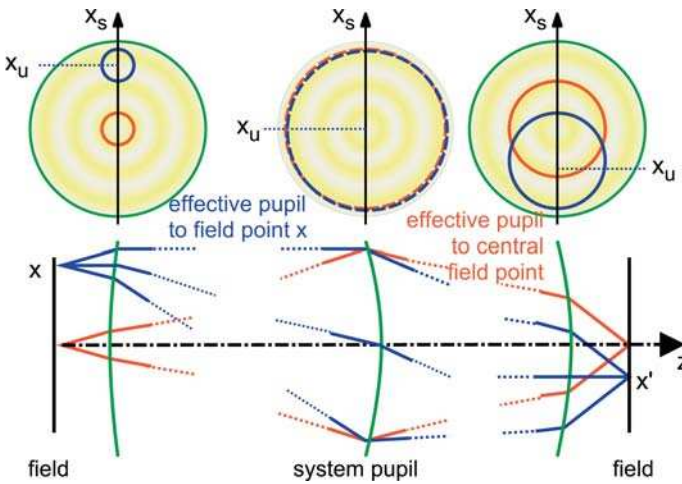


Figure 22-29: Influence of surface errors on the field dependence of aberrations.

The transfer function of the total system can be evaluated by the addition of all contributions to the wave-front error according

$$h(\nu; x) = \int \bar{h}(\nu; x; z) dz \tag{22-52}$$

with

$$\bar{h}(\nu; x; z) = \text{circ}\left(\frac{x_s - x_u(x, z)}{\rho(z)}\right) \cdot e^{\frac{2\pi i}{\lambda} W(x, z)} \tag{22-53}$$

and with the scaling

$$\nu = \frac{x_s - x_u(x, z)}{\rho(z)} \cdot \frac{NA}{\lambda} \quad (22-54)$$

with the contribution to the wave-front aberration $W(x_s; z)$ generally dependent on a lateral coordinate x_s of the optical element at z and limited by the z -dependent effective pupil diameter $\rho(z)$, decentred by the unit ray position x_u . The degree of invariance of the transfer function can be described by the correlation function for the transfer function:

$$C(x) = \frac{\int h(\nu; x') h^*(\nu; x' + x) d\nu}{\int |h(\nu; x')|^2 d\nu} \quad (22-55)$$

For small wave-front aberrations, eq. (22-55) can be simplified for each single optical element to

$$C(x; z) \sim \text{circ}\left(\frac{x_s}{\rho(z)}\right) \cdot W(x_s - x_u(x', z); z) \cdot W(x_s - x_u(x' + x, z); z) dx_s \quad (22-56)$$

The correlation of the wave-front aberration depends, however, on the range of frequencies considered in eq. (22-56). As shown in Figure 22-27, in general, the frequencies of the wave front aberration is separated into low frequency, mid-spatial frequency range and high spatial frequency range. For the discussion, an object field size of Δx is considered. The frequency range for which $C \sim 1$ is known as the low frequency aberration range, and includes errors due to design and alignment. The correlation of wave front error and consequently the correlation of the transfer function is very high and the isoplanatic condition is observed over Δx . In other words, the extension of the coherence function given by

$$\Gamma_{IS}(x_1, x_2) = \iint \Gamma_S(x'_1, x'_2) \cdot H_{LSFR}(x_1 - x'_1) \cdot H_{LSFR}(x_2 - x'_2) dx'_1 dx'_2 \quad (22-57)$$

is invariant over Δx .

For wave-front aberrations of mid-spatial frequency, the correlation of the transfer function is lower, and the coherent image can only be calculated for a reduced object field size Δx . If finally the coherence function according eq. (22-57) varies over the region of interest, the partial image has to be calculated according eq. (22-46) with space-dependent point spread functions $H(x)$ or, as shown below, by statistical methods. This is observed at wavefront errors of high spatial frequency. The correlation of the wave front errors vanishes usually for high spatial frequency errors. High spatial frequency errors are induced e.g. by surface roughness errors of optical elements and may not be considered in the coherent image.

To summarize, since the diameter, position and overlapping regions of the effective pupils in optical systems change with the position z in an optical system (see figure 22-29), errors such as surface errors on the optical surfaces cause different types of aberration:

- The size and overlapping regions of the effective pupils increase close to the pupil position of the optical system. For the ideal pupil, all effective pupils to different field points overlap perfectly. Surface errors at corresponding positions will cause almost field-independent or slowly-varying aberrations. Aberrations, which are constant over the field may be compensated for by coherent filters in the pupil plane.
- The size and overlapping regions of the effective pupils decrease at positions apart from the system-pupil positions, causing field-dependent aberrations. Compensation filters for aberrations depending on the field position must be properly placed at corresponding positions z .
- Close to the field position, even low frequency surface errors cause rapid variation of imaging aberrations over the field, especially at the position where the numerical aperture NA is the smallest (in figure 22-29 this is in the object plane).
- High frequency surface errors at any position must be described by statistical methods.

22.5.2

Random Non-stationary Aberrations: Stray Light and Flare

For the high frequency aberrations, which may change from field point to field point, an exact treatment is not possible, but fortunately is also unnecessary. For high frequency aberrations, the impact of the coherence properties of the source can be neglected and the imaging is therefore computed by the incoherent imaging according to eq. (22-47). High frequency aberrations can be considered with sufficient accuracy by their incoherent point-spread function given by the expectation values of the high-frequency transfer functions. High frequency aberrations are thus considered by statistical methods. The image intensity is therefore given by the partial coherent image I_{LF} considering low and mid-spatial frequencies, convoluted with the scattering point-spread function $\langle I_S \rangle$:

$$\begin{aligned} I(\mathbf{x}) &= I_{LF}(\mathbf{x}) \otimes \langle I_S(\mathbf{x}) \rangle \\ &= \left\{ \iint T(\mathbf{x}_1) \cdot H_{LF}(\mathbf{x} - \mathbf{x}_1) \cdot \Gamma(\mathbf{x}_1, \mathbf{x}_2) \cdot T^*(\mathbf{x}_2) \cdot H_{LF}^*(\mathbf{x} - \mathbf{x}_2) \, d\mathbf{x}_1 d\mathbf{x}_2 \right\} \otimes \langle I_S(\mathbf{x}) \rangle \end{aligned} \quad (22-58)$$

with

$$\langle I_S(\mathbf{x}) \rangle = \iiint |s(\mathbf{v}_q)|^2 \cdot \left\langle e^{i[\Phi(\mathbf{v}_1 + \mathbf{v}_q) - \Phi(\mathbf{v}_2 + \mathbf{v}_q)]} \right\rangle \cdot e^{i2\pi\mathbf{x} \cdot (\mathbf{v}_1 - \mathbf{v}_2)} \, d\mathbf{v}_1 d\mathbf{v}_2 d\mathbf{v}_q \quad (22-59)$$

and the phase error given by the high frequency wave-front error $\Phi(\mathbf{v}) = 2\pi/\lambda \cdot W_{HF}(\mathbf{v})$. The scattering cross coefficient TCC_S corresponds to a weighted correlation of the surface errors W_{HF} :

$$TCC_s(\nu_1, \nu_2) = \int |s(\nu_q)|^2 \cdot e^{i[\Phi(\nu_1+\nu_q) - \Phi(\nu_2+\nu_q)]} \cdot d\nu_q . \quad (22-60)$$

Equations (22-59) and (22-60) can be further simplified. For light sources larger than the typical correlation length of the phase error $\Phi(\nu)$, the weighting with the light source can be neglected and eq. (22-59) can be simplified:

$$\langle I_s(x) \rangle = \iint g(\nu_1 - \nu_2) \cdot e^{i2\pi x \cdot (\nu_1 - \nu_2)} d\nu_1 d\nu_2 = \int g(\Delta\nu) \cdot e^{i2\pi x \cdot \Delta\nu} d\Delta\nu = G(x) \quad (22-61)$$

with the ensemble average of the random mutual phase differences $g(\nu_1 - \nu_2)$

$$g(\nu_1 - \nu_2) = \langle e^{i[\Phi(\nu_1) - \Phi(\nu_2)]} \rangle = \iint e^{i[\Phi(\nu_1) - \Phi(\nu_2)]} d\nu_1 d\nu_2 . \quad (22-62)$$

For weak surface roughness, the random mutual phase difference $g(\Delta\nu)$ is approximated by a series expansion with

$$e^{i[\Phi(\nu_1) - \Phi(\nu_2)]} = \sum_n \sum_m \frac{(i)^{n-m}}{n!m!} \Phi^m(\nu_1) \Phi^n(\nu_2) . \quad (22-63)$$

In second order, the random mutual phase differences $g(\nu_1 - \nu_2)$ can be written as

$$\begin{aligned} g(\nu_1 - \nu_2) &= 1 + \iint \left[i\Phi(\nu_2) - i\Phi(\nu_1) + \Phi(\nu_1)\Phi(\nu_2) - \frac{\Phi^2(\nu_1) + \Phi^2(\nu_2)}{2} \right] d\nu_1 d\nu_2 \\ &= 1 + \sigma^2 \kappa(\nu_1 - \nu_2) - \sigma^2 \end{aligned} \quad (22-64)$$

with the correlation function $\kappa(\nu_1 - \nu_2)$ of the phase fluctuations

$$\kappa(\nu_1 - \nu_2) = \frac{1}{\sigma^2} \iint \Phi(\nu_1)\Phi(\nu_2) d\nu_1 d\nu_2 \quad (22-65)$$

and the mean square average σ of the phase fluctuations W_{HF}

$$\sigma = \int \Phi^2(\nu) d\nu = \left(\frac{2\pi}{\lambda} \right)^2 \int W_{HF}^2(\nu) d\nu . \quad (22-66)$$

Fourier-transformation of $g(\nu_1 - \nu_2)$ gives the scattering intensity point spread function

$$G(x) = (1 - \sigma^2) \delta(x) + \frac{\sigma^2}{\pi A^2} K(x) \quad (22-67)$$

with $K(x)$ given by the Fourier transform of $\kappa(\nu_1 - \nu_2)$. Surface errors of different spatial frequencies ν will certainly give rise to different correlation lengths A and thus to different ranges of the scattered light due to $K(x)$. Therefore, the scattering point-spread function is frequently separated into different ranges of surface error frequencies:

$$G(x) = \left[1 - \sum_k \sigma_k^2 \right] \delta(x) + \sum_k \frac{\sigma_k^2}{\pi \Lambda_k^2} K(x). \quad (22-68)$$

Example: with a Gaussian correlation function, it follows that

$$\kappa(\nu) = e^{-\pi^2 \Lambda^2 \nu^2}. \quad (22-69)$$

Fourier-transformation gives

$$G(x) = (1 - \sigma^2) \delta(x) + \frac{\sigma^2}{\pi \Lambda^2} e^{-\frac{x^2}{\Lambda^2}}. \quad (22-70)$$

Figure 22-30 illustrates the superposition of different scattering point-spread functions to form $G(x)$ with Gaussian surface roughness errors of the correlation lengths $\Lambda = 1, 5$ and 25 and each $\sigma = 0.1$.

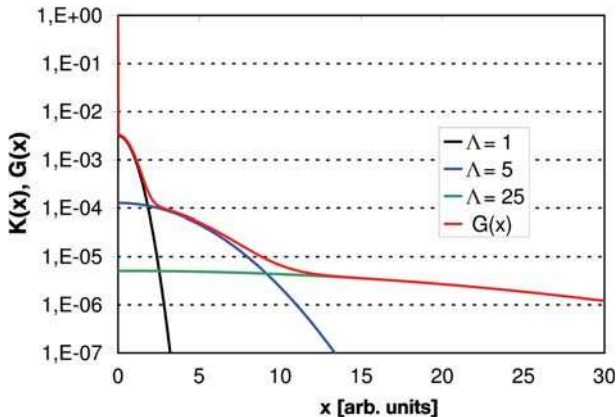


Figure 22-30: Range separation of surface roughness errors.

The emphasis of this chapter is mainly to point out the transition between coherent imaging with aberrations and incoherent consideration of expectation values of high frequency or fluctuating phase errors. For the further treatment of scattering in optical imaging as well as the transition to real surface roughness errors, the reader is referred to the literature, e.g., [22-25], [22-26], [22-27] and [22-28] and volume 6.

22.6

Literature

- 22-1 M. P. Rimmer and B.R. Irving, Calculation of partially-coherent imagery, *SPIE* **237**, 150–157 (1980).
- 22-2 E.C. Kintner, Method for the calculation of partially coherent imagery, *Appl. Opt.* **17**, 2747–2753 (1978).
- 22-3 H.H. Hopkins, On the diffraction theory of optical imaging, *Proc. Royal Soc. of London* **217A**, pp. 408–432 (1953).
- 22-4 Ch. Hofmann, *Die optische Abbildung* (Akademische Verlagsgesellschaft, Leipzig, 1980).
- 22-5 S.T. Yang, R.L. Hsieh, Y.H. Lee, R.F.W. Pease, G. Owen, Effect of central obscuration on image formation in projection lithography, *SPIE* **1264**, 477–485 (1990).
- 22-6 S. Kessler, G. Nietzsche, J. Hebenstreit, Defocused imaging of an amplitude sine wave object under partially coherent illumination with an annular source, *J. Mod. Opt.* **35**, 177–186 (1988).
- 22-7 E.H. Linfoot, *Qualitätsbewertung optischer Bilder*, Vieweg, Braunschweig 1960.
- 22-8 Ch. Hofmann, *Die Bewertung optischer Systeme*, *Fortschr. Physik* **27**, 595–639 (1979).
- 22-9 M. Burkhardt, A. Yen, Ch. Proglar, G. Wells, Illuminator Design for the printing of regular contact patterns, *Microelectronic Engineering* **41/42**, 91–96 (1998).
- 22-10 A.E. Rosenbluth, S. Bukofsky, R.N. Singh, A. Wong Optimum mask and source patterns to print a given shape, *SPIE Vol.* **4346**, (2001).
- 22-11 A.E. Rosenbluth et al., Optimum mask and source patterns to print a given shape, *J. Microlithogr. Microfabr. Microsyst.* **1**, 13–30 (2002).
- 22-12 Y. Granik, Source optimization for image fidelity and throughput, *J. Microlithogr. Microfab. Microsyst.* **3**, 509–522 (2004).
- 22-13 F.M. Schellenberg, Resolution enhancement techniques in optical Lithography, *SPIE Milestone Series Vol. MS 178* (2004).
- 22-14 J. B. DeVelis and G.B. Parrent, Transfer function for cascaded optical systems, *J. Opt. Soc. Am.* **57**, 1486–1490 (1967).
- 22-15 D. Nyssonen Grimes and B.J. Thompson, Two-point resolution with partially coherent Light, *J. Opt. Soc. Am.* **57**, 1330–1334 (1967).
- 22-16 R.J. Becherer and G.B. Parrent, Nonlinearity in optical imaging systems, *J. Opt. Soc. Am.* **57**, 1479–1486 (1967).
- 22-17 R.E. Swing and J.R. Clay, Ambiguity of the transfer function with partially coherent illumination, *J. Opt. Soc. Am.* **57**, 1180–1189 (1967).
- 22-18 G.O. Reynolds and J.B. DeVelis, Review of optical coherence effects in instrument design; parts I, II, and III, *Opt. Engin.* **20**, SR-084–SR-095; SR-124–SR-131; SR-166–SR-173 (1981).
- 22-19 H.O. Bartelt, K.-H. Brenner, A.W. Lohmann, The Wigner distribution function and its optical production, *Opt. Commun.* **32**, 32–38 (1980).
- 22-20 M.J. Bastiaans, Application of the Wigner distribution function to partially coherent light, *J. Opt. Soc. Am. A* **3**, 1227–1238 (1986).
- 22-21 K.-H. Brenner, *Phasenraumdarstellungen in Optik und Signalverarbeitung*, Dissertation, Erlangen 1983.
- 22-22 K.-H. Brenner and A.W. Lohmann, Wigner Distribution Function display of complex signals, *Opt. Commun.* **42**, 310–314 (1982).
- 22-23 M.J. Bastiaans and P.G.J. van de Mortel, Wigner distribution function of a circular aperture, *J. Opt. Soc. Am. A* **13**, 1698–1703 (1996).
- 22-24 D. Dragoman, Wigner distribution function applied to third order aberrations, *Appl. Opt.* **34**, 161–168 (1996).
- 22-25 R.N. Youngworth and B.D. Stone, Simple estimate for the effects of mid-spatial frequency surface errors on image quality, *Appl. Opt.* **39**, 2198–2209 (2000).
- 22-26 J.E. Harvey, K.L. Lewotsky and A. Kotha, Performance predictions of a Schwarzschild imaging microscope for soft x-ray applications, *Opt. Eng.* **38**, 2423–2436 (1996).
- 22-27 R.J. Noll, Effect of Mid- and high-spatial frequencies on the optical performance, *Opt. Eng.* **18**, 137–142 (1979).
- 22-28 E.L. Church, H.A. Jenkinson, and J.M. Zavada, Relationship between surface scattering and microtopographic features, *Opt. Eng.* **18**, 125–136 (1979).

23

Three-dimensional Imaging

- 23.1 Introduction 320
- 23.2 The Ewald Sphere and the Generalized Pupil 321
 - 23.2.1 The Ewald Sphere 321
 - 23.2.2 The Generalized Aperture and the Three-dimensional Point-spread Function 322
- 23.3 The Three-dimensional Transfer Function 327
 - 23.3.1 Born Approximation and the Laue Equation 327
 - 23.3.2 Dändliker's Representation and the Shape of the Three-dimensional Transfer Function 330
 - 23.3.3 Resolution, Depth Resolution and Depth of Focus 335
 - 23.3.4 3D-Transfer Functions in Microscopy 338
 - 23.3.5 Magnification and a Comment on Absolute Instruments 340
- 23.4 Selected Examples of the Three-Dimensional Transfer Function 343
 - 23.4.1 Transfer Function for Incoherent Imaging with $\sigma=1$ 343
 - 23.4.2 Partial Coherent Image Examples 344
 - 23.4.3 'Tayloring' of the 3D-Transfer Function 346
 - 23.4.5 Influence of Aberrations 351
- 23.5 Literature 352

23.1

Introduction

In chapters 20–22 the two-dimensional description of imaging for planar objects and planar images was considered. Now the question arises as to how three-dimensional objects are imaged and how the three-dimensional aerial image appears. The three-dimensional image is of special interest because most objects, such as biological tissue or views of landscapes, are three-dimensional. On the other hand, receptors or detectors are conventionally only two-dimensional and thus only two-dimensional images can be recorded, while the aerial image as interference phenomenon is generally three-dimensional. It is therefore worthwhile to gain a full understanding of the three-dimensional imaging process. The full description of three-dimensional imaging was initiated by E. Wolf [23-1], R. Dändliker and R. Weiss [23-2], and A. Lohmann [23-3] and was finally introduced by N. Streibl [23-4, 23-5] in 1984. It has become very popular, mainly within the field of microscopy [23-6], [23-7], [23-8], [23-9]. The three-dimensional description of imaging is not only the most modern theoretical description of image formation, but also gives a visual explanation of lateral as well as depth resolution.

The three-dimensional description by Streibl is based on much previous work. Already by 1926 Berek had contested Abbe's theory which stated that the interference patterns in the image space would be principally afocal in nature and thus the image plane could not be determined under coherent illumination [23-10]. So, even with a planar object, a three-dimensional interference pattern is generated and thus a three-dimensional aerial image. The general question of longitudinal image formation was also discussed over a long period [23-11], [23-12]. For an incoherent image, image formation can be described as a convolution of the object with the point-spread function which, according to McCutchen, can be obtained by using the generalized pupil [23-13]. The generalized pupil represents a segment of the Ewald sphere. With the Ewald sphere, a three-dimensional discussion of diffraction and image formation becomes straightforward. By varying the incident angles, the spectrum of the transferred object frequencies can be determined [23-2], [23-3] and the three-dimensional transfer function can be derived [23-4, 23-5], [23-14].

The discussion of three-dimensional image formation is a very useful and concrete method of gaining an understanding of the effects of optical imaging. Furthermore, it makes it possible to influence the imaging of certain object details by the careful design of the transfer function, depending on the object spectrum. The goals of such a description, however, may differ greatly. In microscopy the axial resolution should be increased as much as possible, while for lithographical imaging the extent of the image of a planar object in the axial direction should be as large as possible (large depth of focus). This chapter describes the most important features of three-dimensional image formation and gives some examples of the use of images – as was done by E. Abbe over 100 years ago with gratings. Three-dimensional imaging properties are based on consideration of the three-dimensional spectrum representation, introduced in section 17.5. Since this illustrative method is not widely known, the Ewald sphere and the Born approximation are repeated as an introduction.

23.2
The Ewald Sphere and the Generalized Pupil

23.2.1
The Ewald Sphere

As shown in chapter 17, any field amplitude distribution may be expanded into plane waves, or, in general, stationary field distributions $U(x,y,z)$ can be described as a three-dimensional Fourier transformation of the spatial frequency spectrum $u_{3D}(v_x, v_y, v_z)$:

$$U(x, y, z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u_{3D}(v_x, v_y, v_z) e^{2\pi i(xv_x + yv_y + zv_z)} dv_x dv_y dv_z . \tag{23-1}$$

Propagating solutions of equation (23-1) are plane waves, while evanescent waves can be neglected in almost any case of image formation. Propagating solutions or plane waves are given in a space domain by its spatial periodical field distribution with

$$U(x, y, z) = A \cdot e^{-i2\pi(v_x x + v_y y + v_z z)} . \tag{23-2}$$

However, only the real part of the complex field has any physical meaning. Considering an instantaneous shot of the plane wave at the time $t=t_0$, wave peaks and troughs can be observed (figure 23-1). An alternative description of a plane wave can be given by its spatial frequency vector \vec{v} . For frequency vectors \vec{v} of the three-dimensional spectrum of propagating fields – as a consequence of the wave equation – Ewald’s relation is valid,

$$\vec{p} = \lambda \vec{v} = n \vec{s} = \nabla W \tag{23-3}$$

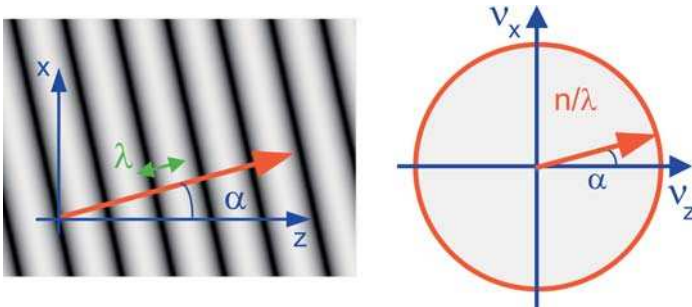


Figure 23-1: Description of a monochromatic planar wave in the spatial and in the frequency domain.

with the optical direction cosine \bar{p} , the refractive index n , the tangential vector \bar{s} to the light ray and the wave front W . The spatial frequency vector \bar{v} of length n/λ_0 describes the propagation direction of a plane wave and is perpendicular to the wavefronts. Since for all propagating monochromatic wave fields all spatial frequency vectors have the same length n/λ_0 they all terminate on a spherical surface, the *Ewald sphere*. The Ewald sphere is therefore equivalent to the *transfer function of free space*. Particularly for the z component of the spatial frequency vector, it follows thus from eq. (23-3)

$$v_z = \pm \sqrt{\left(\frac{n}{\lambda}\right)^2 - v_x^2 - v_y^2}. \quad (23-4)$$

For the discussion of the two-dimensional Fourier theory of optical imaging in section 21 the frequency vectors have been described only by the two transverse components v_x and v_y , since the v_z component is determined by Ewald's equation (23-4). In particular, if the propagation direction of the planar waves with reference to the z axis is unique the descriptions of the two or three-dimensional frequency spectrum, respectively, can be changed arbitrarily by using eq. (23-4). For a positive propagation direction we obtain:

$$U(x, y, z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u_{2D}(v_x, v_y; z=0) e^{2\pi i \left(x v_x + y v_y + z \sqrt{\left(\frac{n}{\lambda}\right)^2 - v_x^2 - v_y^2} \right)} dv_x dv_y \quad (23-5)$$

with

$$u_{2D}(v_x, v_y; z=0) = u_{3D}(v_x, v_y, v_z) / \cos \alpha. \quad (23-6)$$

23.2.2

The Generalized Aperture and the Three-dimensional Point-spread Function

From the three-dimensional frequency spectrum the three-dimensional scalar point response function of an optical system can be calculated directly, according to eq. (23-1). The maximum transverse frequency of the spectrum is given by the aperture angle α of the optical system. The three-dimensional frequency spectrum is thus given by a segment of the Ewald sphere. According to McCutchen this segment of the Ewald sphere is called the *generalized aperture* (figure 23-2) [23-13]. The discussion of the frequency band-width yields simple estimation formulas for the resolution of optical imaging. The distribution of the three-dimensional point image function perpendicular to the z -axis is given by the Fourier transform of the circular opening – the Airy disk:

$$I_{\text{PSF}}(r, z=0) = \left(\frac{J_1\left(2\pi \frac{NA}{\lambda} r\right)}{\pi \frac{NA}{\lambda} r} \right)^2. \quad (23-7)$$

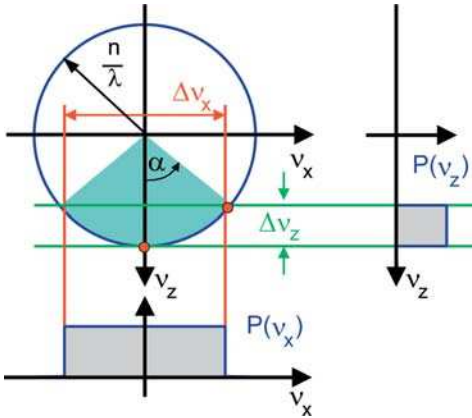


Figure 23-2: Generalized aperture $P(v_x, v_z)$ as a frequency spectrum limited by the aperture angle on the Ewald sphere.

The distribution along the z -axis is given by the Fourier transform of the rectangular function, i.e., the sinc function (normalized):

$$I_{PSF}(r = 0, z) = \left| \text{sinc} \left[\frac{(n - n \cos \alpha) \cdot z}{\lambda} \right] \right|^2. \quad (23-8)$$

The axial extent of the ideal three-dimensional point image is given by the inverse maximum spatial frequency band-width in v_z , and McCutchen's equation for depth of focus is then obtained:

$$\delta z = \frac{1}{\Delta v_z} = \frac{\lambda}{n - n \cos \alpha} = \frac{\lambda}{n - \sqrt{n^2 - NA^2}} \approx \frac{2n \cdot \lambda}{NA^2}. \quad (23-9)$$

McCutchen's equation (23-9) had been given by Berek and provides an approximation for the depth of focus of the optical image [23-10], [23-13]. However, as shown later it can be used only in a limited way. With equation (23-9) the z component is frequently given in the Rayleigh unit with a Rayleigh unit RU corresponding to defocussing δz by a phase $\sim \lambda/4$, thus with eq. (23-9)

$$RU = \frac{1}{4} \frac{\lambda}{n - n \cos \theta} \approx \frac{n \cdot \lambda}{2NA^2}. \quad (23-10)$$

Figure 23-3 compares the three-dimensional point-spread functions in cross-sections in the x - y plane and in the x - z plane for apertures from $NA = 0.4, 0.6, 0.8$ and 1.0 . The amplitude of the point-spread function is modulated in z in accordance with the decentered position of the generalized aperture with respect to the v_x -axis. The amplitude distribution through the focus is illustrated in figure 23-3b. For $NA = 1$ the generalized aperture is given by a half-sphere, and the phase anomaly through the focus becomes visible [23-15]. For a perfect spherical wave, the full

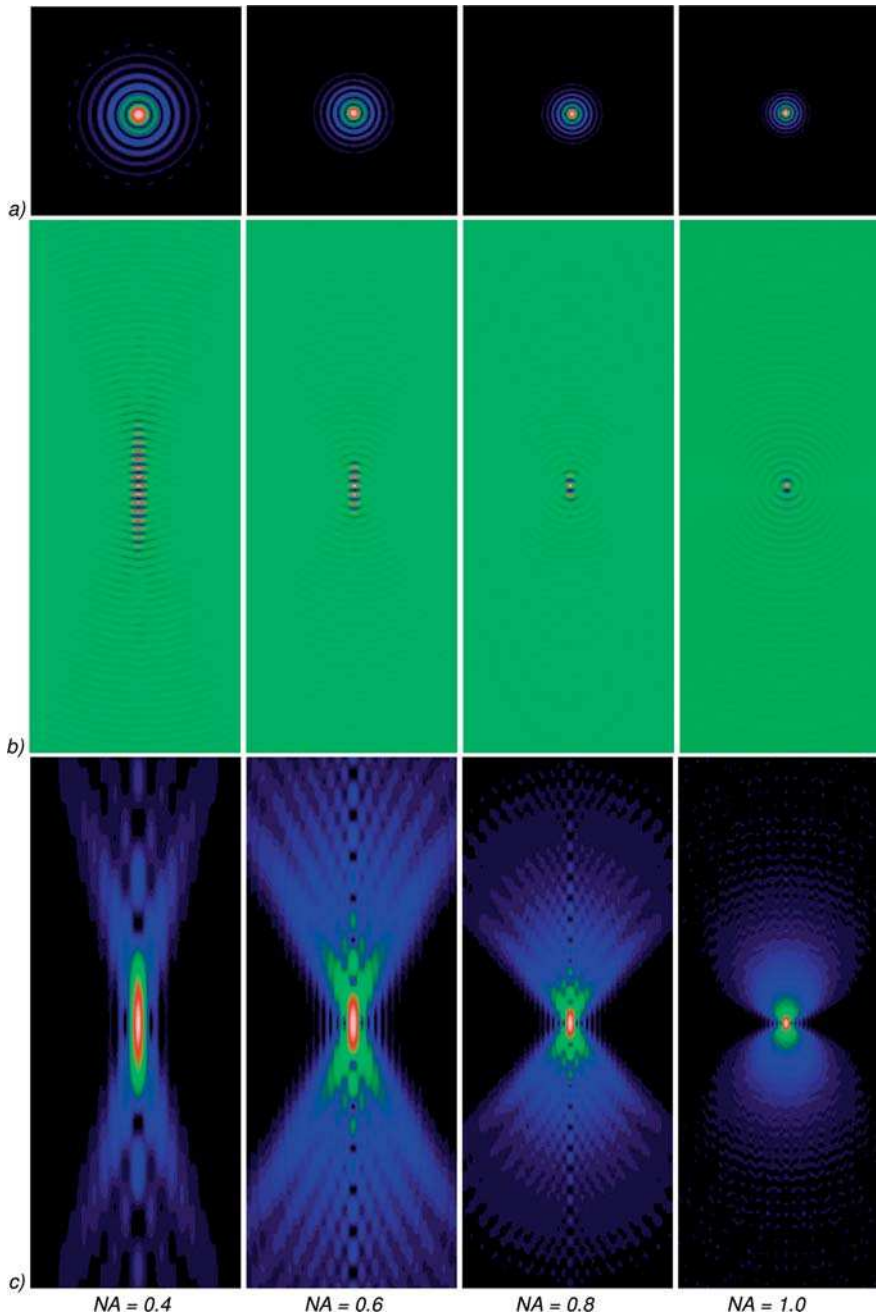


Figure 23-3: a) x - y -scan of point-spread function; b) x - z -scan of the amplitude distribution; c) x - z -scan of the intensity of three-dimensional point-spread function at $NA=0.4$, $NA=0.6$, $NA=0.8$ and $NA=1.0$ with the z -axis pointing downwards (intensity on logarithmic scale).

Ewald sphere, as a frequency spectrum, is required. Figure. 23-4 illustrates the three-dimensional point-spread function with intensity scans perpendicular to the z -axis $I(r; z = 0)$ and along the z -axis $I(r = 0; z)$. While the lateral intensity distribution scales linearly with the numerical aperture NA , the non-linear behavior of the Intensity through the focus becomes visible especially at the examples for $NA = 0.5$ and $NA = 1$. In figure 23-5 the point-spread functions with spherical aberration (Z_9, Z_{16}) and coma (Z_7, Z_{14}) are shown. Symmetrical aberrations of higher order than defocus Z_4 break the longitudinal symmetry of the 3D-PSF, while asymmetrical aberrations, like coma, break the transverse symmetry. For coma of lowest order Z_7 , the intensity through the focus is bent to the shape of a banana, therefore it was also called “bananicity” in optical lithography.

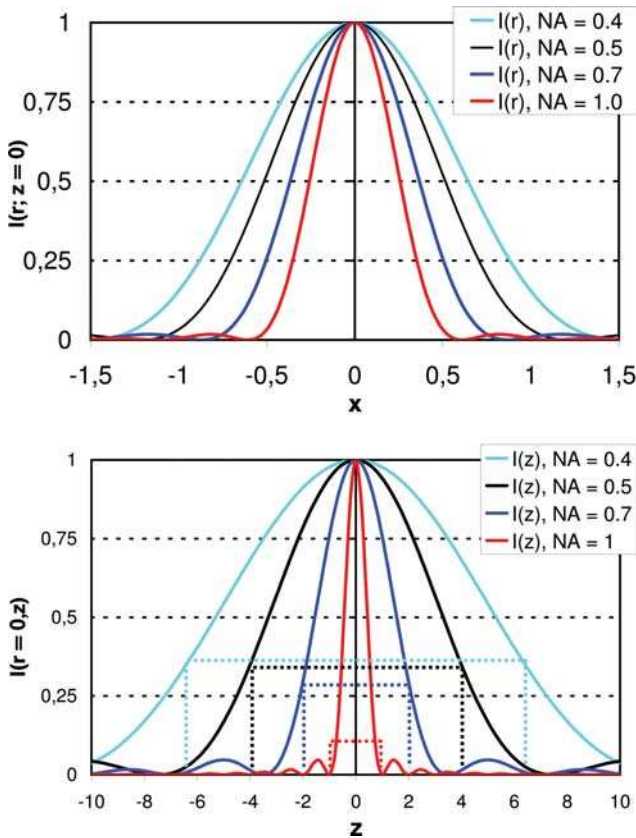


Figure 23-4: a) Airy disk $I(x; z = 0)$; and b) intensity distribution $I(x = 0; z)$ ($\lambda = 1$; z -coordinates in units of λ). The dashed boxes show the depth of focus in a parabolic approximation according to McCutchen.

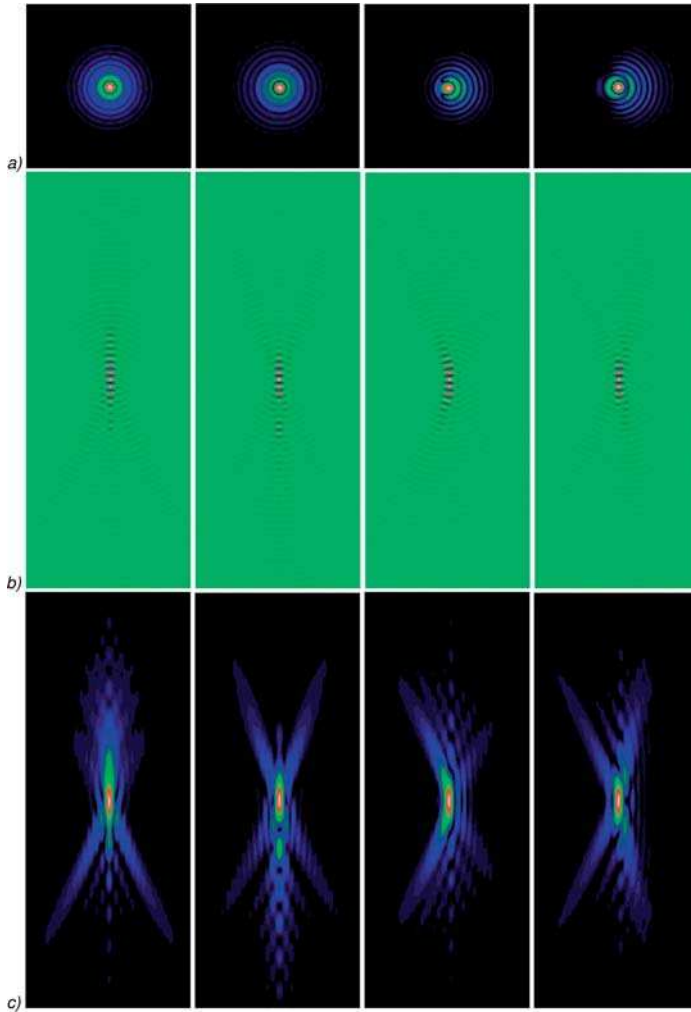


Figure 23-5: Amplitude and Intensity scans of the three-dimensional scalar point-spread functions at $NA = 0.6$ with spherical aberration $Z_9 = \lambda/5$ and $Z_{16} = \lambda/5$ and coma $Z_7 = \lambda/5$ and $Z_{14} = \lambda/5$; a) cross-section x-y-plane; b) amplitude in x-z-plane; c) intensity cross-section through the x-z plane.

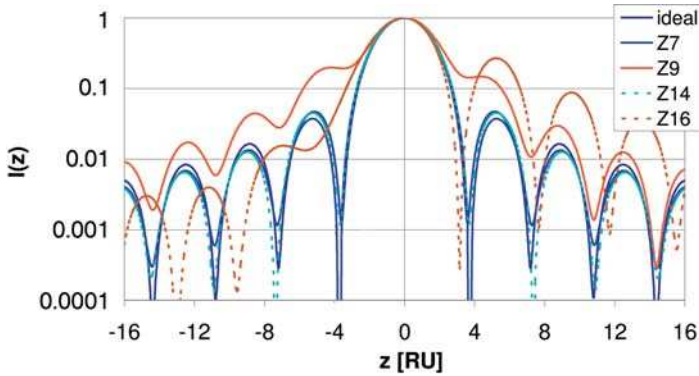


Figure 23-6: Intensity through the focus $I(z)$ on the optical axis for different aberrations of figure 23-6 ($NA = 0.6$, aberration level $\lambda/5$ each).

The axial intensities for different aberrations are compared in figure 23-6 with the axial intensity of the ideal imaging system. Since asymmetrical wave-front aberrations like coma Z_7 and Z_{14} have a point-symmetrical behaviour with respect to the optical axis, the axial intensity distribution $I(z)$ is symmetrical with respect to the origin. For symmetrical wave-front aberrations, the axial intensity is in general not symmetrical with respect to the ideal image plane. Only in the paraxial approximation is the axial intensity distribution also symmetrical [23-15]. The generalized aperture is further applied to annular apertures in chapter 25.2.

23.3

The Three-dimensional Transfer Function

23.3.1

Born Approximation and the Laue Equation

In the two-dimensional Fourier theory of optical imaging, as described in chapters 20–22, the diffracted spectrum is given by the convolution of the incident spectrum with the object spectrum. As it turns out in the three-dimensional treatment of optical imaging, the reason for this is the linearity of diffraction in frequency space, e.g., approximated by the Born approximation of first order (see also chapter 18). According to the latter, the scattered spectrum u_s is given by the convolution of the incident spectrum u_i with the three-dimensional frequency spectrum of the object:

$$u_s(\nu_x, \nu_y, \nu_z) \sim f(\nu_x, \nu_y, \nu_z) \otimes u_i(\nu_x, \nu_y, \nu_z) . \quad (23-11)$$

In physical terms this means that the light is only diffracted, refracted or scattered once by the object, i.e., the light is deflected only by a single interaction. For an incident planar wave eq. (23-11) is reduced such that the spectrum of the scattered wave is given by the object frequency spectrum shifted by the frequency of the incident wave:

$$u_s(\vec{\nu}_s) \sim f(\vec{\nu}_s - \vec{\nu}_i) . \quad (23-12)$$

A consequence of eq. (23-11) is the well-known Laue equation. Only those frequencies $\vec{\gamma}$ are transferred from the object spectrum to the scattered wave, for which the the Laue equation is satisfied:

$$\vec{\nu}_s - \vec{\nu}_i = \vec{\gamma} \quad (23-13)$$

with $\vec{\nu}_s$ being the diffracted and $\vec{\nu}_i$ being the incident frequency vector. The Laue equation results from the Born approximation due to the assumption of single scattering (= single momentum transfer). According to the Laue equation the scattered wave vector $\vec{\nu}_s$ can be determined from the sum of an incident wave vector $\vec{\nu}_i$ and a frequency vector $\vec{\gamma}$ from the scattering object. In the far-field approximation, i.e., if evanescent waves can be neglected, the transfer function of free space has also to be considered. This means that only the frequencies $\vec{\nu}_s$ on the Ewald sphere can propagate over distances large in comparison to the wavelength λ . Figure 23-7 illustrates this in a graphical way. The propagating contributions to the scattered field are limited to vectors on the Ewald sphere. According to the Laue equation (23-13), light is scattered in those directions for which the spatial frequency spectrum of the scattering object has non-vanishing contributions on the Ewald sphere with center $-\vec{\nu}_i$.

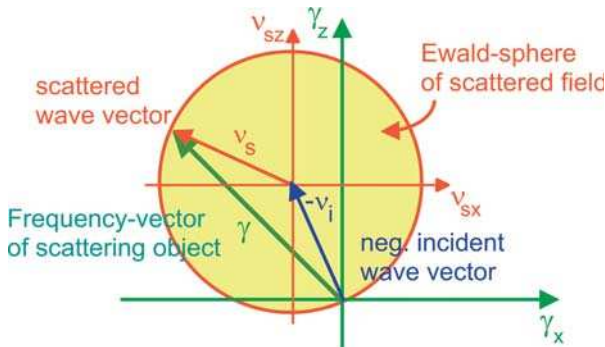


Figure 23-7: Laue construction for determining a diffracted, scattered or refracted wave vector.

For periodical objects like crystal lattices, discrete points in the spatial frequency space which are determined by the inverse lattice vector (see solid state physics literature and figure 23-8) are obtained. According to Huygens principle every lattice point diffracts light in all directions. Due to the periodicity of the light field and the grating, there is constructive interference of the diffracted partial waves in certain directions only, while there is destructive interference in certain other directions. The conditional equation for the directions of constructive interference is the Laue equation (23-13) assuming “conservation” of spatial frequencies.

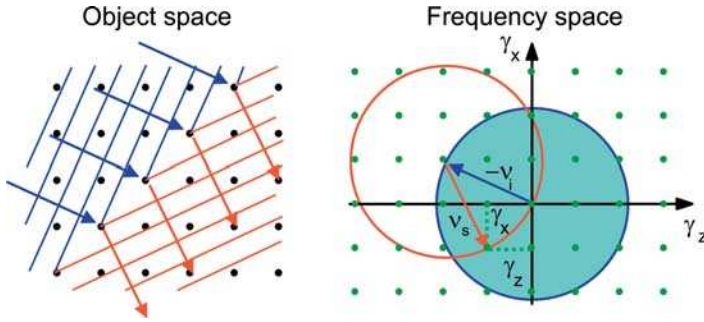


Figure 23-8: Lattice scattering and Laue construction with the Ewald sphere.

Plane diffraction gratings are periodical structures, too. Assuming a planar, infinitely thin grating in the plane $z=0$, with an infinite expansion in the y -direction, the spatial frequency spectrum of the grating in the ν_x -direction is the one-dimensional frequency spectrum of the lattice and in the ν_y -direction it is a delta function $\delta(\nu_y)$, while in the z -direction the frequency spectrum is infinitely extended (figure 23-9; for convenience the symbol ν is applied to all frequencies).

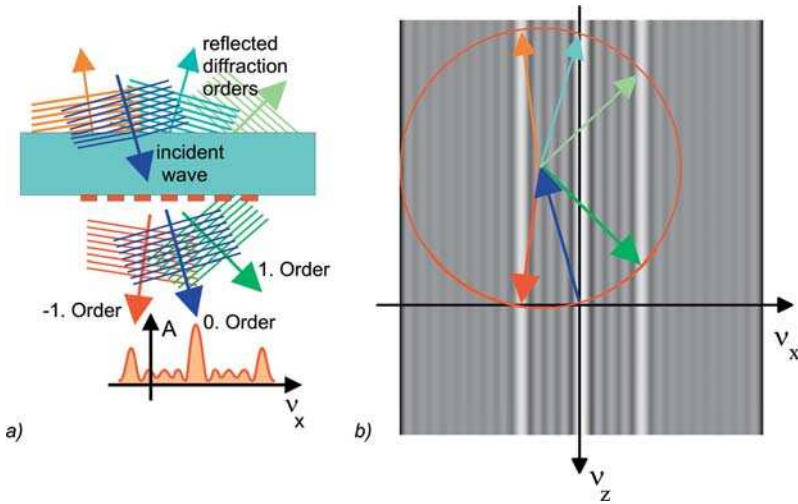


Figure 23-9: a) Diffraction grating; b) spatial frequency spectrum of the grating and diffraction orders as frequencies, where the spatial frequency spectrum of the grating does not vanish.

Again, according to the Laue equation the difference vector of the incident and diffracted wave has to be equal to a grating vector $\vec{\gamma}$. The graphical illustration of this is given in figure 23-9b. For a planar linear grating the Laue vector equation is reduced to:

$$\nu_x = \nu_{ix} + \frac{m}{d}, \quad \nu_y = \nu_{iy}, \quad \nu_z = \sqrt{\left(\frac{n_1}{\lambda}\right)^2 - \nu_x^2 - \nu_y^2} \quad (23-14)$$

with the index i denoting the incident field and m denoting the diffraction order. Only for the simplified special case of the incident frequency vector laying in the ν_x - ν_z plane do all vectors lay in the one plane (see figure 23-9b). For this case an equation for the x -component of the wave vectors follows from the grating diffraction equation above, for the transverse component with the angles measured with reference to the z axis (perpendicular to the grating):

$$\sin \alpha_m = \sin \alpha_0 + m \cdot \frac{\lambda}{p}. \quad (23-15)$$

With $\nu_{iy} \neq 0$, the diffracted beam no longer lies in the incident plane determined by the incident vector and the surface normal. This conical grating diffraction is a pure 3D effect and thus cannot easily be illustrated in a graphical way. In figure 23-9 the incident frequency vector has to be imagined as having a ν_y -component; then the center of the scattered Ewald sphere does not lie in the ν_x - ν_z plane in figure 21-9b but either behind or in front of it. The intercepts with the lines of the grating spectrum are still in the ν_x - ν_z plane, only it does not traverse the scattered Ewald sphere at the “equator” but instead at a “latitude”. From the section plane and the center of the sphere a truncated cone results. The spectral orders are limited to this cone, therefore it is known as conical diffraction.

The Born approximation is only valid in the approximation of single scattering, frequently also applied in the case of weak scattering, i.e., in inhomogeneous media. Multiple scattering can be considered by an expansion in the Born series using a scattered wave vector as the incident wave vector for the next scattering process [23-16]. The Laue equation, however, is valid in a more general way. It can be shown in a quite simple way that, apart from higher-degree effects, a diffracting object can always generate only the same directions for diffracted waves from a given incident wave even after multiple scattering. This means that the Laue equation correctly determines the directions of the diffracted waves, but not the amplitudes. So the limitations of the Born approximation refer to the scattering potential $f(\vec{\nu})$, and for this the scattering potential calculated by rigorous methods can be inserted instead of the Fourier transform of the object. The Born approximation according to eq. (23-11) offers only an approximation which, however, yields better results for phase objects than the Thin Element Approximation TEA with a complex transmission function [23-17].

23.3.2

Dändliker's Representation and the Shape of the Three-dimensional Transfer Function

All accessible information about an object is transferred to the diffracted wave according to the Laue equation. Thus the area in the frequency space can be determined where information about the object can be transmitted by propagating electromagnetic waves. This interpretation of the Laue equation is shown in figure 23-10 at an x - z cross-section in the frequency space.

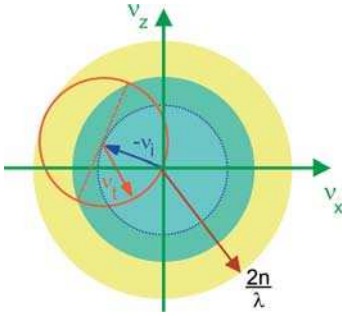


Figure 23-10: Illustration of the Laue equation for the construction of the 3D transfer function [23-2].

Each incident wave vector \vec{v}_i ends on the central Ewald sphere around the center of the coordinate system (blue circle). The scattered wave vectors for each incident wave vector \vec{v}_i lay again on an Ewald sphere whose center point is given by the respective incident wave vector \vec{v}_i (red circle). The scattered field parts are the contributions of the scattering potential $f(\nu)$ on the Ewald sphere. In the Born approximation the scattering potential is simply determined by the spatial frequency spectrum of the object. The radius of the Ewald sphere of the scattered field distribution is then equal to that of the incident field distribution if the incident and the scattered field are in a homogenous space with an identical refractive index.

According to Dändliker and Weiss [23-2] figure 23-10 allows an explanation of the limited resolution power of optical far-field methods. For this discussion it is completely insignificant which optical instrument is used. If one tries to determine the shape and structure of an object from the scattering potential, i.e., from each projection with a planar wave as the incident field distribution, one obtains information about the object along the scattered Ewald sphere. By changing the incident direction or by turning the object (as, e.g., used in computer tomography [23-1], [23-18], [23-19]) eventually the complete frequency spectrum of the object within the large circle (yellow) around the origin can be determined. The resolution limit or minimum object period d_{\min} follows from the maximum spatial frequency component of the object which can be transferred:

$$d_{\min} = \frac{1}{|\gamma_{\max}|} \geq \frac{\lambda}{2}. \quad (23-16)$$

Due to the technical layout, as for conventional optical imaging in translucent light, the reflected parts are mostly lost (except for con-focal microscopy). Considering only forward scattering it follows (see green circle in figure 23-10) that

$$d_{\min} \geq \frac{\lambda}{\sqrt{2}} \approx 0.7 \cdot \lambda. \quad (23-17)$$

Additional limitations of the resolution emerge from limited aperture angles or numerical apertures, respectively. Figure 23-11 shows the limitations of the diffracted wave vectors by the aperture angle of the pupil function of the imaging system.

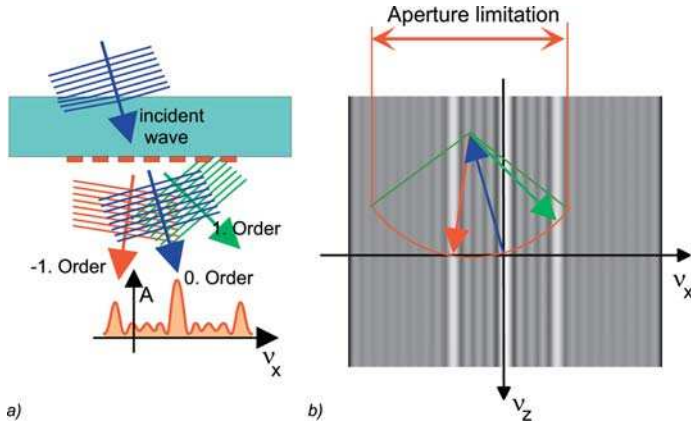


Figure 23-11: The transfer function of a perfectly imaging optical system is – besides a linear scaling corresponding to the change of the scale – given by a low-pass filtering.

For partially coherent optical imaging, the object is illuminated by a limited angular spectrum. Only the information for a finite angular range is gathered. So, besides the low-pass filtering by the imaging system the finite illumination aperture has also to be taken into account. Therefore it has to be assumed again that the images generated by single illumination source points are completely uncorrelated and thus have to be superposed incoherently in the image space. Only those object frequencies are transferred which result from the, now limited, illumination and object apertures (figure 23-12). The resulting volume in frequency space represents the *three-dimensional transfer function* of optical imaging. It describes the possible object frequencies contributing to the image. In figure 23-12 and the following, the shape

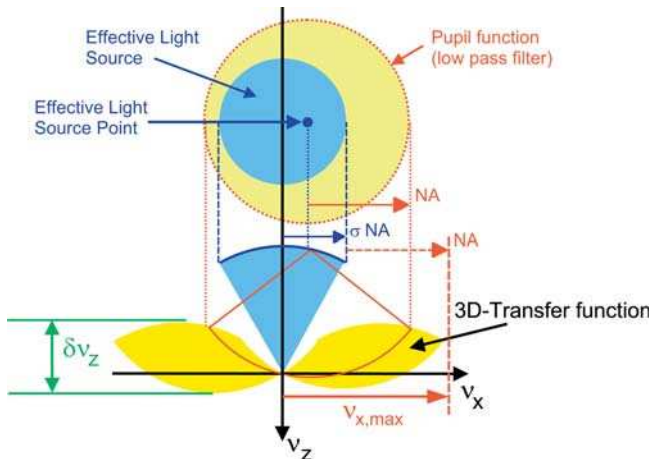


Figure 23-12: Resulting 3D transfer function at infinite illumination and imaging aperture (for telecentric imaging; simplified geometrical construction).

of the 3D-MTF is illustrated according to a simple geometrical construction, in which illumination directions with ν_y -components are neglected. Especially for structured illumination the intersections of the true 3D-MTF through the ν_x - ν_z plane might deviate significantly from this simplified construction, as shown below in the numerical examples of section 23.4.

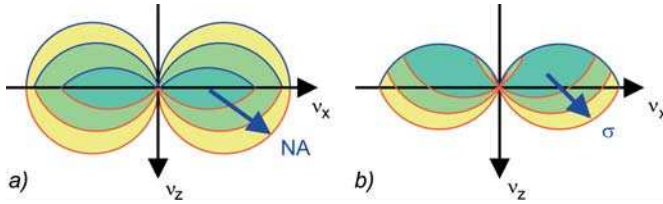


Figure 23-13: Cross-section through 3D transfer functions: a) for various apertures at $\sigma = 1$; and b) various σ for constant numerical aperture.

In figure 23-13 the intersections of the 3D transfer functions with the ν_x - ν_z plane are shown for the normalized aperture angle of the illumination σNA and the detection NA. Figure 23-13a compares the transfer functions for various numerical apertures with $\sigma = 1$, while figure 23-13b compares the transfer functions for a given numerical aperture at various σ -values. Taking the rotational symmetry around the ν_z axis into account, the 3D transfer functions are doughnut-shaped or toroid-like constructions.

For the three-dimensional case, the discussion of the transfer function becomes substantially more visual and concrete than in two dimensions. First scattering takes place at the object according to the Laue equation. Depending on the illumination direction, different object frequencies are transferred to the scattered waves. The scattered field distributions are then low-pass filtered by the numerical aperture. Corresponding interference patterns are formed in the image space to build the image intensity. It should be remembered that the image is generally not given by a simple Fourier transformation of the transfer function multiplied by the object spectrum. Rather the coherence properties of the light source, i.e., the illumination conditions, have to be considered. In Köhler illumination with an incoherent light source, first the intensity for an image amplitude calculated for a light source point must be determined, then all intensities of the partial images to all light source points have to be added.

For comparison, an example of the imaging of a grating beyond the coherent resolution limit is given in figure 23-14. For a numerical aperture of $NA = 0.8$, the coherent resolution limit is at periods $\delta d_{\text{coh}} = \lambda/NA = 1.25 \cdot \lambda$, with $\sigma = 0.7$ the resolution limit is approximately at period $\delta d \sim 0.735 \cdot \lambda$. The period of the image grating of $1.05 \cdot \lambda$ is beyond δd_{coh} , hence the central source point no longer contributes to the imaging and thus destroys the contrast by the false light share. In figure 23-14 the Ewald sphere around the central illumination direction – as well as the information about the extent of the object – intercepts only the zero order of the object frequency spectrum and thus delivers only a constant offset.

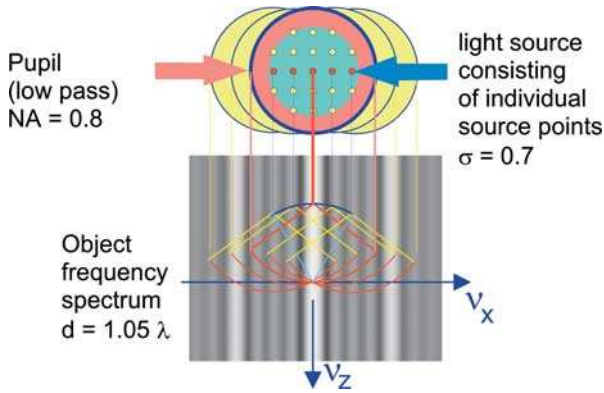


Figure 23-14: Different object frequencies are “read out” for different illumination directions.

Figure 23-15a shows the intensity at the center source point. Using light in the illumination direction mid left and right, respectively, the zero and the first order are intercepted “obliquely” by the Ewald sphere. The interference stripes are described by the difference vector of the two interfering waves, thus the orientation of the interference fringes in the assigned image are tilted to the optical axis (figure 23-15b). The Ewald spheres for the scattered field distributions to the selected peripheral points of the light source, however, intercept the zero and first diffraction order in the frequency spectrum. In this way the difference vector of the two interfering waves is perpendicular to the v_z -axis and so interference stripes parallel to the z -axis are formed (figure 23-15c). The image-forming interference patterns are now given by afocal two-beam interferences which, however, according to the interception points of the Ewald spheres with the

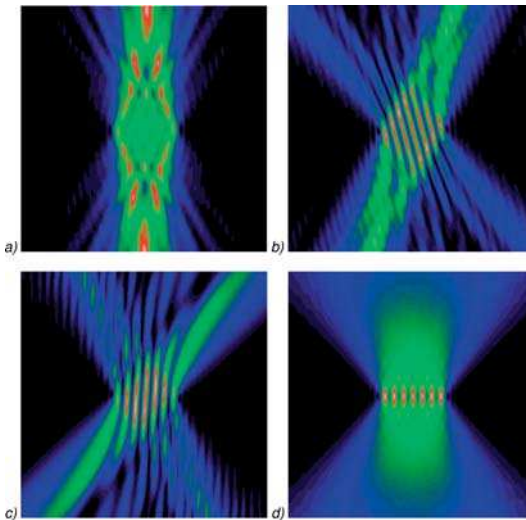


Figure 23-15: Selected partial images of the central (a); mid right (b); outer right light source point (c) and d) the incoherent interference of all partial images.

diffraction orders of the grating, are tilted to the optical axis. Although all two-beam interference patterns have a very large depth extent, the depth of focus decreases greatly, due to the incoherent superposition of the tilted interference patterns (figure 23-15d). Thus the example in figure 23-15 already contains indications for the selection of the illumination direction. So, according to the Siedentopf illumination principle, in order to increase the contrast it is best to leave out those illumination directions which do not contribute to the image formation. To increase the depth of focus even more, the afocal nature of the two-beam interference can be used, if there is only illumination in those directions whose interference patterns are aligned in parallel to the z-axis as shown in figure 23-15c. This example and more are discussed in detail below.

23.3.3

Resolution, Depth Resolution and Depth of Focus

From the 3D transfer function, the lateral resolution and the depth of focus can be determined using the inverse of the maximum spatial frequency of the transfer function. From the maximum transferred transverse frequency the lateral resolution at partially coherent illumination is found to be:

$$\delta d = \frac{1}{\nu_{x, \max}} = \frac{\lambda}{NA(1 + \sigma)} . \quad (23-18)$$

Equation (23-18) is identical to the fundamental equation for the optical resolution (chapter 20). Next the question of depth resolution or depth of focus will be considered. For conventional imaging like photography, usually a depth of focus is desired which is as high as possible, i.e., a depth resolution which is as low as possible, while for some modern microscopic methods the depth resolution is also to be as high as possible. It can be seen from figure 23-12 and 23-13 how the lateral resolution and the depth resolution depend on the illumination aperture, the numerical aperture and the object frequency. With increasing numerical aperture and with increasing incoherence, i.e., a larger illumination aperture σNA , the lateral resolution increases, but a higher depth resolution is mainly achieved by using a larger numerical aperture. In conventional imaging of ever tinier structure widths, larger and larger numerical apertures are used. That also decreases the image's depth of focus. For a large depth of focus it is necessary to obtain the longitudinal ν_z -expansion of the transfer function to be as small as possible, while the lateral resolution ought to stay as high as possible. On a first view both conditions are contradictory and cannot be fulfilled simultaneously. As will be described below, however, large depth resolution as well as large depth of focus and high resolution can be obtained for special objects under oblique illumination.

The McCutchen formula is applicable to coherent imaging or imaging with small sigma, when the 3D-transfer function is simply given by a segment of the Ewald sphere. For the imaging of gratings below the resolution limit, the numerical aperture of an optical system is not fully used and the depth resolution can be considerably lower than the minimum depth resolution according to eq. (23-9). For ampli-

tude gratings with dominating zero and first diffraction orders, the McCutchen formula can be generalized by replacing the numerical aperture by the effectively “used” aperture λ/d of the object with grating period d :

$$\delta z_T = \frac{1}{\Delta \nu_z} \approx \frac{2\lambda}{NA^2} = \frac{2d^2}{\lambda}. \quad (23-19)$$

Objects of discrete spectra are, however, periodical in the direction of the discrete frequencies. Equation (23-19) gives the longitudinal image frequency of the periodically repeated image. This effect is well known as the Talbot Effect [23-20], [23-21]. In figure 23-16 the three-dimensional aerial images of different gratings patterns are compared. The grating period is selected such that the third diffraction order passes the lens at the border of the pupil and contributes to the image. The apertures are $NA_1=0.25$, $NA_2=0.5$ and $NA_3=1$. The periods are thus $d_1=12\lambda$, $d_2=6\lambda$ and $d_3=3\lambda$. The pictures are scaled accordingly such that the lateral image at the best image plane is identical for all three images. However, the Talbot periods scale according eq. (23-19) from $\delta z_{T1}=288\lambda$, $\delta z_{T2}=72\lambda$, and $\delta z_{T3}=18\lambda$. The effective aperture used by the first diffraction order is three times less the imaging aperture, thus the depth of focus according to McCutchen deviates by a factor of 9.

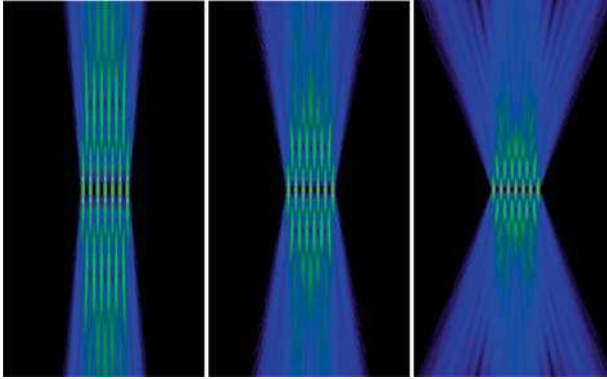


Figure 23-16: Three aerial images at object periods of $3\lambda/NA$ for different NA ; $NA_1=0.25$, $NA_2=0.5$ and $NA_3=1.0$ (from left to right).

Figure 23-17 shows the corresponding axial intensity scans $I(x=0, z)$ and $I(x=d/2, z)$ for the three imaging conditions. The images are symmetrical with respect to the x -axis, thus only positive z -coordinates are shown. It is to be remarked that the McCutchen formula according to eq. (23-19) gives the longitudinal periodicity of the three-dimensional aerial image as the distance between the two maxima of the central intensity peak. Taking a certain image contrast requirement as the limit, the depth of focus is significantly lower. The contrast $V(z)$ can be approximated from both intensity scans through the maximum and minimum of the central object period of figure 23-17. As illustrated in figure 23-18, at the z -positions, according to eq. 23-19, the image contrast is reversed.

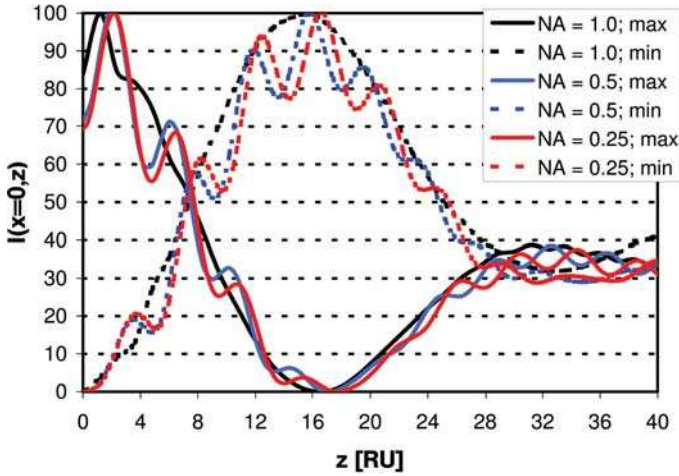


Figure 23-17: Axial intensity scan $I(x=0,z)$ of the three images (max) and through the first minimum of grating image at $z=0$ (min); z -scale in Rayleigh units according eq. (23-10).

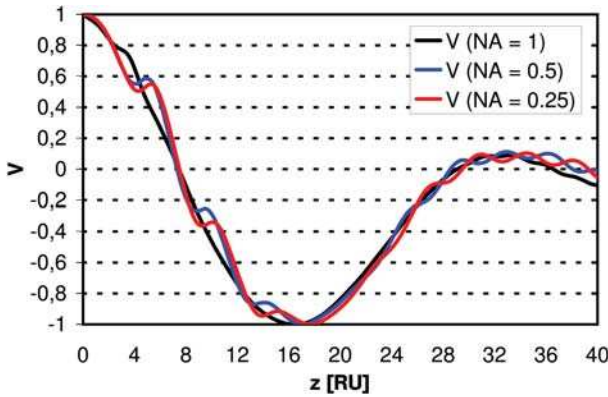


Figure 23-18: Contrast V through defocus z .

From figure 23-13a and b it can be seen that the depth resolution depends on the illumination conditions. Since the McCutchen formula eq. (23-19) is considering the numerical aperture only it may not exactly describe the depth of focus. In addition to the system parameters like the object and illumination aperture, the depth resolution also depends on the lateral object frequency. Figure 23-19 illustrates the maximum frequency $\Delta\nu_z$ as a function of a grating object period d . From the calculation of the maximum frequency $\Delta\nu_z$ by means of the Laue equation (eq. (23-13)) for the image of a grating, the depth resolution can be determined. For large sigma ($\sigma > 1 - \lambda/(NA \cdot d)$) and under a small-angle approximation comparable to the McCutchen formula, a generalized depth resolution formula can easily be derived:

$$\delta z = \frac{d^2}{NA \cdot d - \lambda/2} \quad (23-20)$$

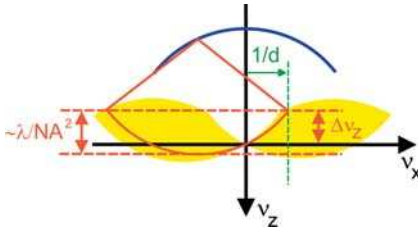


Figure 23-19: Construction of a generalized formula for depth resolution.

Figure 23-20 shows the numerically calculated depth resolution δz for different apertures at $\sigma=1$ and $\lambda=1$ (solid lines) and the results according to equation (23-19) (dashed lines). For incoherent imaging with high aperture angles the approximation $1-\cos\alpha \sim 0.5\sin^2\alpha$ collapses and the approximation in eq. (23-19) predicts a significantly larger depth of focus. The minimum depth resolution is reached for object periods at the coherent resolution limit $d = \lambda/NA$ for which eq. (23-20) becomes the McCutchen-formula eq. (23-19).

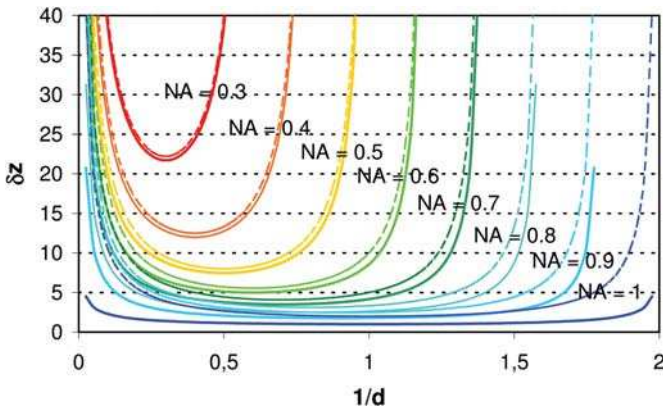


Figure 23-20: The depth resolution δz is a function of the object frequency $1/d$ and the numerical aperture ($\sigma=1, \lambda=1$); solid lines: exact calculation; dashed lines: small angle approximation according to eq. (23-19).

23.3.4

3D-Transfer Functions in Microscopy

The 3D-transfer function is frequently applied to the different imaging conditions of microscopy [23-22], [23-23]. As examples, conventional imaging in reflection and transmission and also confocal imaging, are discussed. In figure 23-21 sections through the v_x-v_z plane of the most relevant examples of 3D-transfer functions for microscopy are compared, in which the illumination directions from v_y are neglected.

For microscopy in transmission or optical lithography, but also, for example, in a slide projector, the optical system collects the radiation which is transmitted through the object. In this case one obtains a transfer function similar to that of figure 23-12.

The dark-field 3D-MTFs do not intersect the zero order of the object spectrum at $\nu_x = \nu_z = 0$. In the case of weak scattering, i.e., dominating zero order, the frequency read-out for the illumination directions contains only the contributions from the scatterer and the image therefore has a high contrast. With an illumination setting $\sigma_{\text{out}} > 1$ for dark-field illumination, the maximum transverse frequency ν_x is higher and therefore the transverse frequency bandwidth of the 3D-MTF increases at the expense of the low frequencies.

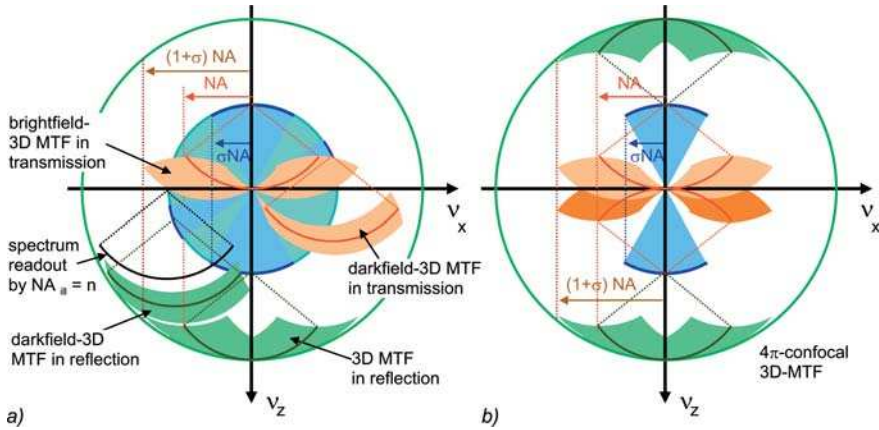


Figure 23-21: a) Transfer functions in reflection, transmission for bright and dark-field illumination; and b) transfer function for 4π -confocal imaging.

For imaging with reflected light, an optical system (photo objective, microscope, eye) receives the radiation reflected by the object. With this imaging an area of the object spectrum shifted in the ν_z -direction is transferred (figure 23-21a). The depth resolution, however, does not increase in this transformation since the frequency bandwidth $\Delta\nu_z$ does not increase. The amplitudes are just phase-modulated in the z -direction according to a carrier frequency which vanishes after forming the intensity.

A significant increase in depth resolution is obtained by scanning confocal microscopy, where the illumination and the detection contribute to the imaging process in an identical manner [23-24], [23-25]. However, this type of imaging requires a scanning imaging operation, i.e., every image point is detected separately and the image is numerically computed. So far, the highest depth resolution is achieved by collecting reflected as well as transmitted radiation and causing them to interfere coherently for image formation (figure 23-21b and 23-22). Due to the collection, in principle, of the full Ewald sphere of scattered light, this is called 4π -confocal microscopy according to Stelzer and Hell [23-26], [23-8].

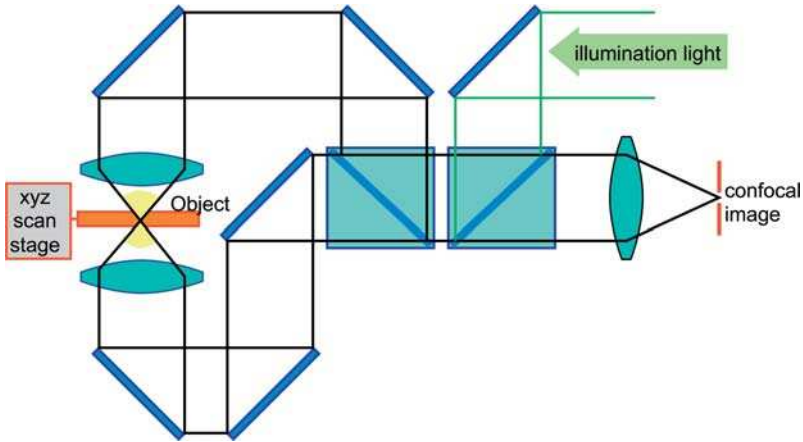


Figure 23-22: Confocal 4π -microscope according to Stelzer and Hell.

23.3.5

Magnification and a Comment on Absolute Instruments

An image similar to the object is only achievable by optical systems obeying the sine condition, i.e., the lateral frequency spectrum of the image is scaled with the inverse of the magnification β of the imaging. In general, from the necessary transverse linear scaling of the frequency spectrum, the sine condition, follows a non-linear scaling of the longitudinal spatial frequencies ν_z – with the one exception of $\beta=1$. It is therefore in general impossible to achieve a three-dimensional image similar to the three-dimensional object. This obvious fact was already derived by Maxwell and Carathéodory. For media of constant indices of refraction the plane mirror is the only perfect or absolute optical instrument [23-27], [23-28].

The impact on image formation of imaging with magnification $\neq 1$ is illustrated in figure 23-23. For a transverse similar image, the transverse spatial frequencies of the image are scaled according to

$$\nu'_x = \frac{1}{\beta} \nu_x \quad \text{or} \quad \beta = \frac{\nu_x}{\nu'_x} = \frac{n \sin \varphi}{n' \sin \varphi'} \tag{23-21}$$

whereby the longitudinal spatial frequencies in the object and image space are given according to the Ewald equation by:

$$\nu_z = \sqrt{\left(\frac{n}{\lambda}\right)^2 - \nu_x^2}, \tag{23-22}$$

$$\nu'_z = \sqrt{\left(\frac{n'}{\lambda}\right)^2 - \nu_x'^2} = \sqrt{\left(\frac{n'}{\lambda}\right)^2 - \frac{\nu_x^2}{\beta^2}}. \tag{23-23}$$

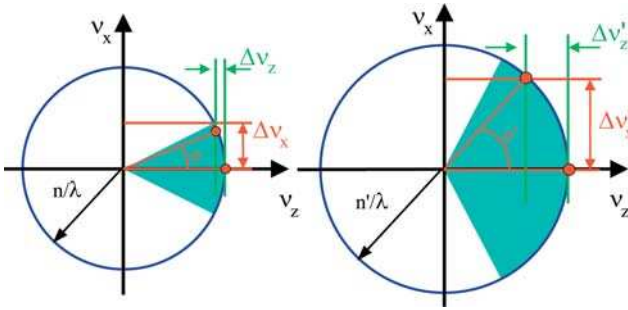


Figure 23-23: Impact of imaging, according to the sine condition, on the longitudinal frequency scaling; the frequency spectrum in the object space with refractive index n (left side) and in the image space with refractive index n' (right side).

The longitudinal magnification α is thus given by

$$\alpha = \frac{\Delta v_z}{\Delta v'_z} = \frac{\frac{n}{\lambda} - v_z}{\frac{n'}{\lambda} - v'_z} = \frac{\frac{n}{\lambda} - \sqrt{\left(\frac{n}{\lambda}\right)^2 - v_x^2}}{\frac{n'}{\lambda} - \sqrt{\left(\frac{n'}{\lambda}\right)^2 - v_x'^2}}. \quad (23-24)$$

The longitudinal magnification α is in general a function of the transverse frequency v_x . After insertion of the propagation angles φ with respect to the optical axis, the geometrical-optical formulation of eq. (23-24) is derived:

$$\alpha = \frac{n(1 - \cos \varphi)}{n'(1 - \cos \varphi')} = \frac{n \sin^2 \frac{\varphi}{2}}{n' \sin^2 \frac{\varphi'}{2}}. \quad (23-25)$$

A longitudinally similar image is obtained only if the longitudinal magnification α is constant for all angles φ . Only in the paraxial approximation of the square-roots in eq. (23-24), is the longitudinal magnification given by a constant – the square of the lateral magnification:

$$\alpha \approx \frac{n'}{n} \beta^2. \quad (23-26)$$

With the paraxial approximation, after insertion in eq. (23-25) the Herschel condition [23-29] follows

$$\beta \stackrel{!}{=} \frac{n \sin \frac{\varphi}{2}}{n' \sin \frac{\varphi'}{2}}. \quad (23-27)$$

For physical reasons, i.e., the wave-nature of light, it is impossible to satisfy the Herschel condition eq. (23-27) strictly. Furthermore, the Herschel condition and the sine condition are incompatible and cannot both be exactly satisfied simultaneously – again with the one exception of $\beta=1$ and $n=n'$.

In an optical imaging system therefore, the lateral frequency is scaled linearly, while the longitudinal spatial frequency is scaled according to eq. (23-24). The consequences of imaging with $|\beta| \neq 1$ are illustrated in figure 23-24. The illumination directions are indicated by dark blue circle segments in the negative ν_z direction. Illumination during transmission with $\sigma = 1$ is assumed. The blue transfer functions indicate the frequency spectrum gathered from the object in object space, with individual apertures in object space (blue circle segments in the positive ν_z -direction). After imaging with $|\beta| = 2$ in figure 23-24a and $|\beta| = 0.5$ in figure 23-24b, the transmitted spectrum illustrated in red is laterally scaled, respectively. The red circle segments indicate the generalized apertures in the image space with the ν_z -components scaled according to eq. (23-24). The transfer function in image space can be re-scaled in object coordinates. The scaling is illustrated in figure 23-25 at the example of $\beta = 0.5$, whereby the wavelength is scaled with β^2 to maintain equation (23-24).

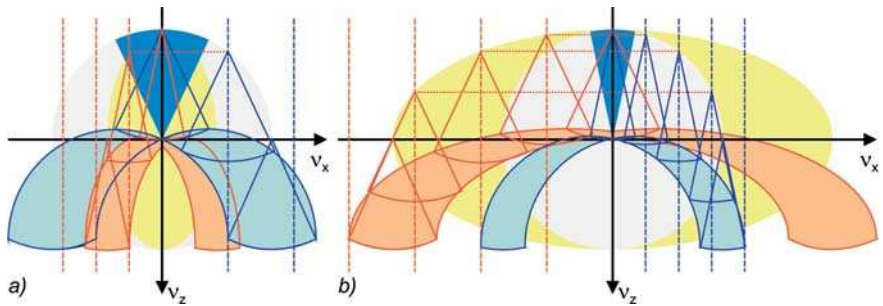


Figure 23-24: Scaling of 3D-transferred spatial frequency spectrum; a) with $|\beta| = 2$; and b) with $|\beta| = 0.5$.

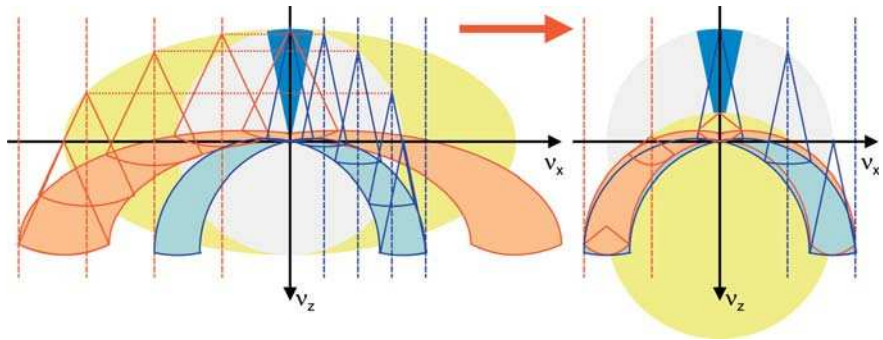


Figure 23-25: Scaling of the 3D-transferred spatial function according to imaging in frequency coordinates of the object space with $|\beta| = 0.5$.

With thin gratings, considered as objects in the remainder of this chapter, the spectrum is invariant in the ν_z direction, therefore the effect of the scaling in ν_z is negligible. Similar to the transverse treatment of Fourier imaging, the transverse scaling with $1/\beta$ can be applied directly to the object spectrum and the image formation can be simulated in image space – with the one exception of dark-field illumina-

tion in illumination directions $\sigma > NA_{\text{image}} \cdot \beta$. Simulation of reduction imaging in the image space allows only for illumination directions up to NA_{image} , which corresponds in the object space to a maximum illumination direction $NA_{\text{image}} \cdot \beta$.

23.4

Selected Examples of the Three-Dimensional Transfer Function

The degree of coherence is adjusted by the illumination. The smaller the effective light source image in the pupil becomes, the more coherent the image becomes. In figure 23-14 an example with $\sigma = 0.7$ has already been shown. In the following, different image settings are discussed. As a structure a grating consisting of seven periods is chosen. Thus the grating spectrum is superimposed by a sinc-function which is generated by the envelope of the grating with seven periods. The discussion is again limited to ideal imaging in telecentric systems, although an extension to non-telecentric imaging can be carried out quite easily.

23.4.1

Transfer Function for Incoherent Imaging with $\sigma = 1$

As an example, imaging for the period of 1.4λ at the resolution limit with an $NA = 0.4$ and $NA = 0.7$ are compared. For incoherent imaging with $\sigma = 1$ the 3D-transfer function becomes symmetrical to the $\nu_x - \nu_y$ -plane. Figure 23-26 shows cross-sections of the 3D-transfer function through the $\nu_x - \nu_z$ plane, figure 23-27 shows the frequency spectrum of the grating transferred by this. In figure 23-28, the resulting aerial images are illustrated. The example shows clearly that the depth of focus can be increased in the optical system by stopping down until, theoretically, an infinite

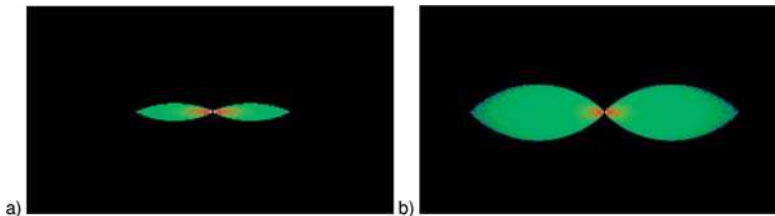


Figure 23-26: 3D-Transfer function with $NA = 0.4$ (a) and $NA = 0.7$ (b).

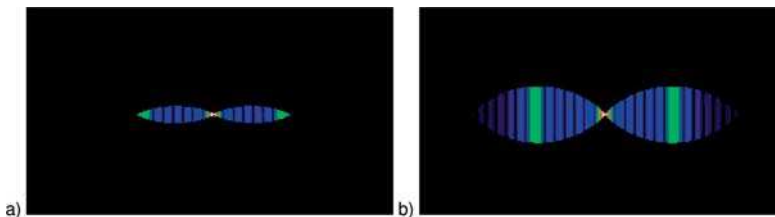


Figure 23-27: Transferred frequencies with $NA = 0.4$ (a) and $NA = 0.7$ (b).

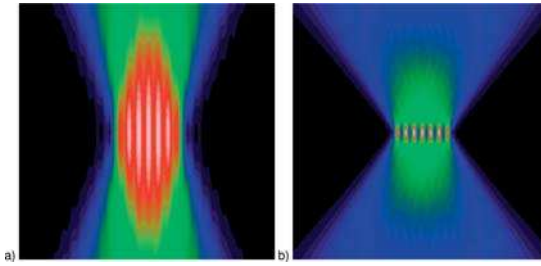


Figure 23-28: x - z cross-sections through the aerial image with $NA=0.4$ (a) and $NA=0.7$ (b).

depth of focus can be reached at the resolution limit. However, the contrast for the imaging with conventional illumination is significantly lower with approximately $V \sim 7\%$ at $NA = 0.4$ than at $NA = 0.7$ with $V \sim 70\%$ and vanishes at the resolution limit.

23.4.2

Partial Coherent Image Examples

Figure 23-29 shows the Laue construction for partial coherent imaging with $NA = 0.8$ and $\sigma = 0.7$. The resolution limit is approximately $\delta d_\sigma \sim 0.735 \cdot \lambda$. Since the selected period of $d = 1.3 \cdot \lambda$ is still above the coherent resolution limit $\delta d_{coh} \sim 1.25 \cdot \lambda$, the central source point contributes to the image (figure 23-30a) and limits the depth of focus by the three-beam interference to approximately $\delta z \sim 3 \cdot \lambda$ (figure 23-30c) according to McCutchen's formula.

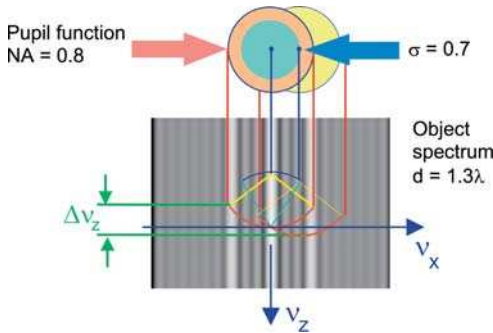


Figure 23-29: Laue construction of 3D-diffracted spectra.

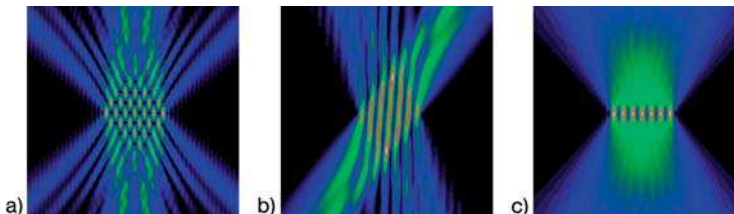


Figure 23-30: Interference patterns for individual source points: a) central source point; b) right source point; c) incoherent superposition of interference patterns.

To obtain large lateral resolution, the maximum frequency is not to be curtailed although it seems to be quite obvious to cut down the lower frequencies, i.e., to curtail the pupil in the center. Central obscuration is unavoidable for some optical systems, such as reflection telescopes. The construction of the transfer function is illustrated in figure 23-31 for light source points with $\nu_y = 0$. If the light source points with $\nu_y \neq 0$ are considered the transfer function and the transferred object spectrum of the grating are changed only insignificantly for a central obscuration of $NA = 0.2$ (figure 23-32), though in the example the contrast decreases from $V \approx 86\%$ to $V \approx 80\%$ without increasing the depth of focus significantly (figure 23-33).

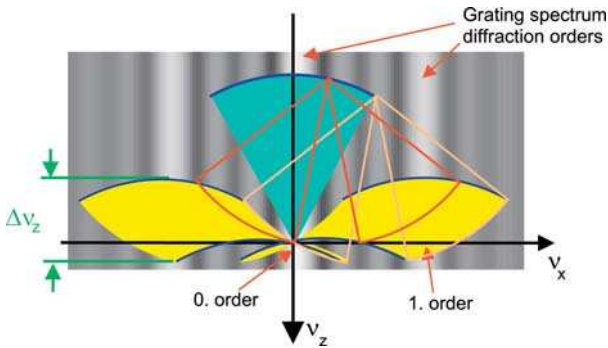


Figure 23-31: 3D-Transfer function under central obscuration.

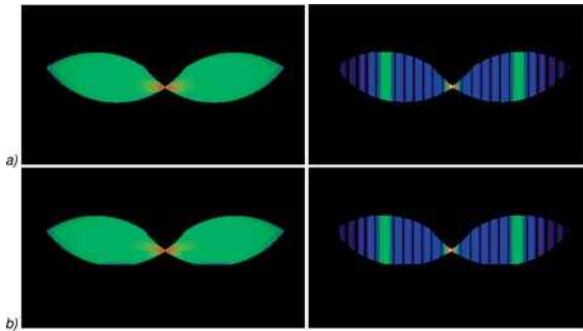


Figure 23-32: 3D-transfer function and transferred object spectrum: a) with conventional system; b) with central obscuration of $NA_{in} = 0.2$ ($NA = 0.8$, $\sigma = 0.7$).

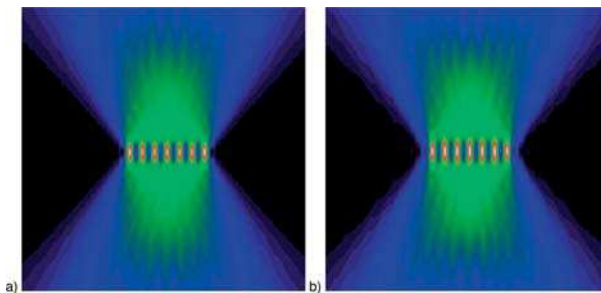


Figure 23-33: Aerial image without (a) and with (b) central obscuration.

23.4.3

'Tayloring' of the 3D-Transfer Function

For the investigation of some objects, such as biological objects, or for determining the ideal image plane, a large depth of resolution is desirable. On the contrary, for the microscopical investigation of planar objects or lithographic exposure of ideal gratings, a very low depth of resolution is advantageous. In the case of a considerable field curvature of the imaging lens, a thin planar object like a transparency may not be focused for all positions in the object plane simultaneously and therefore the image suffers from blurring, due to defocus aberration. Therefore, it is desirable that the image should not change over a focus range which is as large as possible. As will be shown in this chapter, with the 3D-transfer function the proper illumination and imaging conditions for either large or low depth resolution can be obtained, but the object spectrum has to be taken into account.

According to the Siedentopf principle, a higher contrast can be achieved by, e.g., annular illumination. A cross-section through the 3D-transfer function under annular illumination is shown in figure 23-34. The lateral resolution does not change according to the maximum illumination angle $\sin\alpha_{max} \sim \sigma \cdot NA$. The longitudinal extent of the transfer function, though, is constricted. Correspondingly the depth of focus is increased (for a two-dimensional treatment see e.g. [23-30]). In figure 23-34 the 3D-transfer function under annular illumination as a typical example is shown for illumination directions with $\nu_y = 0$. In the example the depth of focus can in principle become infinite. For this the illumination angles have to be selected such that the Ewald spheres of the diffracted field distributions intercept the diffraction orders of the object near the $\nu_x - \nu_y$ plane at $\nu_z = 0$. Then the extent of the transferred system in the ν_z -direction $\Delta\nu_z$ becomes a minimum, making the depth of focus $DOF = \delta z = 1/\Delta\nu_z$ a maximum. This of course follows from all two-beam interferences running almost parallel to the z -axis. Although the depth of focus is considerably higher, as for conventional illumination, due to the contribution to the transfer function for annular illumination directions with $\nu_y \neq 0$, the contrast and depth of focus is diminished. Figure 23-35 shows the 3D-transfer function and the transferred object spectrum for a grating period of $1.05 \cdot \lambda$ ($NA = 0.8$, $\sigma = 0.55 - 0.65$), the corresponding image is compared with conventional illumination in figure 23-36.

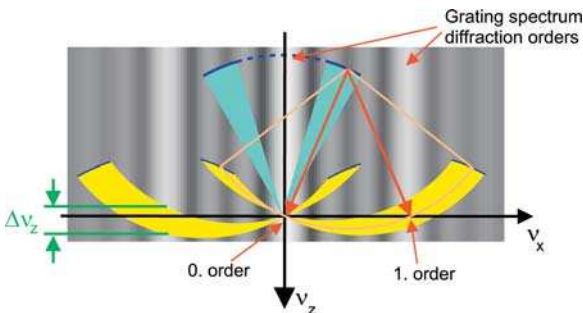


Figure 23-34: 3D-Transfer function under annular illumination.

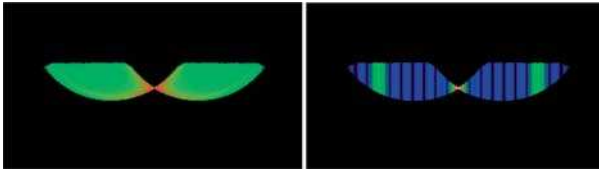


Figure 23-35: 3D-Transfer function under annular illumination.

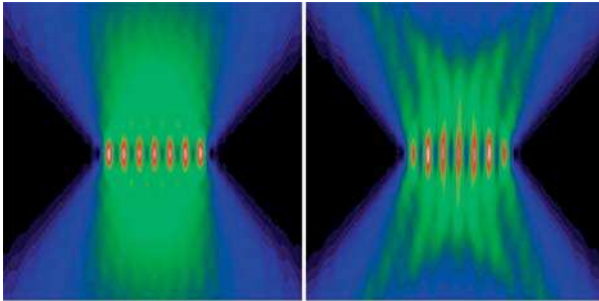


Figure 23-36: Aerial image with conventional setting ($\sigma=0.8$; left) and annular illumination setting ($\sigma=0.55-0.65$; right).

As Hopkins discovered, the annular illumination, however, still contains illumination directions which do not lead to an image contrast [23-31]. In order to fulfill the condition for maximum depth of focus perfectly, i.e., to ‘read out’ the object spectrum only at coordinates $v_z=0$, the v_x -component of the illumination has to be equal to half the inverse grating period:

$$d \cdot \lambda \cdot v_{ix} = d \cdot \sin \alpha_i = \frac{\lambda}{2}. \quad (23-28)$$

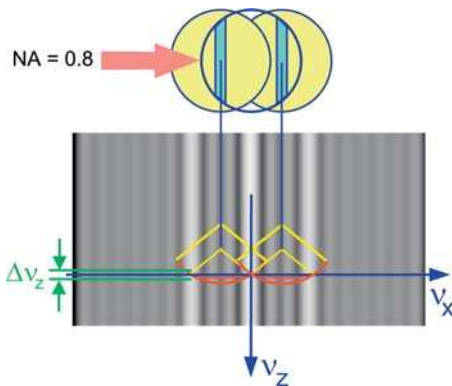


Figure 23-37: 3D-Transfer function of the dipole illumination.

The illumination condition eq. (23-28) for large depth of focus is equivalent to the illumination condition for imaging at maximum contrast, i.e., when neighboring

openings of the idealized grating are illuminated in phase opposition (see chapter 24) [23-32], [23-33]. So the ideal illumination for the imaging of an idealized grating with a large depth of focus consists of two line-shaped light sources separated by the grating frequency. Figure 23-37 shows the construction, figure 23-38 the transfer function and the transferred grating spectrum with a grating of 1.05λ at an aperture of $NA = 0.8$. The v_x -components of the light source are in the v_x -direction between $\sigma = 0.55$ and $\sigma = 0.65$. Image simulation results are given in figure 23-39 for 7 and 15 grating periods.

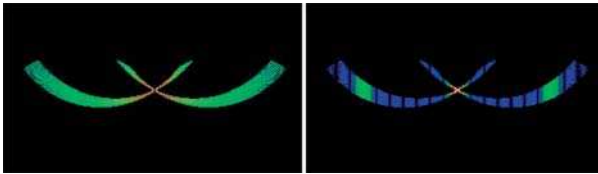


Figure 23-38: Transfer function and transferred spectrum with a line-shaped light source with v_x between $\sigma = 0.55 NA$ and $\sigma = 0.65 NA$.

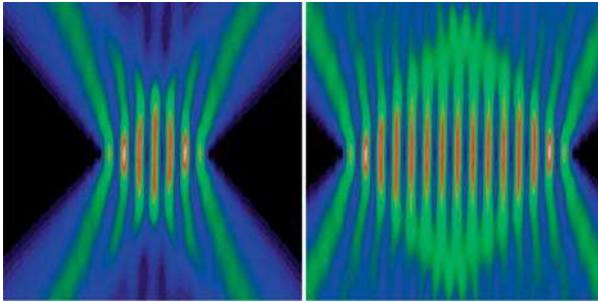


Figure 23-39: Aerial image with line-shaped light source, left with 7 grating periods, right with 15 grating periods; the depth of focus depends on the extent of the grating.

When, on the other hand, large depth resolution is required, i.e., to find precisely the ideal image plane, the illumination directions in accordance with the object period can be selected with the help of figure 23-40.

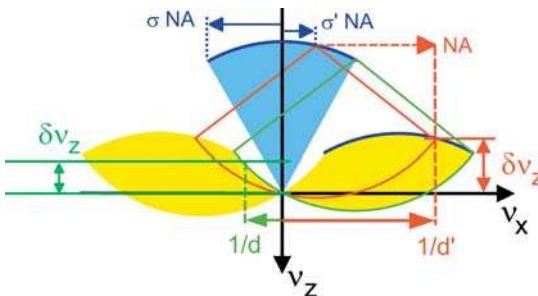


Figure 23-40: Two illumination directions σ' and σ for different object period d and d' for which the depth resolution is a maximum ($\lambda = 1$).

For spatial frequencies ν_x of the object below $NA \cdot (1-\sigma)/\lambda$ the maximum depth resolution is obtained with maximum obliqueness of the illumination direction, shown in the example of figure 23-40 on the left-hand side (green curves). From the maximum frequency $\delta\nu_z$ the depth resolution is derived as

$$\delta z \approx \frac{\lambda}{\lambda - 2\sigma \cdot NA \cdot d} \cdot \frac{2d^2}{\lambda} . \quad (23-29)$$

For spatial frequencies ν_x of the object above $NA \cdot (1-\sigma)/\lambda$ (see figure 23-40, right side) the illumination direction for maximum depth resolution is given by

$$n \sin \alpha_i = \sigma' \cdot NA = \frac{\lambda}{d} - NA . \quad (23-30)$$

The depth resolution δz for the object frequency $\nu_x = 1/d > NA \cdot (1-\sigma)/\lambda$ is thus

$$\delta z \approx \frac{\lambda}{2NA \cdot d - \lambda} \cdot \frac{2d^2}{\lambda} . \quad (23-31)$$

As an alternative to selecting correct illumination directions, the 3D-spectrum can be limited by using an appropriate pupil filter. In the example of central obscuration and coherent illumination, the zero diffraction order is filtered and only the higher diffraction orders may pass the pupil. For object periods beyond the coherent resolution limit, only ± 1 diffraction orders pass the pupil (figure 23-41). The corresponding transfer functions are shown in figure 23-42. Besides the frequency doubling of the interference pattern almost every other piece of information about the object is lost at too strong a central obscuration (figure 23-43d). If the central obscuration is reduced to, e.g., $NA_{in}=0.2$ at $NA_{out}=0.8$, an interference pattern with double frequency limited to the object area is obtained (figure 23-43b). The suppression of the zero diffraction order was proposed by the application of a pupil filter for spatial frequency doubling lithography [23-34].

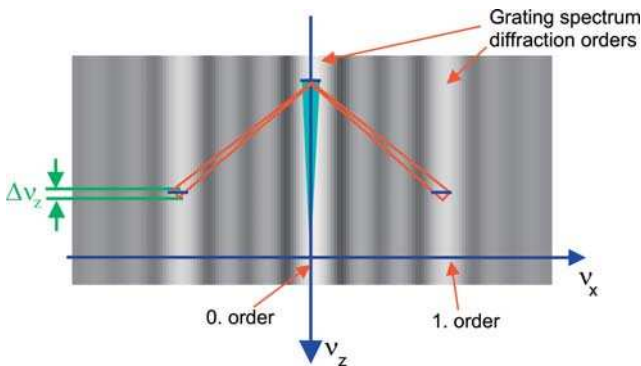


Figure 23-41: Transfer function for an annular pupil and coherent illumination.

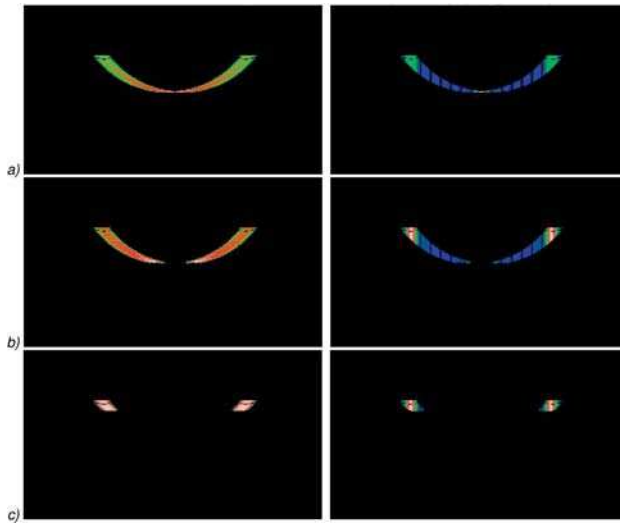


Figure 23-42: Transfer functions and transferred spectra: a) without central obscuration; b) with $NA_{in} = 0.2$; and c) with $NA_{in} = 0.7$; $NA = 0.8$, $d = 1.3 \cdot \lambda$; $\sigma = 0.1$.

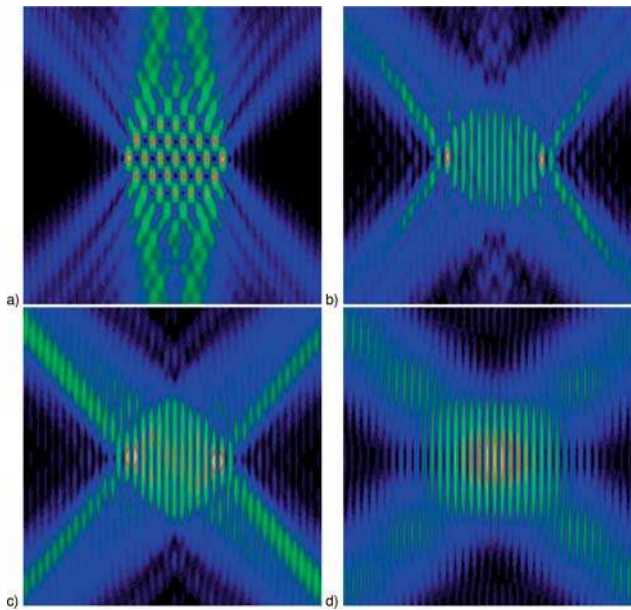


Figure 23-43: a) Aerial images without central obscuration; b) Central obscuration of zero order with small central obscuration ($NA_{in} = 0.2$); c) with $NA_{in} = 0.4$; and d) large central obscuration $NA_{out} = 0.8$. $NA = 0.8$, $d = 1.3 \cdot \lambda$, $\sigma = 0.1$.

Since the point-spread function of an optical system with an annular pupil is, however, disadvantageous for arbitrary objects of different shape and size, the zero order is frequently suppressed by the appropriate design of the object with phase shifting structures (e.g., alternating phase-shift masks, [23-35], [23-36]). With an alternating phase shift of π , neighboring openings of the grating are in phase opposition and an analogous effect is obtained as with oblique illumination, according to eq. (23-28). Since there are higher diffraction orders, and the phase-shifting structures have the desired effect of $\Delta\varphi \sim \lambda/2$ only in a small angle range, phase-shifting masks are mainly used under coherent illumination [23-36].

23.4.5

Influence of Aberrations

Three diffraction orders with their phase position and direction define a wavefront by whose curvature radius and angle of tilt, the position of the image pattern can be determined (figure 23-44a). On the other hand two diffraction orders determine only the orientation and the phase position of the interference fringes (figure 23-44b). Wavefront errors lead to a different interference pattern with different distortion and defocus for each illumination direction. The superimposed intensity patterns with non-fitting phase position thus form an image which is additionally faded by aberrations.

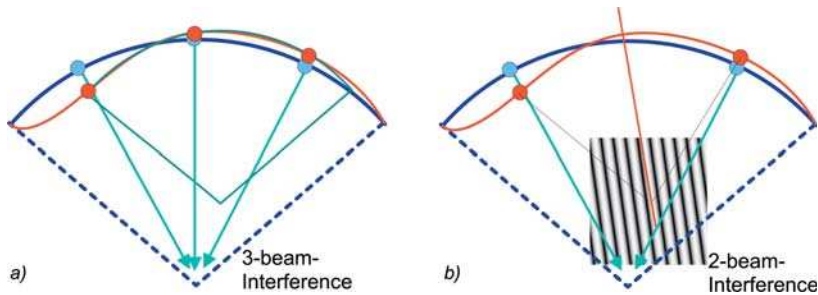


Figure 23-44: a) Asphere is determined by three diffraction orders; b) two diffraction orders only determine the orientation and the phase position of the interference pattern.

As an example, figure 23-45 shows the effect of the symmetrical wavefront error, like spherical aberration $Z_9 = \lambda/2$. Figure 23-46 shows the anti-symmetrical wavefront error coma $Z_7 = \lambda/2$ (at $d = 1.3 \cdot \lambda$, $NA = 0.8$, $\sigma = 0.7$ for both examples). In contrast to the three-beam interference produced by the illumination direction of the central light source point, the impact of the strong aberration on each two-beam interference pattern can hardly be seen. Superposition of the intensities from different light source points leads – even without the three-beam interferences – to a faded image.

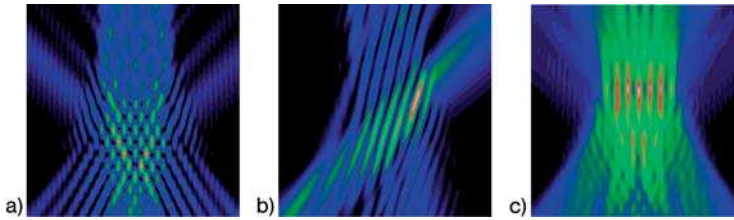


Figure 23-45: Interference images with spherical aberration of $\lambda/2$ to: a) the central light source point; b) the peripheral point of the effective source; c) the superposition of all interference patterns delivers the faded image.

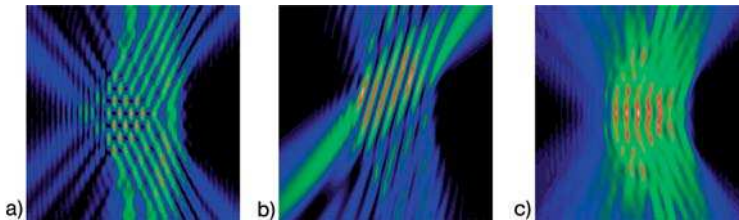


Figure 23-46: Interference image with coma of $\lambda/2$ to: a) the central light source point; b) the peripheral point of the effective light source; c) the superposition of all interference patterns delivers the faded image.

23.5

Literature

- 23-1 E. Wolf, Three-dimensional structure determination of semi-transparent objects from holographic data, *Opt. Commun.* **1**, 153–156 (1969).
- 23-2 R. Dändliker und R. Weiss, Reconstruction of the three-dimensional refractive index of scattered waves, *Opt. Commun.* **1**, 323–328 (1970).
- 23-3 A. W. Lohmann, Three-dimensional properties of wave-fields, *Optik* **51**, 105–117 (1978).
- 23-4 N. Streibl, Fundamental restrictions for 3D light distributions, *Optik* **66**, 341–354 (1984).
- 23-5 N. Streibl, Depth transfer by an imaging system, *Opt. Acta* **31**, 1233–1241 (1984).
- 23-6 N. Streibl, Three-dimensional imaging by a microscope, *J. Opt. Soc. Am. A* **2**, 121–127 (1985).
- 23-7 C.J.R. Sheppard, The spatial frequency cutoff in three-dimensional imaging, *Optik* **72**, 131–133 (1986).
- 23-8 M. Gu and C.J.R. Sheppard, Three-dimensional transfer functions in 4Pi confocal microscopes, *J. Opt. Soc. Am. A* **11**, 1619–1627 (1994).
- 23-9 Y. Kawata, R. Juskaitis, T. Tanaka, T. Wilson, S. Kawata, Differential phase contrast microscope with a split detector for readout system of a multilayered optical memory, *Appl. Opt.* **35**, 2466–2470 (1996).
- 23-10 M. Berek Über Kohärenz und Konsonanz des Lichtes, IV. Die optische Abbildung nichtselbstleuchtender Objekte, *Z. f. Physik* **37**, 420–449 (1926).
- 23-11 B. Roy Frieden, Longitudinal image formation, *J. Opt. Soc. Am.* **56**, 1495–1501 (1966).
- 23-12 B. Roy Frieden, Optical transfer of the three-dimensional object, *J. Opt. Soc. Am.* **57**, 56–66 (1967).
- 23-13 C.W. McCutchen, Generalized Aperture and Three-Dimensional Diffraction Image, *J. Opt. Soc. Am.* **54**, 240–244 (1964).
- 23-14 J. Bille, B. Schmitt, E. Beck, Reconstruction of three-dimensional light microscopic im-

- ages comparing frequency and spatial domain methods, *SPIE* **697**, 349–356 (1986).
- 23-15** Jakob J. Stamnes, *Waves in focal regions* (Adam Hilger, Bristol and Boston, 1986).
- 23-16** A. Rohrbach, W. Singer, Scattering of the scalar field at a dielectric surface by Born series expansion, *J. Opt. Soc. Am. A* **15**, 2651–2659 (1998).
- 23-17** M. Testorf, On the zero-thickness model of diffractive optical elements, *J. Opt. Soc. Am. A* **17**, 1132–1133 (2000).
- 23-18** A.J. Devaney, A filtered backpropagation algorithm for diffraction tomography, *Ultrasonic Imaging* **4**, 336–350 (1982).
- M.H. Maleki, A.J. Devaney, A. Schatzberg, Tomographic reconstruction from optical scattered intensities, *J. Opt. Soc. Am. A* **9**, 1356–1363 (1992).
- 23-19** L.-J. Gelius, J.J. Stamnes, Diffraction tomography: potentials and problems, in *Scattering in Volumes and Surfaces*, M. Nieto-Vesperinas and J. C. Dainty, eds. (Elsevier, Amsterdam, 1990), pp. 91–109.
- 23-20** H.F. Talbot, *Philos. Mag.* **9**(56), 401–407 (1836).
- 23-21** W. D. Montgomery, Self-imaging Objects of Infinite Aperture, *J. Opt. Soc. Am.* **57**, 772–778 (1967).
- 23-22** C.J.R. Sheppard and X.Q. Mao, Three-dimensional imaging in a microscope, *J. Opt. Soc. Am. A* **6**, 1260–1269 (1989).
- 23-23** T. Noda, S. Kawata and S. Minami, Three-dimensional phase contrast imaging by an annular illumination microscope, *Appl. Opt.* **29**, 3810–3815 (1990).
- 23-24** C.J.R. Sheppard and A. Choudhury, Image formation in the scanning microscope, *Optica Acta* **24**, 1051–1073 (1977).
- 23-25** G.J. Brakenhoff, P. Blom, P. Barends, Confocal scanning light microscopy with high numerical immersion lens, *J. Microsc.* **117**, 219–232 (1979).
- 23-26** S. Hell, E. H.K. Stelzer, Properties of a 4Pi confocal fluorescence Microscope, *J. Opt. Soc. Am. A* **9**, 2159–2166 (1992).
- 23-27** M. Born, *Optik*, Reprint of the Original from 1932 (Springer, Berlin/Heidelberg/New York, 1985).
- 23-28** M. Born, E. Wolf, *Principles of Optics*, 6th ed. (Pergamon Press, Oxford/New York, 1980).
- 23-29** Ch. Hofmann, *Die optische Abbildung* (Akademische Verlagsgesellschaft, Leipzig, 1980).
- 23-30** S. Kessler, C. Nitzsche and J. Hebenstreit, Defocused imaging of an amplitude sine wave object under partially coherent illumination with an annular source, *J. Mod. Optics* **35**, 177–186 (1988).
- 23-31** H.H. Hopkins, On the diffraction theory of optical imaging, *Proc. Royal Soc. of London* **217A**, pp. 408–432 (1953).
- 23-32** K. Michel, *Die Grundzüge der Theorie des Mikroskops* (Wissenschaftliche Verlagsgesellschaft M.B.H., Stuttgart, 1981).
- 23-33** Ernst Abbe, *Die Lehre von der Bildentstehung im Mikroskop*, edited by O. Lummer and F. Reiche (Vieweg, Braunschweig, 1910).
- 23-34** T. Jewell and D. L. White, Spatial frequency doubling lithography SFDL, *Journal of Lithwave Technology* Vol. 7, No. 9, 1386–1393 (1989).
- 23-35** H. Hänsel, W. Polack, Verfahren zur Herstellung einer Phasenmaske mit Amplitudenstruktur, German Democratic Republic Patent No. 126.361 (1976).
- 23-36** M.D. Levenson, Improving Resolution in Photolithography with a phase-shifting mask, *IEEE Transactions on Electron Devices* **ED-29**, 1828–1836 (1982).

24

Image Examples of Selected Objects

- 24.1 Introduction 356
- 24.2 Two-point Resolution 356
 - 24.2.1 Incoherent Versus Coherent Two-point Resolution 356
 - 24.2.2 Image of a Double Slit for Coherent and Incoherent Illumination 360
 - 24.2.3 Phase Shift and Oblique Illumination 364
- 24.3 The Image of an Edge 365
 - 24.3.1 The Coherent Image of an Amplitude and Phase Edge 365
 - 24.3.2 The Incoherent Image of an Amplitude Edge 369
 - 24.3.3 Partially Coherent Edge Image 370
 - 24.3.4 The Determination of the Optical Transfer Function from the Edge Image 375
- 24.4 The Line Image 376
 - 24.4.1 The Line Image of a Rotational-symmetrical Lens 376
 - 24.4.2 Coherent Line or Slit Image 377
 - 24.4.3 Incoherent Line or Slit Image 380
- 24.5 The Grating Image 381
 - 24.5.1 The Coherent Linear Grating Image 381
 - 24.5.2 The Coherent Grating Image with Aberrations 384
 - 24.5.3 The Influence of the Coherence Parameter σ on the Grating Image 386
 - 24.5.4 Influence of the Shape of the Effective Light Source on the Grating Image 389
 - 24.5.5 Wigner Distribution Function for Gratings, Talbot Effect and Propagation-invariant Fields 394
- 24.6 Pinhole Imaging and Quasi-point Sources 399
 - 24.6.1 Introduction 399
 - 24.6.2 Incoherent Image of a Circular Object 400
 - 24.6.3 Quasi-point Source 402
 - 24.6.4 Pinhole with Coherent Illumination 404
 - 24.6.5 Pinhole with Partial Coherent Illumination 405
 - 24.6.6 Defocusing Planes and Deconvolution 406
- 24.7 Literature 407

24.1

Introduction

In this chapter, coherent, incoherent and partially coherent imaging are compared for the examples of two-point resolution (section 24.2), the edge image (section 24.3), the line image (section 24.4) and the grating image (section 24.5). In the examples the methods for numerical evaluation of the image formation problem according to the theoretical descriptions given in chapters 21–23 are further explained and illustrated.

24.2

Two-point Resolution

24.2.1

Incoherent Versus Coherent Two-point Resolution

In the case of rotational-symmetric imaging with a circular pupil according to eq. (20-54) the ideal image of an object point is given by the Airy disc. As shown in chapter 20, section 20.3.4, the incoherent image of two points cannot be distinguished from the imaging of a line element until the points are at a certain distance. The critical distance for incoherent illumination is at $dw = 0.47$ with normalized coordinates $w = NA/\lambda \cdot r$. For coherent imaging, the resolution limit is at approximately $dw = 0.73$, and the coherent resolution limit for two-point imaging is at $\delta r \approx 0.75 \lambda/NA$. Figure 24-1 compares the intensity scans for the coherent and incoherent images of two points.

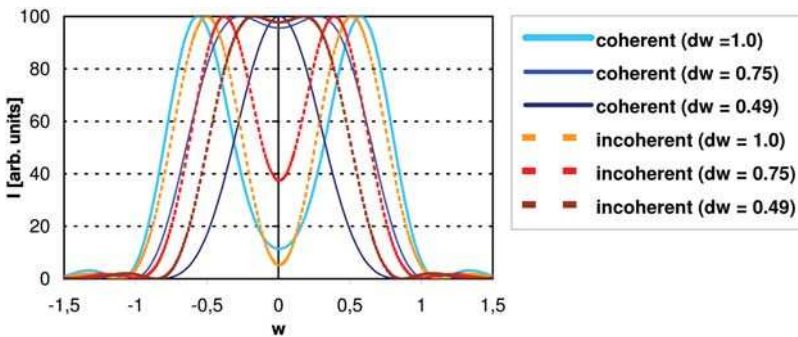


Figure 24-1: Incoherent and coherent two-point resolution.

Figure 24-2 illustrates the difference between coherent and incoherent image intensity with increasing distance between two single object points (with $NA = 0.8$). The incoherent image shows high contrast for both object point distances. For the narrow distance of $d = \lambda$, the coherent two-point image is hardly resolved. The coherent image suffers generally from larger interference effects, visible in the side maxima, due to the addition of the coherent amplitude. For a larger distance, e.g., a dis-

tance of 3λ in figure 24-2a, the interference effects between the two points may also cancel out. Figure 24-2b compares the image for a line segment of length 4λ , to the two-point image. Here, with a different distance to that in figure 24-2a, the coherence effects of the coherent two-point image do not cancel out and are much larger for the coherent image. The coherent line image length appears much shorter, while for the incoherent line image the 50% intensity value corresponds sufficiently well with the line length. At the resolution limit, the visible size of objects therefore depends on the illumination conditions.

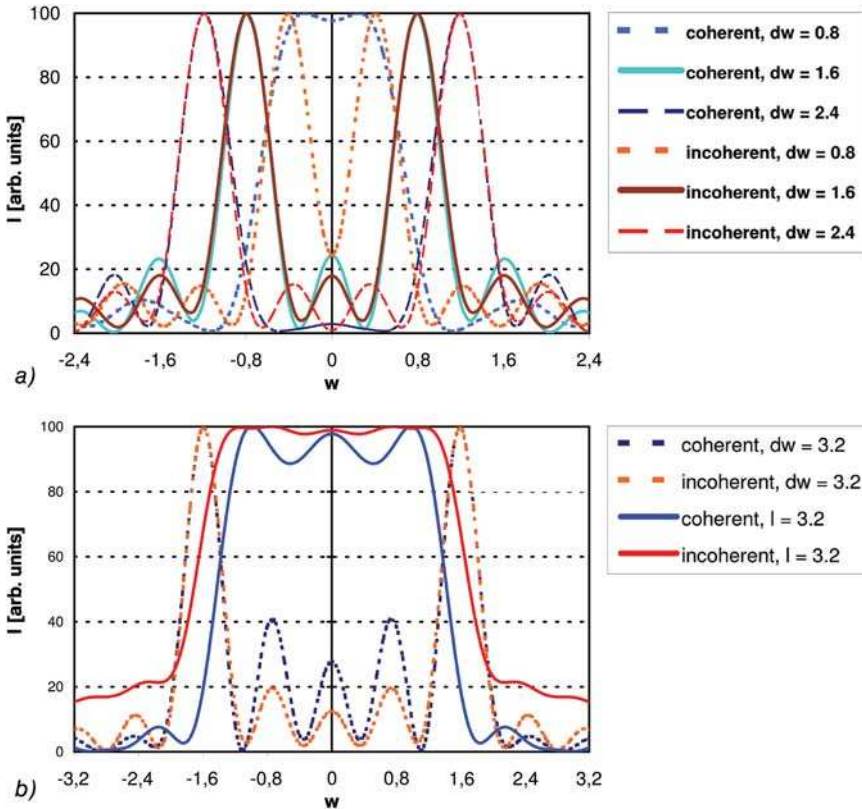


Figure 24-2: Coherent and incoherent images with numerical aperture of $NA=0.8$ with distances of: a) two point images 1λ , 2λ and 3λ ; b) two-point (4λ distance) and line image (4λ length) in comparison.

However, when taking the interference pattern into account, the coherent image is more sensitive to changes in the object. Figure 24-3 compares, on a logarithmic intensity scale, the incoherent and coherent images of two point objects with increasing distance for ideal imaging. While the incoherent images are given by linear superposition of the Airy pattern, the coherent images suffer from interference effects. In certain regions the intensity becomes much brighter, while in other regions both amplitudes cancel each other. The resulting specific interference pattern depends on the point distance.

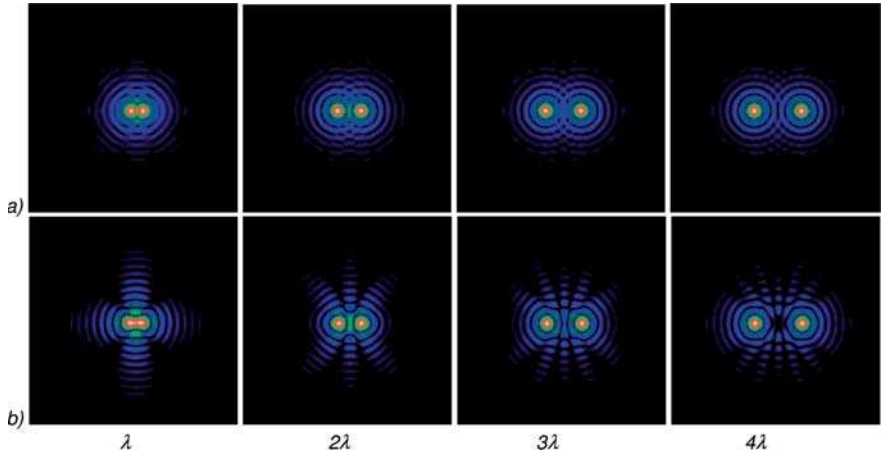


Figure 24-3: a) Incoherent and b) coherent two-point images (logarithmic scale).

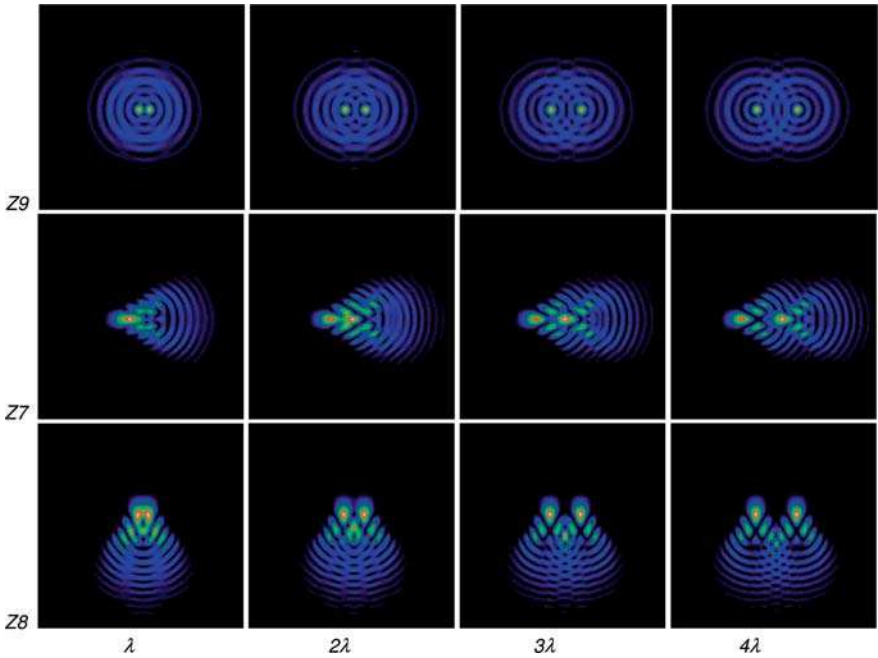


Figure 24-4: Incoherent two-point images with aberrations differing by 1λ .

Figures 24-4 and 24-5 illustrate the difference between incoherent and coherent imaging for the example of two object points at increasing distance with spherical aberration Z_9 , coma Z_7 and Z_8 of 1λ each. In each case the coherent addition of the amplitude distributions in figure 24-5 gives significantly different image intensities with the interference fringes superimposed.

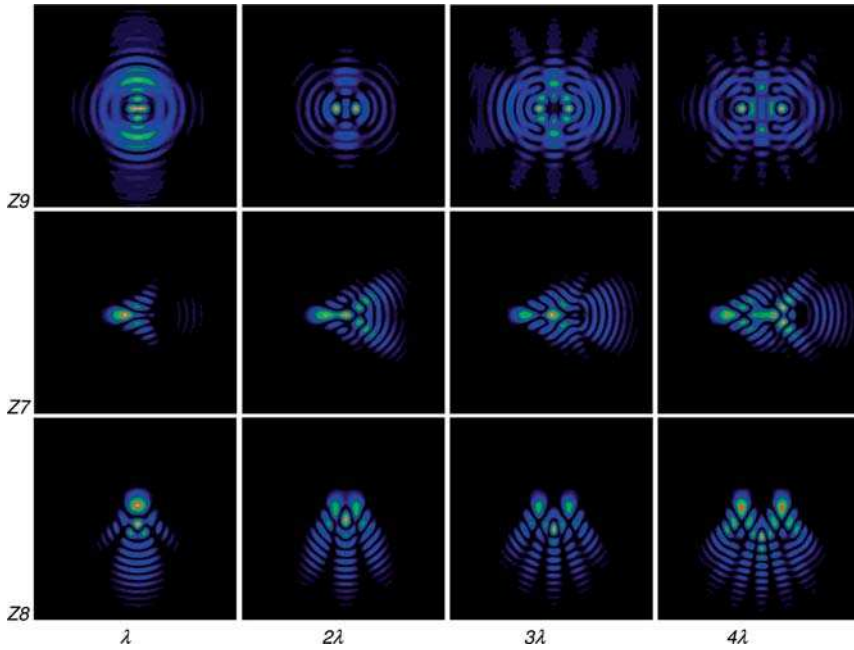


Figure 24-5: Coherent two-point images with aberrations differing by 1λ .

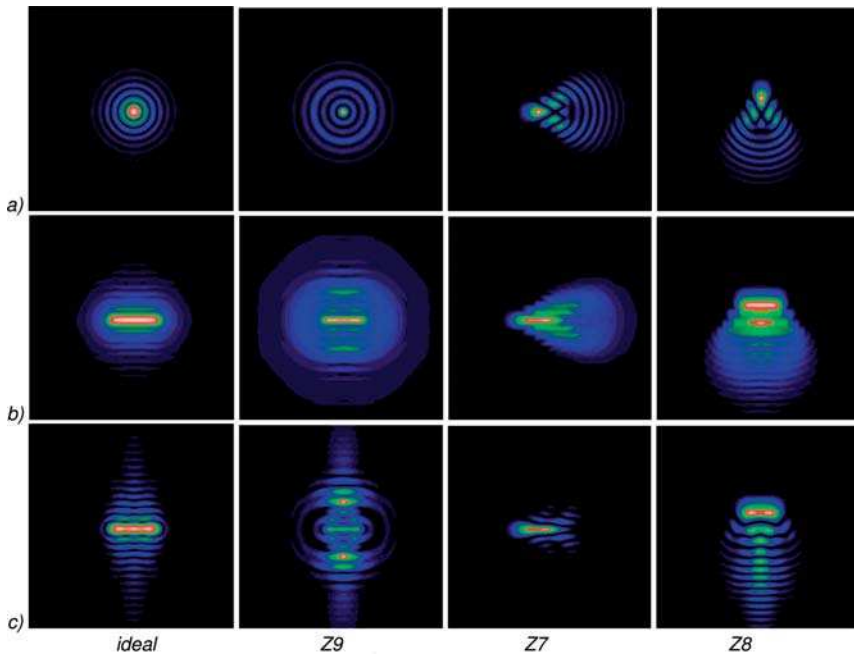


Figure 24-6: a) Point-spread function $|H(x)|^2$; b) incoherent line image; and c) coherent line image of a 4λ -line ideal and with different aberrations of 1λ .

Figure 24-6 illustrates the difference between coherent and incoherent imaging for the example of a line image with line length 4λ ($NA = 0.8$). In the example, the point-spread function $H(x)$ is disturbed by a 1λ spherical aberration Z_9 , 1λ coma Z_7 and Z_8 . The difference between coherent (c) and incoherent (b) imaging, particularly for spherical aberration, is significant!

24.2.2

Image of a Double Slit for Coherent and Incoherent Illumination

The double slit image is a convenient example of image computation using the coherence and Wigner distribution functions. The discussion is limited to the x -dimension (the double slit is assumed to be infinitely extended in the y -direction). Figure 24-7 shows the coherence function of the light source for coherent illumination with $\sigma = 0.05$ and incoherent illumination with $\sigma = 1$. In both cases – under one-dimensional consideration – the coherence function of the light source at the position of the object is given by a sinc function of width $\sim 1/2\sigma$.

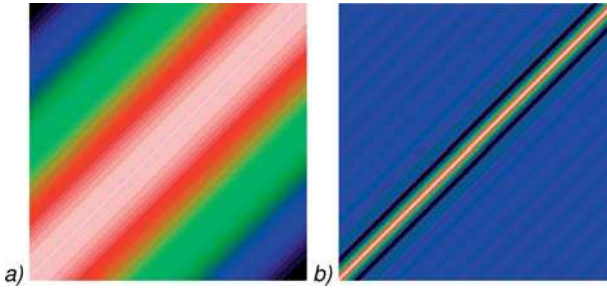


Figure 24-7: Coherence function a) of a coherent and b) an incoherent light source.

With the openings of the double slit at the coordinates x_0 and $-x_0$ given by δ -functions

$$T(x) = \delta(x - x_0) + \delta(x + x_0) \quad (24-1)$$

then for the coherence function of the double slit with a small light source S with $\sigma = 0.05$ it follows that:

$$\Gamma_{OS}(x_1, x_2) = \Gamma_S(x_1, x_2) \cdot [\delta(x_1 - x_0) + \delta(x_1 + x_0)] \cdot [\delta(x_2 - x_0) + \delta(x_2 + x_0)] \quad (24-2)$$

So the coherence function of the object wave under coherent illumination consists of four delta peaks at the positions (x_0, x_0) , $(x_0, -x_0)$, $(-x_0, x_0)$ and $(-x_0, -x_0)$ (figure 24-8a). At partially coherent illumination specified by the light source size σNA , it follows for the counter-diagonal elements of the four delta peaks of the coherence function that

$$\Gamma_{OS}(x_0, -x_0) = \int |s(\nu)|^2 e^{i2\pi 2x_0\nu} d\nu = I_S \operatorname{sinc}\left(\frac{\sigma NA}{\lambda} 2x_0\right) = I_S \operatorname{sinc}\left(\frac{\sigma NA}{\lambda} D\right) \quad (24-3)$$

with the slit distance D (in the example $D = 2x_0 = 2\lambda$). The amplitude of the disturbing contributions on the counter-diagonal of the coherence function vanishes for arguments of the sinc function which are equal to an integer:

$$\sigma = m \cdot \frac{\lambda}{D \cdot NA} \quad m = 1, 2, 3, 4, \dots \quad (24-4)$$

For the example with $NA = \sigma = 1$, $D = 2\lambda$ the condition according to eq. (24-4) is fulfilled (figure 24-8b). In general the amplitudes of the disturbing contributions are inversely proportional to the argument of the sinc function, and thus grow smaller with increasing size of the light source, with σNA and the slit distance D , respectively. After Fourier transformation there is a superposition of two cosine functions perpendicular to each other obtained in the frequency space though under the chosen conditions according to eq. (24-4) for the case of incoherent illumination one cosine function vanishes (figure 24-8d).

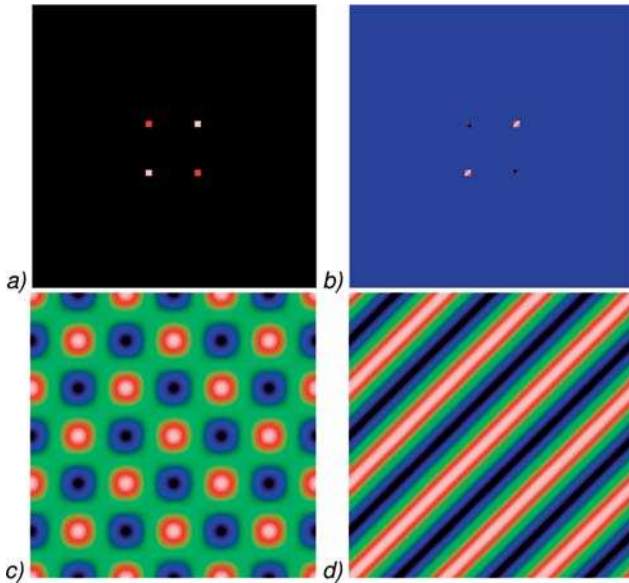


Figure 24-8: Coherence function of the double slit in the object plane with coherent (a) and incoherent (b) illumination; and after Fourier-transformation in the entrance pupil (c and d, respectively).

The coherence functions are low-pass filtered by the coherence transfer function $k(\nu_1, \nu_2) = h(\nu_1) \cdot h^*(\nu_2)$ in two directions. For a perfect optical system multiplication by the coherence transfer function corresponds to a multiplication by a two-dimensional rect function. As a consequence, in the image plane the delta functions of the coherence function are blurred by a two-dimensional sinc function. The intensity in the image space – given by the diagonals of the coherence function in figure 24-9 – consists of the two single-slit images blurred by the point-spread function and an additional superposition with the tails of the two contributions on the counter-diagonal. The disturbances correspond to the undesired interference effects with coherent illumination and depend on the ratio of the slit distance to the aperture. They are minimal for illumination divergences according to eq. (24-4).

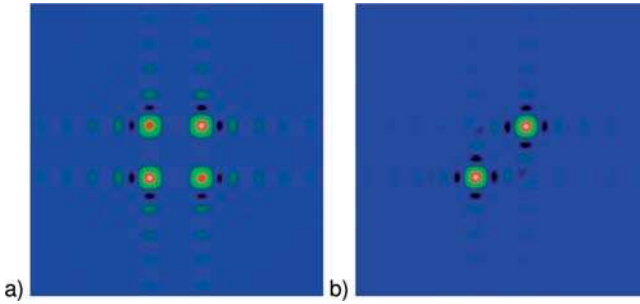


Figure 24-9: Coherence function of the double slit in the image plane for a) coherent and b) incoherent illumination.

The Wigner distribution function can be obtained from the coherence function. The coherence function $j(x, \Delta x)$ after coordinate transformation is shown in figure 24-10b starting from the coherence function Γ in the object plane (figure 24-10a). Accordingly the Wigner distribution function is obtained after a Fourier transformation in the vertical Δx direction (figure 24-10c). The analogous procedure in the image plane is shown in figure 24-11.

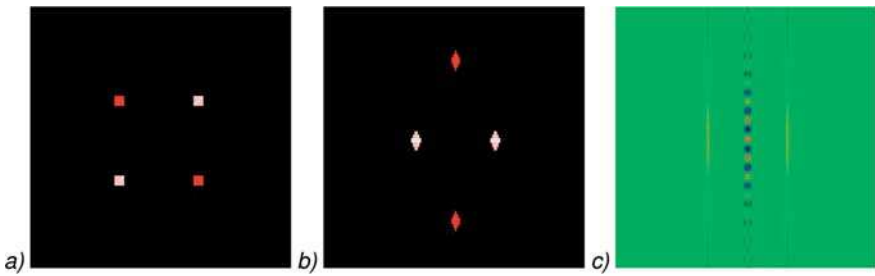


Figure 24-10: Determining the Wigner distribution function of the object wave from the coherence function Γ (a), after coordinate transformation to $j(x, \Delta x)$ (b) and Fourier transformation over Δx (c).

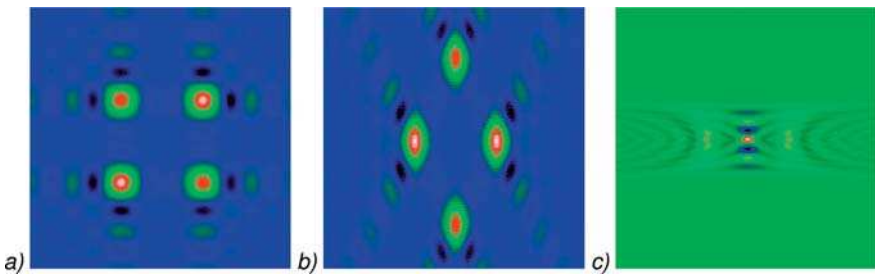


Figure 24-11: Determining the Wigner distribution function in the image plane from the coherence function Γ (a), after coordinate transformation to $j(x, \Delta x)$ (b) and Fourier transformation over Δx (c).

Figure 24-12 compares the Wigner distribution function of the double slit in the object and the image plane for coherent and incoherent illumination. Since the pro-

jection of the Wigner distribution function in the frequency direction gives the intensity, the high-frequency contributions at $x = 0$ must cancel. For incoherent illumination according to eq. (24-4) the high-frequency part of the Wigner distribution function vanishes at $x = 0$ (24-12b).

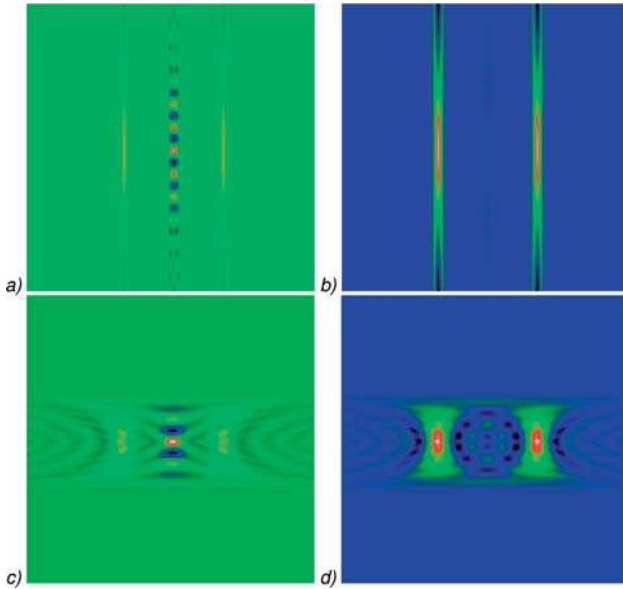


Figure 24-12: Wigner distribution function of the double slit in the object plane with: a) a coherent, b) an incoherent light source; and in the image plane with: c) a coherent, d) an incoherent light source.

The differences between coherent and incoherent illumination become particularly visible at defocusing. Figure 24-13 compares intensity scans at different defocus positions in normalized coordinates $w = \lambda/NA \cdot x$ and Rayleigh units $RU = \lambda/2NA^2$. The Wigner distribution function in figure 24-12c at $x = 0$ shows a modulation in the ν direction at $x = 0$ which for free-space propagation between the two peaks of the double slit leads to increased modulations (figure 24-13a). These modulations are less visible at incoherent illumination (figure 24-13b).

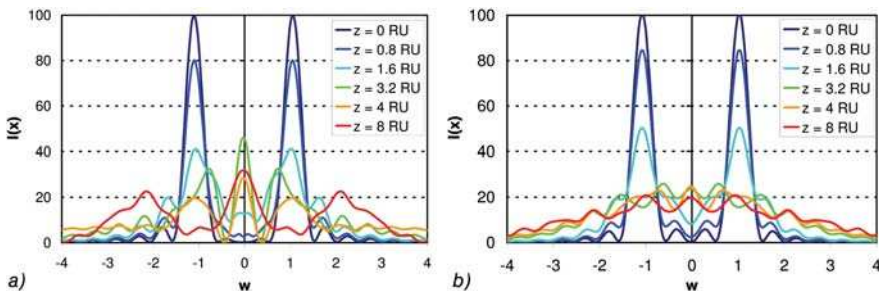


Figure 24-13: Diffraction effects at the double slit for defocusing: a) at coherent illumination; b) at incoherent illumination.

24.2.3

Phase Shift and Oblique Illumination

It has been recognized that the two-point resolution depends on the phase relation between the two waves emitted from, e.g., a double slit. As illustrated in figure 24-14, different phases of the waves emitted from the two slits may be obtained either by oblique illumination or by phase shifting elements in the slit apertures [24-1], [24-2].

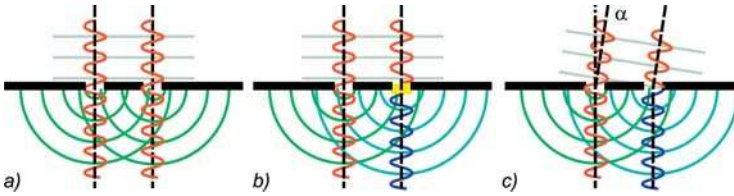


Figure 24-14: a) Conventional double slit; b) with phase shifter; and c) for oblique illumination.

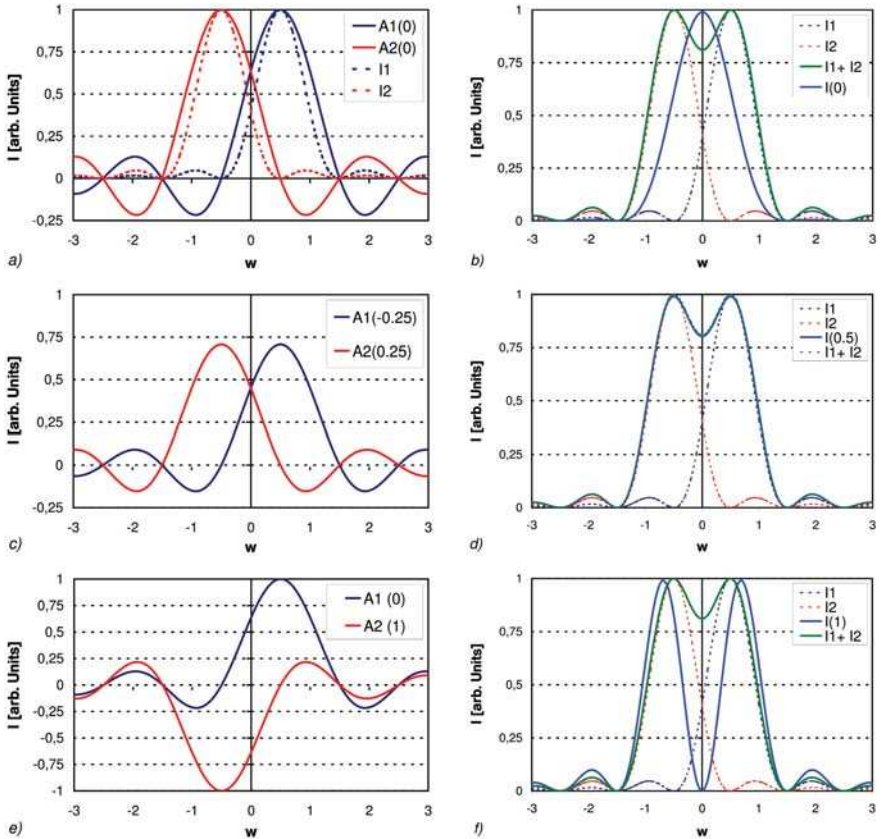


Figure 24-15: Incoherent and coherent images $I(\varphi)$ of two lines at a distance w with a phase difference of a) and b) $\varphi = 0$, c) and d) $\varphi = \pm 0.25\pi$, and e) and f) $\varphi = \pi$.

Figure 24-15 compares the resulting intensities $I(\varphi)$ of the image of the double slit at a distance $d = 1w = \lambda/NA$ with the different phases: $\varphi = 0$, $\varphi = 0.5\pi$ and $\varphi = \pi$. For incoherent imaging the intensities simply add up ($I_1 + I_2$). For coherent imaging the complex amplitudes are added before the intensity is formed by the square of the sum of amplitudes.

Without a phase difference, the incoherent double-slit image is not resolved (24-15a–b). For a phase difference of $\varphi = \pi/2$, the incoherent case is reproduced (figure 24-15c and d). For an amplitude in phase opposition $\varphi = \pi$, the coherent slit image is fully resolved at maximum contrast, but the maxima appear at a larger slit distance. The opposing phase condition for a double slit distance d or grating period d is reached, when the illumination angles to the optical axis satisfies the following condition [24-1], [24-2].

$$n \sin \alpha \cdot d = \pm (2n + 1) \frac{\lambda}{2} . \quad (24-5)$$

As shown in Figure 24-15, the intensity distribution for the coherent image $I(\varphi)$, especially the apparent slit distance for the image, depends on the relative phase difference φ of the amplitudes A_1 and A_2 of the single lines. In particular, for partially coherent image formation, the appearance of the image, e.g., the image line length or the distance of two narrow point or line images and the ‘contrast’ (i.e., the minimum intensity value between the double slit) is influenced by the coherence properties and the obliqueness of the illuminating light. This effect, further discussed below for the examples of the edge, slit and grating images, has to be considered when measuring objects with fine detail. Since in the vector model when considering exact boundary conditions and diffraction effects, the resulting phase shifts often cannot be uniquely determined, and optical metrology of fine structures is in general of limited accuracy. On the other hand, the dependence of the image on the phase or illumination conditions can be utilized in, e.g., lithographic imaging for contrast enhancement of fine details by the application of phase shifters and oblique coherent illumination [24-3], [24-4].

24.3

The Image of an Edge

24.3.1

The Coherent Image of an Amplitude and Phase Edge

The image of a linear edge is illustrated in figure 24-16 and figure 24-17. The steps for coherent image formation consist of Fourier transformation, low-pass filtering, and again Fourier transformation.

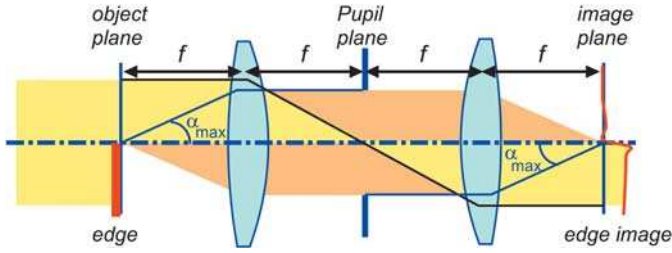


Figure 24-16: Coherent image of an amplitude edge.

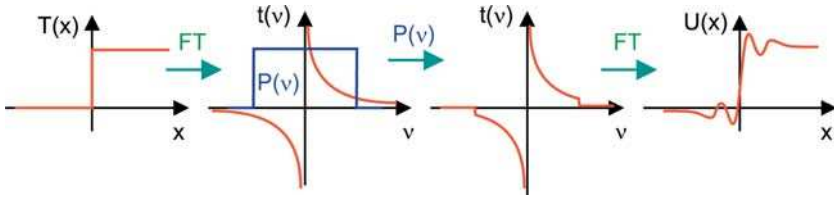


Figure 24-17: Illustration of the imaging of an edge in Fourier steps.

The transmission function of an amplitude edge can be described by the step function $\Theta(x)$

$$T(x) = 0.5 \cdot (\text{sign}(x) + 1) = \Theta(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0. \end{cases} \quad (24-6)$$

According to eq. (21-34), in the simplest case of perpendicular illumination with a plane wave, the image amplitude is given by

$$U(x) = \int t(v) \cdot h(v) \cdot e^{2\pi i v x} dv = \int t(v) \cdot P(v) \cdot e^{\frac{i2\pi}{\lambda} W(v)} e^{2\pi i v x} dv . \quad (24-7)$$

After insertion of the diffraction spectrum of a linear, infinitely extended edge given by

$$t(v) = 0.5 \cdot \left(\delta(0) + \frac{1}{\pi i v} \right) \quad (24-8)$$

into eq. (24-7) it follows that the coherent image amplitude is

$$\begin{aligned} U(x) &= \frac{1}{2} \cdot \int_{v=-\frac{NA}{\lambda}}^{\frac{NA}{\lambda}} \left[\delta(0) + \frac{1}{\pi i v} \right] \cdot e^{\frac{i2\pi}{\lambda} W(v)} \cdot e^{2\pi i v x} dv \\ &= \frac{1}{2} + \frac{1}{2\pi i} \int_{v=-\frac{NA}{\lambda}}^{\frac{NA}{\lambda}} \frac{1}{v} \cdot e^{\frac{i2\pi}{\lambda} W(v)} \cdot e^{2\pi i v x} dv . \end{aligned} \quad (24-9)$$

For the aberration-free optical system eq. (24-9) can be solved with the integral-sinus $\text{Si}(x)$:

$$U(x) = \frac{1}{2} + \frac{1}{\pi} \int_{\nu=0}^{\frac{NA}{\lambda}} \frac{\sin(2\pi\nu x)}{\nu} d\nu = \frac{1}{2} + \frac{1}{\pi} \text{Si}\left(2\pi \cdot \frac{NA}{\lambda} \cdot x\right). \quad (24-10)$$

The image intensity is thus given by

$$I_{\text{edge}}^{(\text{coh})}(x) = \left[\frac{1}{2} + \frac{1}{\pi} \cdot \text{Si}\left(\frac{2\pi \cdot NA \cdot x}{\lambda}\right) \right]^2. \quad (24-11)$$

Figure 24-18 shows the amplitude and intensity of the amplitude edge image in normalized units of $\nu = 2\pi \cdot x \cdot NA/\lambda$. The amplitude $U(\nu)$ shows oscillating behaviour on both sides of the edge due to the low-pass filtering of the optical system. In the intensity image, the diffraction effect is intensified on the bright side of the edge, while in the shadow region it is only visible on a logarithmic scale (figure 24-19).

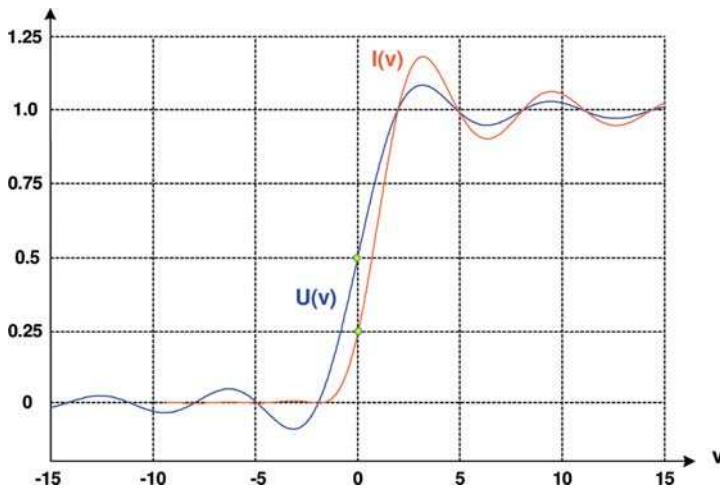


Figure 24-18: Amplitude and intensity distribution of the coherent image of an amplitude edge.

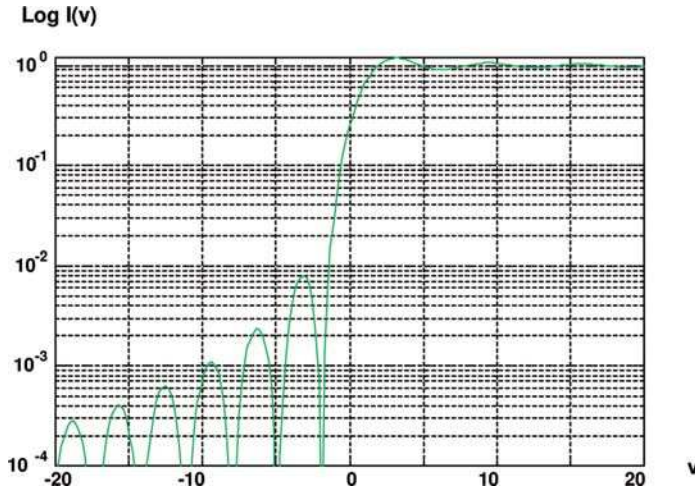


Figure 24-19: Intensity distribution of the coherent image on a logarithmic scale.

As indicated in figure 24-18, the amplitude $U(0)$ at the geometrical image position of the edge is 0.5, and consequently the intensity $I(0) = 0.25$. For the determination of the position of an edge, frequently the 50%-value of the intensity distribution is taken. Coherent images appear smaller than the objects, since in the coherent image the edge is shifted into the bright region by

$$\Delta x = 0.212 \cdot \frac{\lambda}{NA}. \quad (24-12)$$

In the special case of a coherent image of a phase edge with a phase difference of π , the transmission function is simply given by the sign-function:

$$T(x) = \text{sign}(x) \quad (24-13)$$

The image amplitude is given by

$$U(x) = \frac{2}{\pi} \text{Si} \left(2\pi \frac{NA}{\lambda} x \right) \quad (24-14)$$

and the intensity by

$$I_{\pi\text{-edge}}^{\text{(coh)}}(x) = \left[\frac{2}{\pi} \cdot \text{Si} \left(\frac{2\pi \cdot NA \cdot x}{\lambda} \right) \right]^2. \quad (24-15)$$

24.3.2

The Incoherent Image of an Amplitude Edge

The incoherent image intensity of an amplitude edge can be derived from

$$I(x) = |T(x)|^2 \otimes |H(x)|^2. \quad (24-16)$$

The square of the amplitude edge is simply given by the edge function itself, and the Fourier transform of the point-spread function is inserted in the spectrum representation of eq. (24-16) the optical transfer function $g(\nu)$ as:

$$I_{\text{edge}}^{\text{incoh}}(x) = \int t(\nu) \cdot [P(\nu) \otimes P(\nu)] e^{2\pi i x \nu} d\nu = \int t(\nu) \cdot g(\nu) \cdot e^{2\pi i x \nu} d\nu. \quad (24-17)$$

In a linear one-dimensional system approximation with $g(\nu)$ according to eq. (21-64), it follows for the intensity in image plane that

$$\begin{aligned} I_{\text{edge}}^{\text{incoh}}(x) &= \frac{1}{2} \cdot \int_{\nu=-\frac{2NA}{\lambda}}^{\frac{2NA}{\lambda}} \left[\delta(0) + \frac{1}{\pi i \nu} \right] \cdot \left[1 - \frac{\lambda}{2NA} \text{sign}(\nu) \cdot \nu \right] e^{2\pi i x \nu} d\nu \\ &= \frac{1}{2} + \frac{1}{\pi} \text{Si} \left(4\pi \frac{NA}{\lambda} x \right) + \frac{1}{\pi} \frac{\left[\cos \left(4\pi \frac{NA}{\lambda} x \right) - 1 \right]}{4\pi \frac{NA}{\lambda} x}. \end{aligned} \quad (24-18)$$

In contrast to the coherent image, the incoherent image is almost free of interference effects and the modulations of the coherent image vanish. However, as shown in figure 24-20, the edge image is blurred by the low-pass filtering of the optical imaging. The 50% threshold value coincides with the geometrical edge position, therefore the 50% threshold value gives reasonably good image positions of edges.

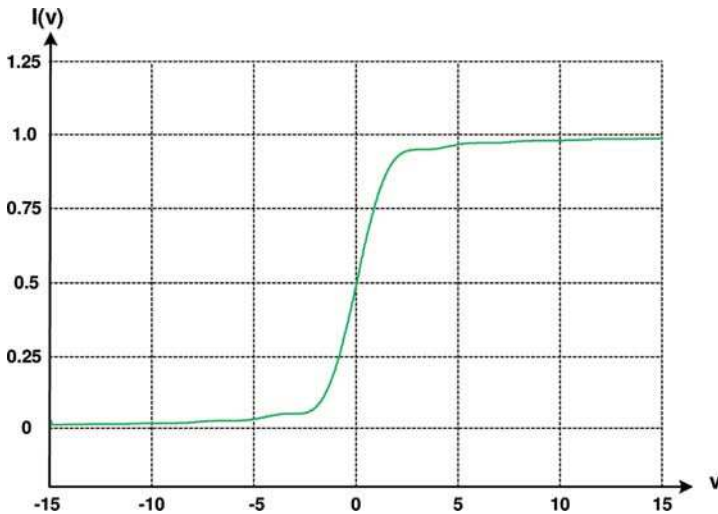


Figure 24-20: Example of an intensity distribution of the incoherent 1D-image of an edge.

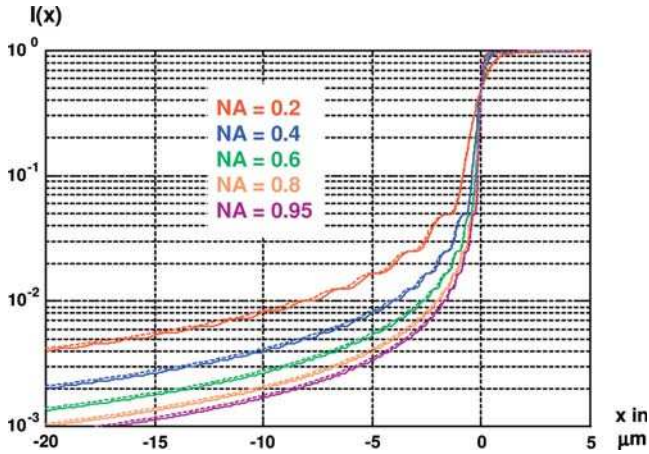


Figure 24-21: Intensity distributions of the incoherent 1D- image of an edge for different numerical apertures for monochromatic illumination (solid lines, $\lambda = 660 \text{ nm}$) and illumination with finite bandwidth of $\pm 30 \text{ nm}$ (dashed lines).

Figure 24-21 shows the intensity distribution of the incoherently imaged edge on a logarithmic scale for different numerical apertures. With decreasing numerical aperture, the impact of the low-pass filtering increases, leading to an increased blurring of the edge. In figure 24-21 the incoherent polychromatic intensity for a light source with bandwidth $\pm 30 \text{ nm}$ is shown in comparison with the monochromatic images. The small bandwidth has a negligible impact on the image. With increasing wavelength the diffraction angles also increase, therefore the effective numerical aperture decreases with increasing wavelength. Therefore, since the diffraction effects increase with the wavelength, the polychromatic image of an amplitude edge shows coloured fringes in the neighbourhood of the edge, which are dominated by the longer wavelengths.

24.3.3

Partially Coherent Edge Image

The partially coherent edge image can be calculated with reasonable accuracy by assuming a linear one-dimensional imaging system by application of the Hopkins transmission cross-coefficient [24-5], [24-6]. Usually the application of the transmission cross-coefficient is reduced to the linear part as it is illustrated in figure 22-1. The linear *TCC* is then only a function of one frequency. In general, however, the *TCC* is a four-dimensional function and reduces to a two-dimensional function for a one-dimensional object. Some examples have been given by Sheppard [24-7]. In the special case of one-dimensional imaging the Hopkins transmission cross-coefficient and thus the image intensity computation according to eq. (22-3), can be approximated analytically. For conventional partially coherent illumination with an effective source of diameter $2\sigma \cdot NA/\lambda$ in one dimension, the transmission cross-coefficient is given by

$$TCC(v_1, v_2) = \int \text{rect}\left(\frac{v_q}{2\sigma NA/\lambda}\right) \cdot \text{rect}\left(\frac{v_1 + v_q}{2NA/\lambda}\right) \cdot \text{rect}\left(\frac{v_2 + v_q}{2NA/\lambda}\right) dv_q. \quad (24-19)$$

Figure 24-22 compares different one-dimensional imaging TCC for different light-source diameters given by 2σ .

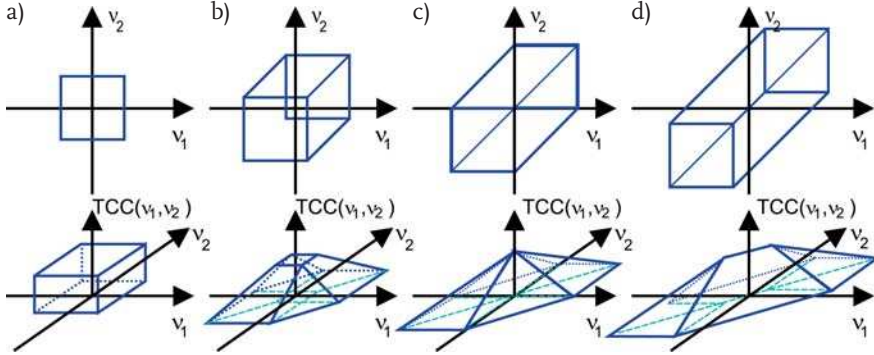


Figure 24-22: One-dimensional imaging Transmission Cross-Coefficient for different degrees of coherence: a) coherent illumination; b) partially coherent illumination with $0 < \sigma < 1$; c) incoherent illumination with $\sigma = 1$; d) incoherent illumination with $\sigma > 1$.

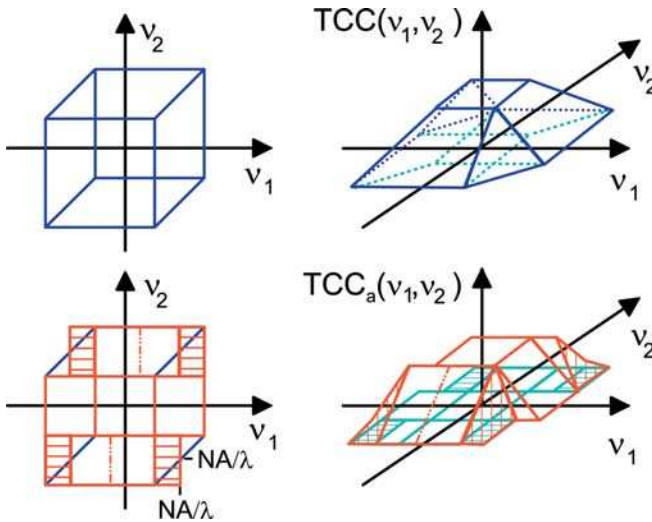


Figure 24-23: One-dimensional imaging Transmission Cross-Coefficient and approximation.

As illustrated in figure 24-23, the 1D- TCC can, according to Glindemann [24-5], [24-6], be approximated by separation of variables v_1 and v_2 :

$$TCC_a(v_1, v_2) = TCC(v_1, 0) \cdot \text{rect}\left(\frac{\lambda \cdot v_2}{2(1 - \sigma)NA}\right) + TCC\left(v_1, \frac{NA}{\lambda}\right) \cdot \text{rect}\left(\frac{\lambda \cdot v_2 - NA}{2\sigma NA}\right) \cdot \frac{NA(1 + \sigma) - \lambda \cdot v_2}{\sigma NA}. \quad (24-20)$$

The linear $TCC(v_1, 0)$ of eq. (24-20) can easily be derived, and the $TCC(v_2, NA/\lambda)$ corresponds to a shifted linear TCC for oblique illumination with obliquity $\sigma_0 = (NA - \sigma)/2$. After insertion into eq. (22-3) we obtain for the real part of the intensity I_R :

$$\begin{aligned}
 I_R(x) = & \int TCC(v_1, 0) \cdot t(v_1) \cdot e^{i2\pi x \cdot v_1} dv_1 \cdot \text{frect}\left(\frac{\lambda \cdot v_2}{2(1 - \sigma)NA}\right) \cdot t^*(v_2) \cdot e^{-i2\pi x \cdot v_2} dv_2 \\
 & + \int TCC\left(v_1, \frac{NA}{\lambda}\right) \cdot t(v_1) \cdot e^{i2\pi x \cdot v_1} dv_1 \\
 & \cdot \text{frect}\left(\frac{\lambda \cdot v_2 - NA}{2\sigma NA}\right) \cdot \frac{NA(1 + \sigma) - \lambda \cdot v_2}{\sigma NA} \cdot t^*(v_2) \cdot e^{-i2\pi x \cdot v_2} dv_2.
 \end{aligned}
 \tag{24-21}$$

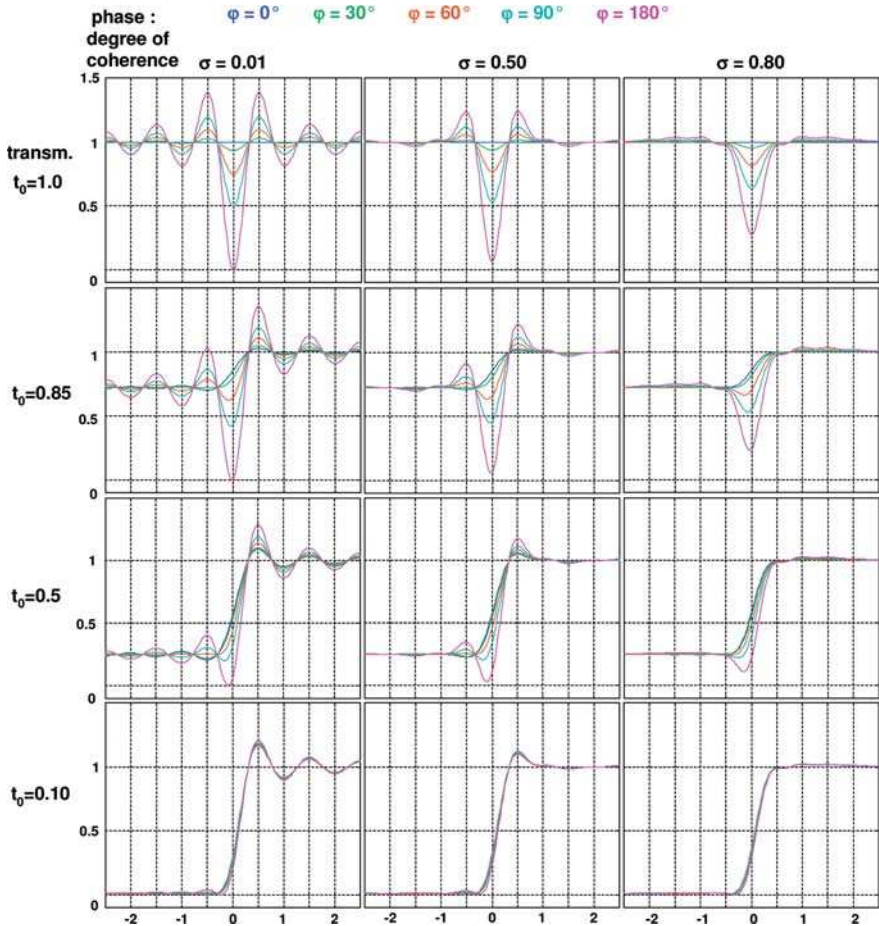


Figure 24-24: Partially coherent one-dimensional image of a straight edge of transmission t_0 , phase φ and partial coherence σ .

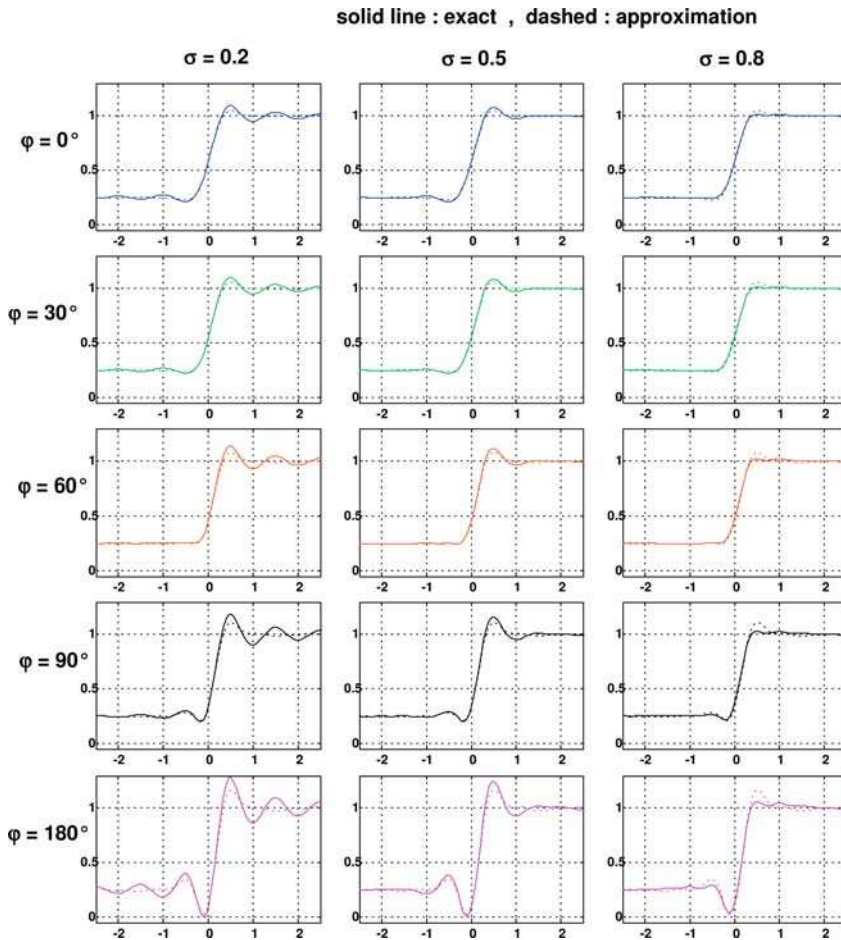


Figure 24-25: Approximation according to eq. (24-21) in comparison with exact one-dimensional partially coherent image of an edge.

For further details of the solution of eq. (24-21) the reader is referred to Glindemann [24-5], [24-6]. In the following, solutions for the imaging of a linear straight edge obtained by the solution of eq. (24-34) will be given. Figure 24-24 compares the partially coherent images of generally complex linear edge functions in dependence on the coherence parameter σ , the phase step of the edge and the transmission t_0 of the opaque side of the edge. Figure 24-25 illustrates the difference from the correct simulation. There is a reasonable agreement only for low σ and for small phase shifts for the diffraction effects, while the edge position is found to be in good agreement with the one-dimensional simulation result. As illustrated in figure 24-26a, for a pure amplitude edge, the position Δx , which is determined as the 50%-value of the intensity distribution, as well as the slope dI/dx of the partially coherent edge image, depends on the coherence parameter σ . For telecentric imaging, the edge

image position Δx remains invariant under defocus and spherical aberration, while asymmetrical aberrations like coma, influence the apparent edge position (figure 24-26b).

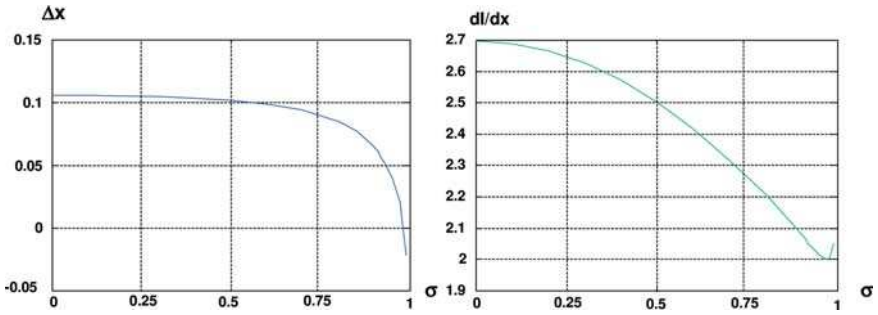


Figure 24-26a: For an amplitude edge, the position Δx as well as the slope dl/dx of the partially coherent edge image depends on the coherence parameter σ .

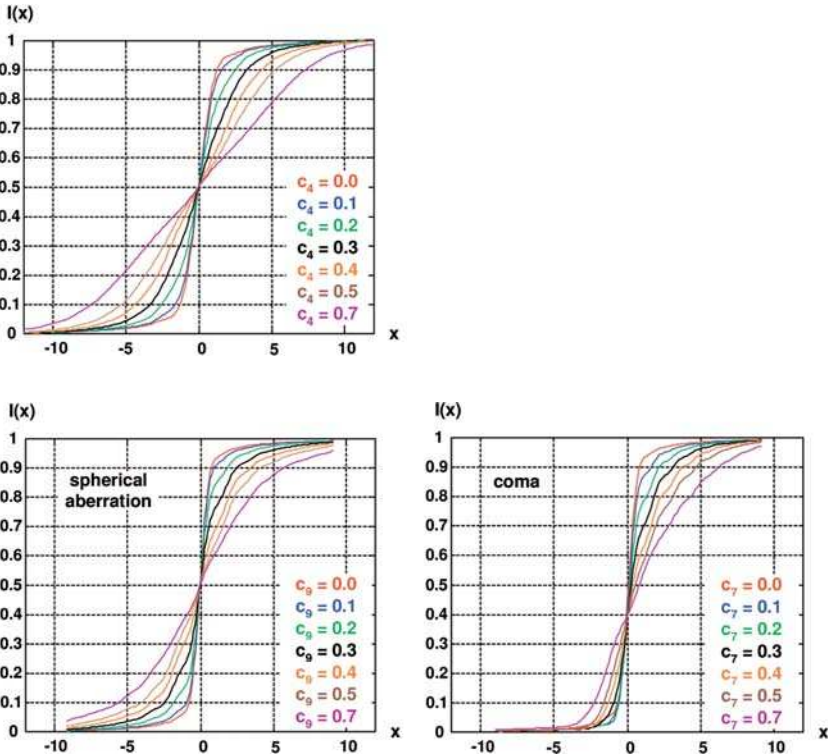


Figure 24-26b: Incoherent one-dimensional images of an amplitude edge under: defocus (top); spherical aberration (bottom left); and coma (bottom right), with Zernike coefficients c_4 , c_9 and c_7 .

24.3.4

The Determination of the Optical Transfer Function from the Edge Image

As illustrated in figure 24-27, the edge image provides a method for the determination of the transfer function of optical systems. After determination of the edge image the line-spread function (LSF) is achieved, and the optical transfer function (OTF) is obtained by Fourier transformation.

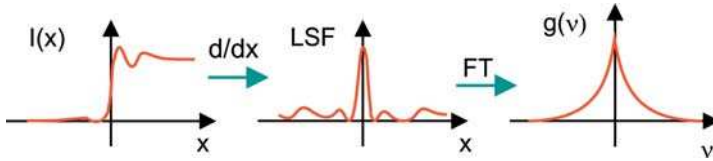


Figure 24-27: Schematic illustration of the differentiation method for determining the transfer function of optical systems.

According to eq. (24-17) the edge image intensity $I(x)$ is given by

$$I_{\text{edge}}^{\text{incoh}}(x) = \frac{1}{2} + \frac{1}{2\pi i} \int_{-\infty}^{\infty} g(v) e^{2\pi i x v} dv \quad (24-22)$$

it follows, with the differentiation rule, that

$$\frac{d I_{\text{edge}}^{\text{incoh}}}{dx} = LSF(x) = \int g(v) e^{2\pi i x v} dv \quad (24-23)$$

with the line-spread function LSF (see chapter 24.4). The incoherent optical transfer function OTF can be calculated by inverse Fourier transformation

$$g(v) = FT[LSF(x)] = FT \left[\frac{d I_{\text{edge}}^{\text{incoh}}}{dx} \right]. \quad (24-24)$$

As illustrated, for the determination of the OTF from measured edge images, the intensity has to be differentiated, which is difficult as it is sensitive to noise at the edge itself, especially for large numerical apertures. Therefore, for determining the optical transfer function, usually methods other than the differentiation method are preferred. Nevertheless, it gives a good example of the application of Fourier theory in optical imaging.

24.4
The Line Image

24.4.1
The Line Image of a Rotational-symmetrical Lens

The idealized line image as the image of a delta-like line is the special case of the slit image of infinitely small width. The general slit function is considered below. The line image or line-spread function LSF is obtained after integration of the point-spread function PSF in one direction (figure 24-28):

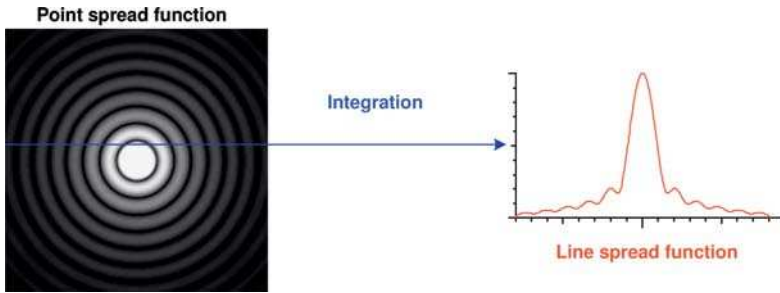


Figure 24-28: Line-spread function LSF given by integration of the PSF in one direction.

$$I_{LSF}(x) = \int I_{PSF}(x, y) dy. \tag{24-25}$$

According to the result of the integration the line-spread function does not have any zeros and the intensity is less modulated than for the point-spread function. The intensity decays in steps. Figure 24-29 illustrates the line-spread function in more details.

The line-spread function is given by the Fourier transform of an intersection through the incoherent transfer function, as can be derived after Fourier transformation of eq. (24-25)

$$\begin{aligned} I_{LSF}(x) &= \int \left| \iint P(v_x, v_y) \cdot e^{-2\pi i(v_x x + v_y y)} dv_x dv_y \right|^2 dy \\ &= \iint P(v_x, v_y) \otimes P^*(v_x, v_y) \cdot e^{-2\pi i(v_x x + v_y y)} dv_x dv_y dy \\ &= \int g(v_x, 0) \cdot e^{-2\pi i v_x x} dv_x. \end{aligned} \tag{24-26}$$

With the ideal transfer function for $v > 0$ in normalized coordinates ($NA = 1$)

$$g(v_x, 0) = \frac{2}{\pi} \left(\arccos v_x - v_x \cdot \sqrt{1 - v_x^2} \right) \tag{24-27}$$

and from the fact that $g(v_x, 0)$ is real, it follows that

$$I_{LSF}(x) = 2 \int_0^1 g(v_x, 0) \cdot \cos(2\pi \cdot x \cdot v_x) dv_x. \tag{24-28}$$

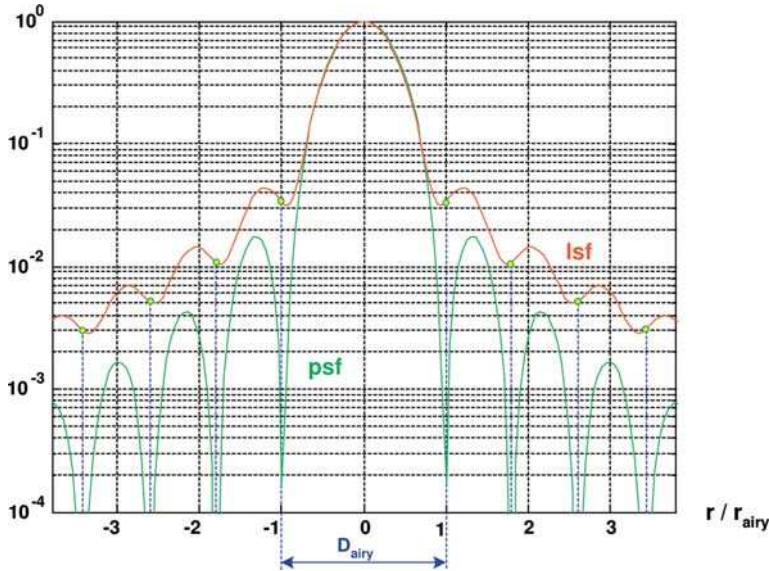


Figure 24-29: Line-spread function LSF of an ideal imaging system.

With the help of the Struve function of order m , defined by

$$H_m(z) = \frac{2(z/2)^m}{\sqrt{\pi} \cdot \Gamma(m + 1/2)} \cdot \int_0^1 (1 - t^2)^{m-1/2} \cdot \sin(zt) dt \quad (24-29)$$

and the recursion formula

$$H_{m-1}(z) + H_{m+1}(z) = \frac{2m}{z} \cdot H_m(z) + \frac{(z/2)^m}{\sqrt{\pi} \cdot \Gamma(m + 3/2)} \quad (24-30)$$

one obtains after insertion of (24-27) into (24-26) and partial integration the final result for the idealized line-spread function, normalized to one [24-8]:

$$I_{LSF}^{(ideal)}(x) = \frac{3\pi \cdot H_1\left(\frac{2\pi \cdot NA}{\lambda} \cdot x\right)}{2 \cdot \left(\frac{2\pi \cdot NA}{\lambda} \cdot x\right)^2} \quad (24-31)$$

24.4.2

Coherent Line or Slit Image

The complex transfer functions of a slit or a line (= bar) can now be constituted by step functions (see figure 24-30 and figure 24-31):

$$T_{slit}(x) = \text{rect}\left(\frac{x}{2a}\right) = \Theta(x + a) - \Theta(x - a), \quad (24-32a)$$

$$\begin{aligned}
 T_{\text{bar}}(x) &= \text{rect}\left(\frac{x}{2a}\right) = \Theta(x-a) + \Theta(-x-a) \\
 &= 1 - \Theta(x+a) + \Theta(x-a).
 \end{aligned}
 \tag{24-32b}$$

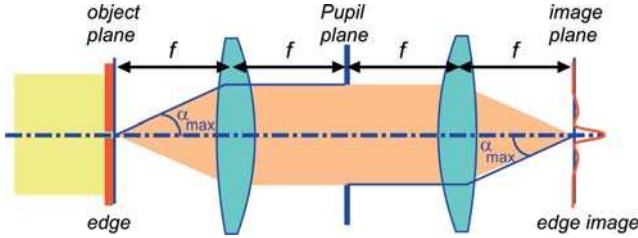


Figure 24-30: Imaging of a slit (transparent line) by a 4f-system.

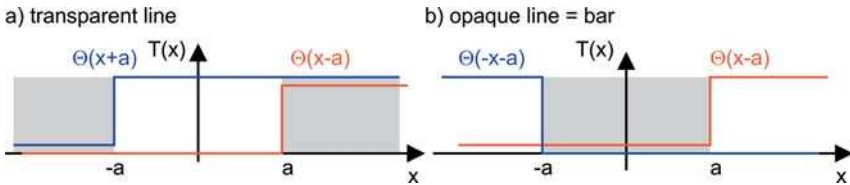


Figure 24-31: Transparent line (a) and line (= bar) (b) constituted by step functions.

The coherent images are consequently given by the linear superposition of the amplitude slit images. The intensity is given by the squared value:

$$I_{\text{slit}}^{(\text{coh})}(x) = \frac{1}{\pi^2} \left[\text{Si}\left(2\pi \frac{NA}{\lambda}(x+a)\right) - \text{Si}\left(2\pi \frac{NA}{\lambda}(x-a)\right) \right]^2.
 \tag{24-33a}$$

$$I_{\text{bar}}^{(\text{coh})}(x) = \frac{1}{\pi^2} \left[\text{Si}\left(2\pi \frac{NA}{\lambda}(x-a)\right) - \text{Si}\left(2\pi \frac{NA}{\lambda}(x+a)\right) \right]^2.
 \tag{24-33b}$$

Figure 24-32 shows an example of a slit image with $a = 20$. From eq. (24-33) it can be seen that the difference in the arguments of the integral-sine-functions $\text{Si}(x)$ scale with the product of $NA \cdot a$. For $NA \cdot a = 0$ the intensity vanishes. With increasing slit width or increasing numerical aperture the difference of the arguments increase, and, as a consequence, the diffraction effects decrease. This is illustrated in figure 24-33, in scaled coordinates x/a , for different slit widths in units of λ/NA .

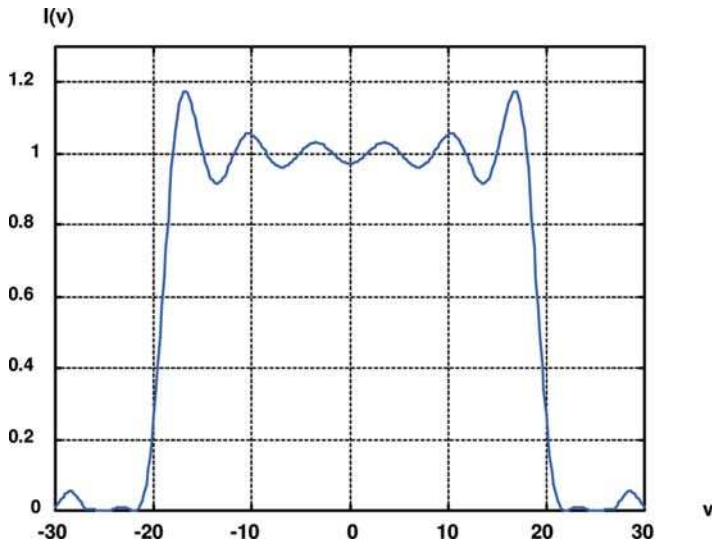


Figure 24-32: Example of a coherent slit image in normalized coordinates with $a = 20$.

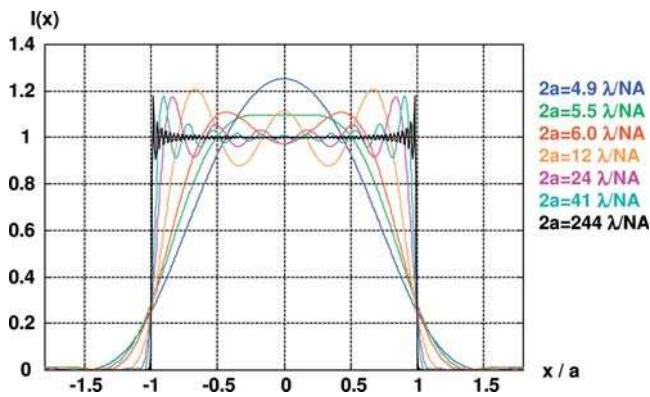


Figure 24-33: Coherent slit images for different numerical apertures in scaled units x/a .

With the 50% threshold of the intensity the apparent slit width deviates from the original with decreasing $a \cdot \lambda / NA$ (figure 24-34). The derivation of the slit width from the coherent image gives, for a threshold value of approximately 25%, one order of magnitude better results than for the usual 50% threshold. The result of figure 24-34 is normalized to intensity while figure 24-33 shows the images with constant energy.

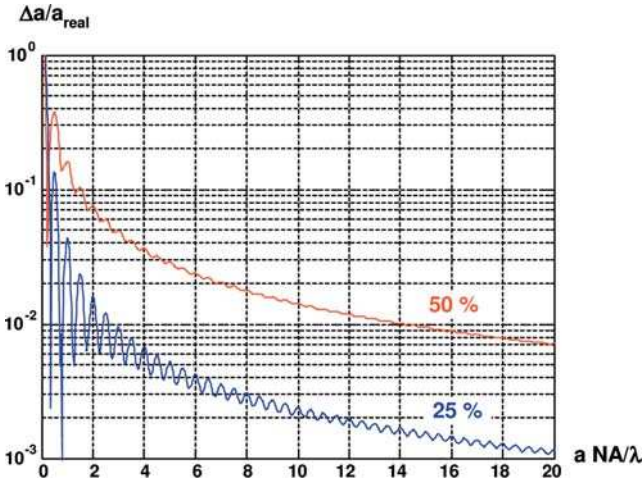


Figure 24-34: Relative difference of apparent slit width of the coherent image from the object slit width for the 25% and the 50% intensity threshold.

24.4.3

Incoherent Line or Slit Image

The incoherent image of a slit, illustrated in figure 24-35, can now be evaluated after insertion of eq. (24-33a) into eq. (24-17). For the intensity it follows that

$$I_{\text{slit}}^{\text{incoh}}(x) = \frac{1}{\pi} \left[\text{Si}(x+c) - \text{Si}(x-c) + \frac{[\cos(x+c) - 1]}{(x+c)} - \frac{[\cos(x-c) - 1]}{(x-c)} \right] \quad (24-34)$$

with the normalized slit half-width $c = 4\pi \cdot a \cdot NA/\lambda$. Figure 24-35 illustrates incoherent slit images for some examples of c .

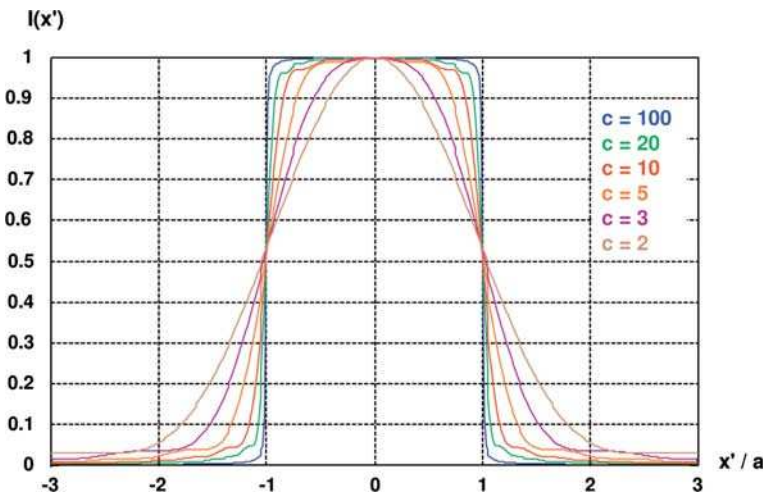


Figure 24-35: Incoherent slit image for different slit half-widths c in scaled coordinates x/a .

24.5
The Grating Image

24.5.1
The Coherent Linear Grating Image

In case of symmetrical aberrations or aberration-free imaging, the image intensity of a grating can be written, according to eq. (22-28), as a series of cosine functions:

$$I(x, y) = b_0 + b_1 \cdot \cos\left(\frac{2\pi}{d} \cdot x\right) + b_2 \cdot \cos\left(\frac{4\pi}{d} \cdot x\right) + \dots + b_n \cdot \cos\left(\frac{2n \cdot \pi}{d} \cdot x\right) \tag{24-35}$$

with the coefficients b_n given from eq. (22-27a). The maximum value for n is given by the number of diffraction orders which are passing the lens, minus one. Considering the zero and first diffraction orders only, the coefficients b_n are given by

$$b_0 = |g_0|^2 \cdot TCC(0, 0) + |g_1|^2 \cdot TCC(1, 1) + |g_{-1}|^2 \cdot TCC(-1, -1), \tag{24-36a}$$

$$b_1 = 2 \cdot \text{Re}[g_1 g_0^* \cdot TCC(1, 0) + g_0 g_{-1}^* \cdot TCC(0, -1)], \tag{24-36b}$$

$$b_2 = 2 \cdot \text{Re}[g_1 g_{-1}^* \cdot TCC(1, -1)]. \tag{24-36c}$$

For an aberration-free imaging system passing a finite number of diffraction orders without truncation of one of the diffraction orders, the TCC 's are equal for all m and n and the image contrast of a binary Ronchi grating with aspect ratio 1:1 is $V = 100\%$, independent of shape and size of the light source. An example with three diffraction orders is illustrated in figure 24-36a. Two exemplary intensity distributions formed by interference patterns with three diffraction orders where $m \leq 1$ and five diffraction orders where $m \leq 3$, are illustrated in figure 24-37. Here, with the Ronchi grating as object the $\pm 2^{\text{nd}}$ diffraction orders vanish. The image formed by five diffraction orders in 24-36b differs from the image formed by just three diffraction orders only in the higher series expansion terms with $n > 3$.

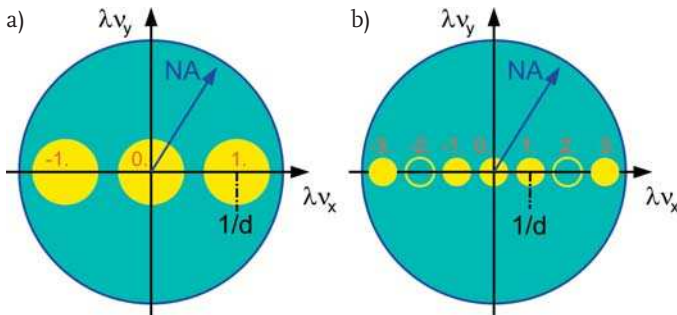


Figure 24-36: Three and five (seven) diffraction orders passing the pupil without truncation.

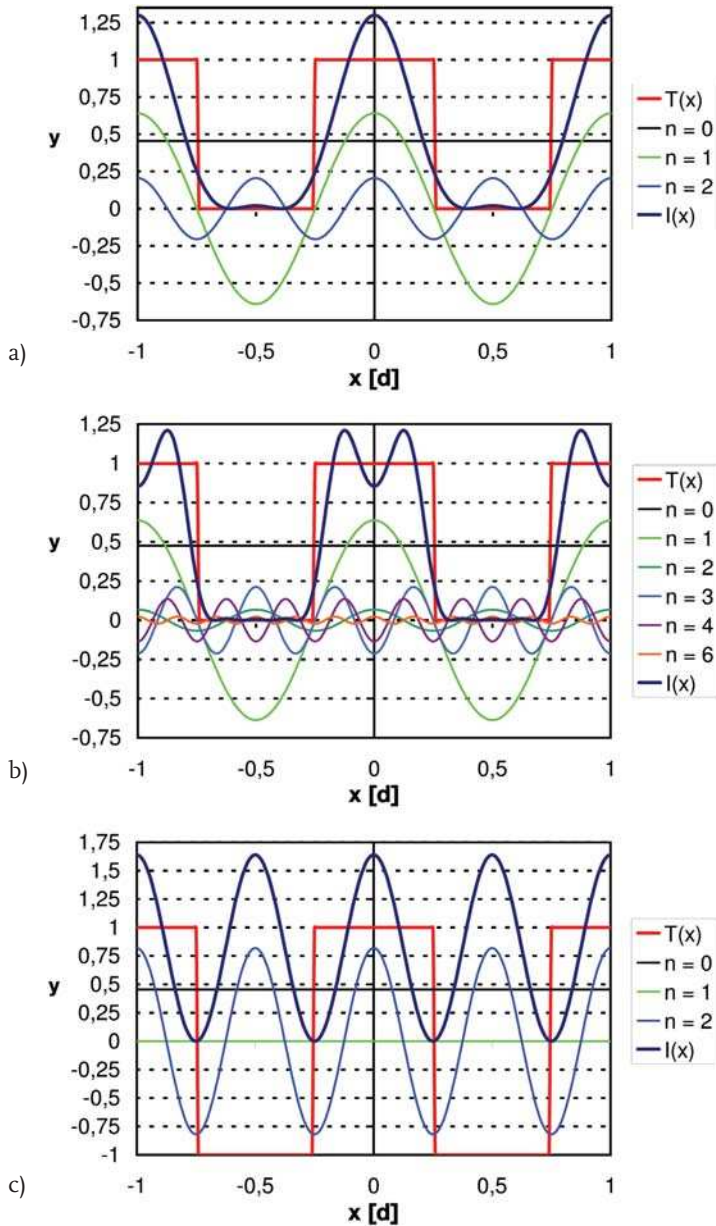


Figure 24-37: Interference patterns of order n forming the image intensity $I(x)$ in comparison with the object transmission for: a) three and b) seven passing diffraction orders; c) phase grating image formed by two diffraction orders.

For a phase grating, the zero-order diffraction efficiency g_0 and thus the series term b_1 vanishes. The coefficients reduce to

$$b_0 = |g_1|^2 \cdot TCC(1, 1) + |g_{-1}|^2 \cdot TCC(-1, -1), \quad (24-37a)$$

$$b_1 = 0, \quad (24-37b)$$

$$b_2 = 2 \cdot \text{Re}[g_1 g_{-1}^* \cdot TCC(1, -1)]. \quad (24-37c)$$

After insertion into eq. (24-35), the image intensity of a phase grating is given by

$$I(x, y) = b_0 + b_2 \cdot \cos\left(\frac{4\pi}{d} \cdot x\right). \quad (24-38)$$

If only first diffraction orders contribute to the image, the phase object still gives an image of high contrast, although the modulation frequency of the image is twice the grating frequency $1/d$ (see figure 24-37c). For symmetrical gratings with phase $\varphi \sim \cos(2\pi/d \cdot x)$, the coefficients in eq. (24-38) can further be simplified by

$$b_2 = b_0 = 2 \cdot |g_1|^2 \cdot TCC(1, 1). \quad (24-39)$$

The image of a symmetrical phase grating thus has a maximum contrast or visibility of $V = 1$ – independent of symmetrical aberrations (like, e.g., spherical aberration or defocus).

24.5.2

The Coherent Grating Image with Aberrations

In general the TCC is given by eq. (22-30)

$$\begin{aligned} TCC(m, n) = & C \iint |u_L(v_x, v_y)|^2 P\left(v_x + \frac{m}{d}, v_y\right) P\left(v_x + \frac{n}{d}, v_y\right) \\ & \cdot e^{\frac{i2\pi}{\lambda} [W(v_x + \frac{m}{d}, v_y) - W(v_x + \frac{n}{d}, v_y)]} dv_x dv_y. \end{aligned} \quad (24-40)$$

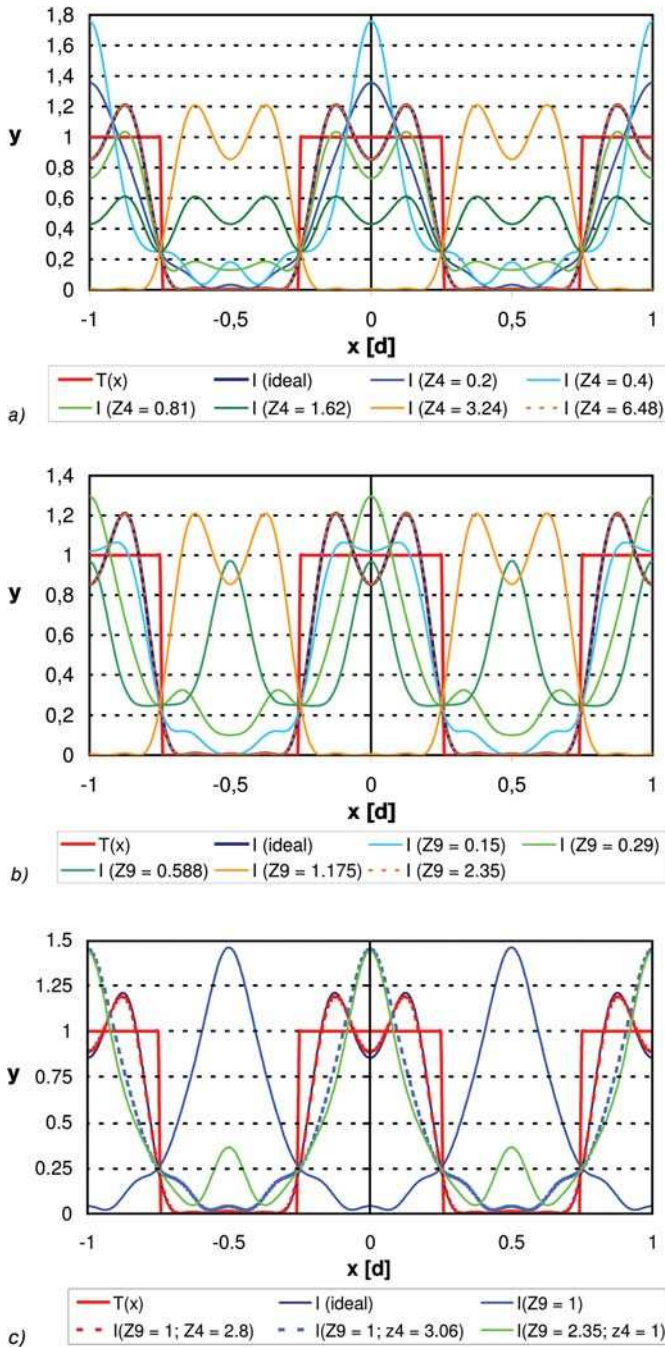


Figure 24-38: a) Influence of defocus Z_4 and b) spherical aberration Z_9 on the grating image with grating period $d = 3.6\lambda/NA$; c) Influence of spherical aberration and defocus on the image of a grating with $d = 3.5\lambda/NA$.

Wave-front errors thus lead in general to complex TCC and cause additional phase modulations in the image intensity. Due to the different phase modulation of individual series contributions which form the image according to eq. (22-27), the intensity distribution of the resulting image period might be asymmetrically deformed. The influence of symmetrical aberrations is illustrated in figure 24-38 at the examples of defocus Z_4 and spherical aberration Z_9 on the image of a Ronchi grating with period $3.6 \lambda/NA$ and $3.5\lambda/NA$, to which consequently five diffraction orders contribute. For ease of computation, the coherent TCC is applied in good approximation for partially coherent image settings with small σ . For a regular Ronchi grating, Talbot periods appear more frequent than they would, according to eq. (17-81), and thus already at $z_T/4$ a repetition of the original grating pattern is reproduced under defocus. For certain object frequencies the grating pattern is reproduced even with spherical aberration (24-38b). In general, the effect of Z_9 on the grating image scales due to a larger defocus (by a factor 3) compared with Z_4 , but the Talbot planes are not found as expected due to the additional aberrations (24.38c).

In figure 24-39, the distortion Z_2 and the asymmetrical deformation of the image period due to coma of lowest order Z_7 is illustrated. Three diffraction orders contribute

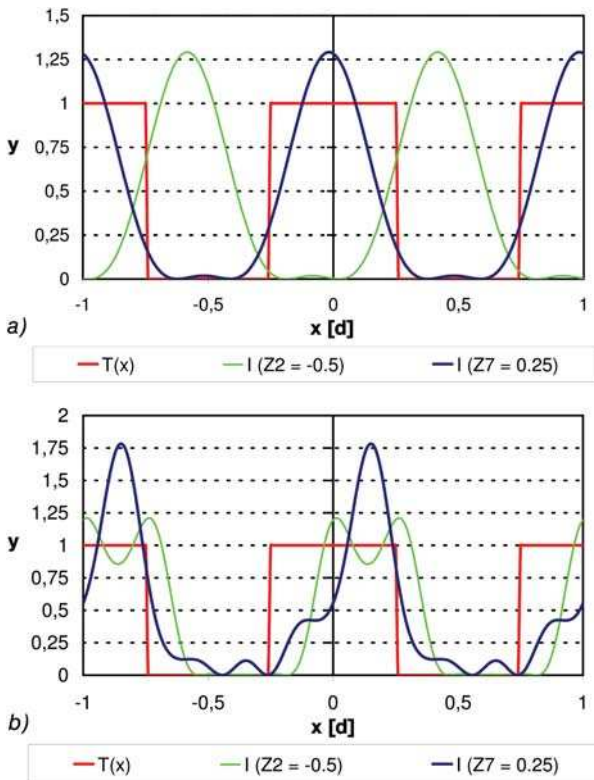


Figure 24-39: 'Real' distortion due to Z_2 in comparison with coma-induced dispersion for the examples of linear grating: a) with $d = 1.2\lambda/NA$; and b) with $d = 3.6\lambda/NA$.

to the grating image with $d = 1.2 \lambda/NA$ and the image intensity thus shows only lateral displacement for both wave-front aberrations Z_2 and Z_7 , but with different lateral displacement depending on the object period. The grating image with $d = 3.6\lambda/NA$ and distortion according to Z_2 is shifted, while for Z_7 the image period is additionally deformed. Due to the different phase modulations for the different series terms, the centre of gravity of the expected period is approximately at $-0.11/d$. Since this effect is observed especially by asymmetric wave-front errors like coma, the effect is called coma-induced distortion. Coma-induced distortion depends on imaging conditions such as illumination, aberration level and object period and orientation.

24.5.3

The Influence of the Coherence Parameter σ on the Grating Image

In this paragraph, the effect on the grating image of partially coherent imaging with different light sources, is discussed. The comparison of the results of partially coherent image simulation is restricted to the iterative solution of eq. (21-40) without consideration of rigorous diffraction effects or polarization. Further, only the aerial image will be given as a simulation result, the effect of detectors will be neglected. A grating consisting of seven periods is taken as the object. All aerial images are given in an image scan through focus, i.e., x - z -scans with lateral coordinate x and axial coordinate z . The x -axis always points to the right, the z -axis points downwards through the focus. For the details of the computation of the through-focus aerial image scan, the reader is referred to chapter 23.

First the image formed by a common circular light source with different sizes σNA with $\sigma = 0.2$, $\sigma = 0.6$ and $\sigma = 1.0$ ($NA = 0.75$) are discussed. Partially coherent imaging with small sizes of the effective light source with, e.g., $\sigma < 0.3$ shows similar behavior to the coherent image. These illumination conditions are thus frequently considered as 'coherent' illumination. The resolution limits for the images and the grating periods which, for higher diffraction orders, participate in the imaging, are compiled in table 24-1. The coherent resolution limit is given for λ/NA at 1.33λ . Figure 24-40 compares simulation results for grating periods beyond the coherent resolution limit, figure 24-41 for periods above the resolution limit.

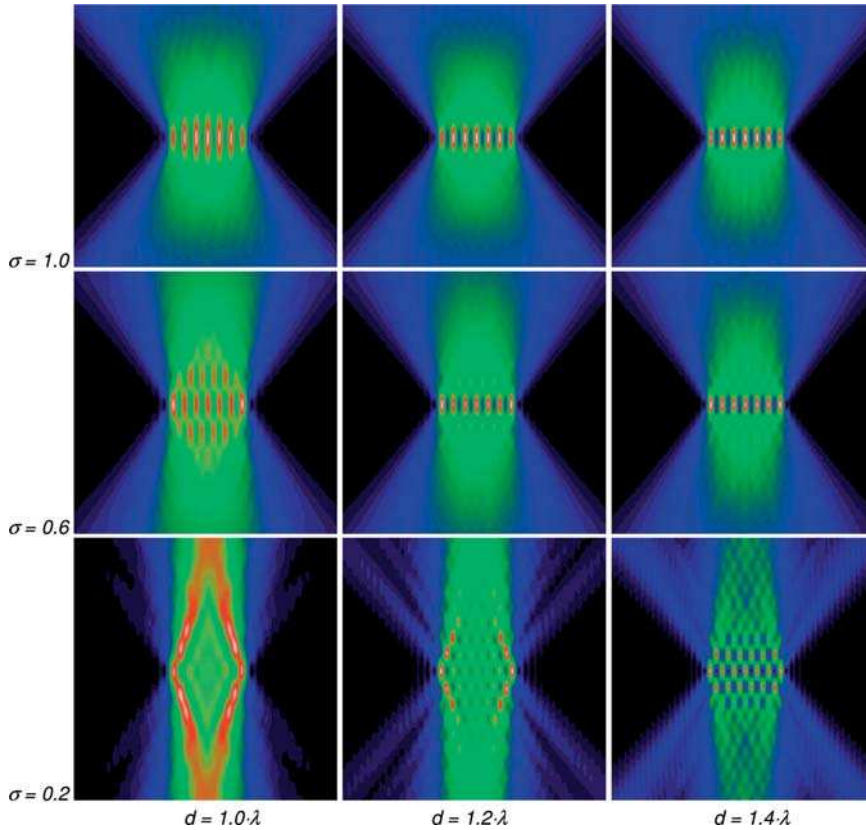


Figure 24-40: Aerial images from a simulation of partially coherent image formation for different grating periods d above the coherent resolution limit with $\sigma=0.2, 0.6$ and 1.0 .

Table 24-1: Resolution limits and grating periods, for which higher diffraction orders participate in the imaging ($NA=0.75$).

σ	1st diffraction order (limit resolution)	2nd diffraction order	3rd diffraction order
Coherent ($\sigma=0$)	1.33λ	2.66λ	4λ
0.2	1.11λ	2.22λ	3.33λ
0.6	0.83λ	1.66λ	2.5λ
1.0	0.66λ	1.33λ	2λ

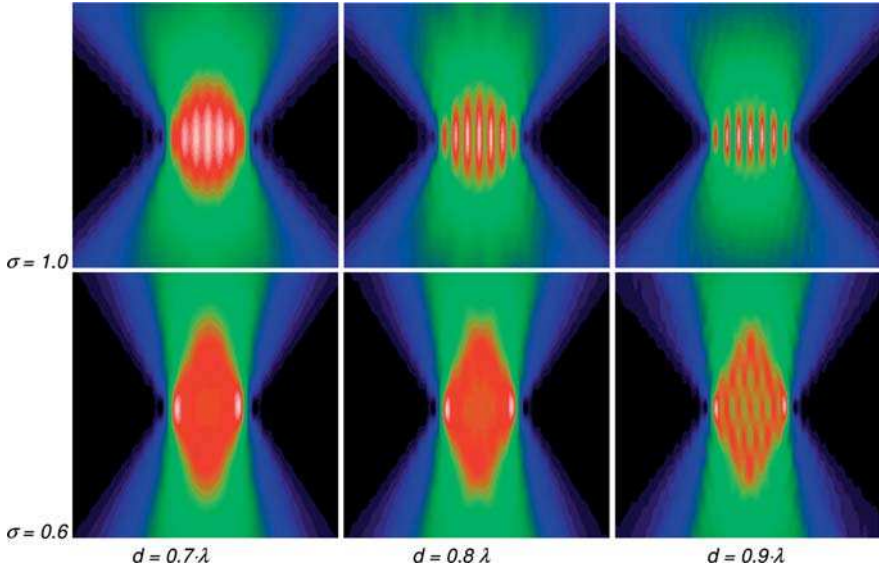


Figure 24-41: Aerial images from a simulation of partially coherent image formation for different grating periods d beyond the coherent resolution limit with $\sigma=0.6$ and $\sigma=1.0$.

For incoherent imaging, the finest object periods can be resolved in the vanishing contrast limit. The incoherent image beyond the coherent resolution limit is formed by symmetrical two-beam interferences whose afocal nature becomes visible for object periods smaller than $d=\lambda$. With increasing object period, the contrast increases while the depth of focus is reduced (figures 24-40 and 24-41, $\sigma=1.0$). For object periods larger than the coherent resolution limit, the image is formed also by three-beam interferences. For the discussed object periods, these as well as the higher diffraction orders, have a minor impact on the incoherent imaging.

The simulation results for partially coherent illumination with $\sigma=0.6$ and object grating periods d below the resolution limit at $d=0.832\cdot\lambda$ only the envelope of the seven periods is imaged. At $\sigma=0.6$ the afocal nature of the two-beam interferences below the coherent resolution limit is less distinctive since the diffraction orders do not pass the pupil symmetrically (see chapter 23).

Under almost coherent imaging with $\sigma=0.2$, the object details are resolved only above larger periods $d=1.11\cdot\lambda$. For object periods above the coherent resolution limit, the aerial images are now characterized by three-beam interferences. Particularly for a smaller illumination divergence the images show periodical repetition of the image grating (though with a reduced number of periods) due to the Talbot effect (figure 24-40, $d=1.2\lambda$ and 1.4λ), while for a large illumination divergence with, e.g., $\sigma=0.6$, the Talbot planes are faded.

Figure 24-42 compares the image contrast or visibility V for the central maximum and the adjacent minima of the aerial images in the ideal focus plane in normalized

frequency units of NA/λ . As explained in section 21.2.3 the image contrast is mainly reduced by that share of the zeroth order which is not contributing to the image formation. For imaging with a common circular light source with $\sigma=1$ the contrast is thus already decreasing for low frequencies of the object grating with increasing grating frequency, while for small σ -settings the contrast theoretically stays at 100% until one of the ± 1 diffraction orders touches the border of the pupil. At the coherent resolution limit, in normalized units at $w = NA/\lambda \cdot d = 1$, the contrast decreases steeply down to the corresponding resolution limit $d = \lambda / NA \cdot (1+\sigma)$.

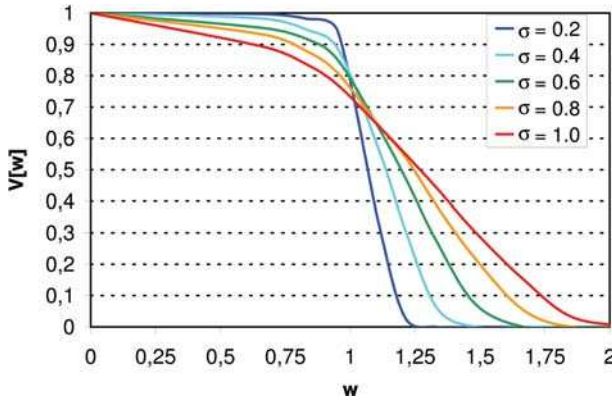


Figure 24-42: Contrast of the central maximum of the aerial images according to the simulation of partially coherent image formation for different light-source dimensions σ .

24.5.4

Influence of the Shape of the Effective Light Source on the Grating Image

The conventional partially coherent and incoherent images have high resolution at the expense of increasing contrast loss. The contrast loss beyond the coherent-resolution limit, however, can be avoided according to Siedentopf, by an appropriate illumination distribution. Using this, the image contrast even for high grating frequencies can theoretically be $V = 100\%$. The size and shape of the effective light source have a significant impact on the image contrast. As examples, the annular illumination, i.e., with the source in the shape of a ring, and the Hopkins dipole illumination are compared with conventional illumination for large and small σ value. The Hopkins dipole illumination is derived from annular illumination using the Hopkins pair of shutters with a structure-dependent shutter position ν_h (figure 21-17). The exemplary light-source shapes under consideration are illustrated in figure 24-43 for comparison.

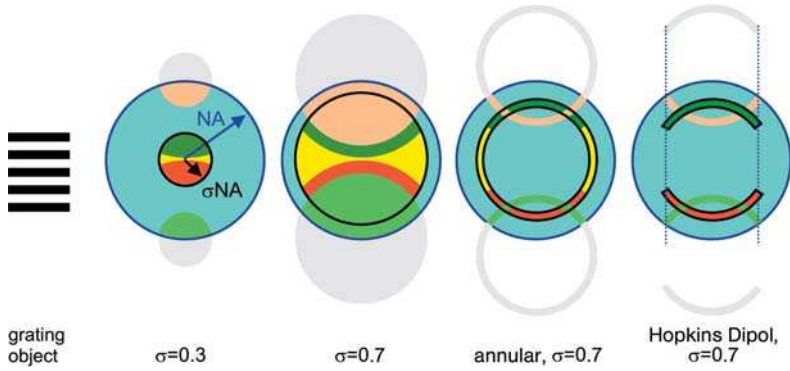


Figure 24-43: Different light-source shapes with two diffraction orders each.

Figures 24-44 to 24-47 compare the simulation results for different shapes of light source. Figure 24-44 shows the image contrast depending on the normalized object frequency w . Again the share of the zeroth order, which is not contributing to image formation or the direct light-source image in the pupil plane, under conventional incoherent illumination with a large light source ($\sigma=0.7$) is already relatively large for large object structures, i.e., for low normalized frequencies w . Under coherent illumination with a small light source, on the other hand, the contrast stays almost ideally at 100% up to frequencies of $w \sim 0.75$, then it decreases steeply and vanishes at the resolution limit $w = (1 + \sigma) = 1.3$. Annular illumination with $\sigma_{\text{mean}}=0.7$ with the inner radius $\sigma_{\text{in}}=0.65$ and the outer radius $\sigma_{\text{out}}=0.75$ behaves approximately like incoherent illumination at smaller object periods, but at higher object frequencies it has a reduced false-light share and thus delivers a higher contrast at frequencies of about $w=1$ and higher. For the structure-adjusted dipole illumination, according to Hopkins, above the normalized frequencies of $w=0.7$ the increased contrast becomes noticeable (below object periods of $1.4 \cdot \lambda$). It stays at more than 80% for frequencies up to $w=1.5$ and decreases, subsequently also vanishing at the

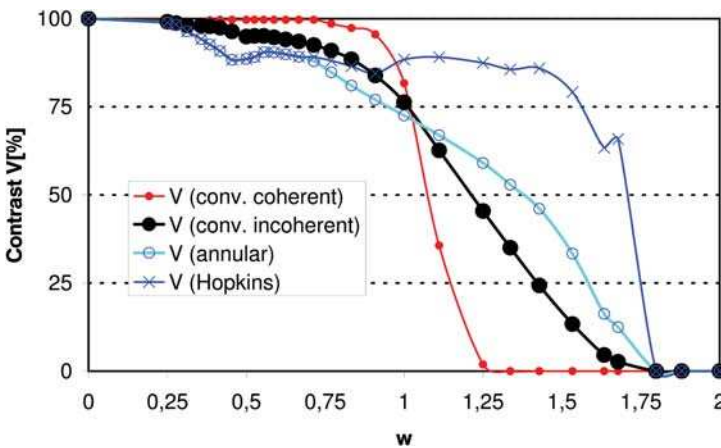


Figure 24-44: Image contrast depending on the object frequency for imaging of a grating structure with 7 periods.

resolution limit at $w = 1.75$. At $w = 0.85$ and $w = 0.425$ the contrast curves with annular illumination and Hopkins' dipole illumination, respectively, show small collapses at those frequencies where the 2nd and the 4th diffraction orders start to pass the pupil. These are quite weak in the example of the amplitude grating with an aspect ratio of the grating width to the groove width of 1:1 due to this structure.

Figure 24-45 shows the contrast distribution at defocus in Rayleigh units $RU = \lambda/(2 \cdot NA^2)$. The depth of focus obviously depends on the object period and the selection of the illumination. For an object period of $d = \lambda$ (dashed lines) Hopkins' illumination aperture only influences the contrast and shifts the contrast distribution basically parallel to the curve of annular illumination. At $d = 2 \cdot \lambda$ the Hopkins pair of apertures no longer has any effect and the contrast is identical to the contrast of annular illumination. With a larger object period, the contrast curve is significantly lower for annular illumination. The question of the depth resolution of the optical image is discussed in more detail in chapter 23, concerning three-dimensional image formation.

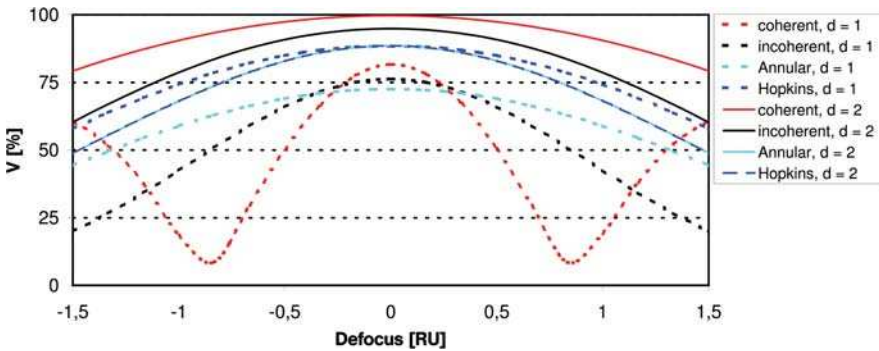


Figure 24-45: Image contrast for grating periods $d=1$ and $d=2$ in units of λ versus the defocusing for different light sources.

Figures 24-46 and 24-47 show the aerial images as results of the partially coherent simulation. It can be clearly seen from the distribution of the interference pattern through the focus when three and more or even just two diffraction orders contribute to the image formation. The small object periods of $d = 0.6\lambda - 0.8\lambda$ can be imaged only with an appropriate dimension of the secondary light source σ . Here a low depth resolution is achieved using the afocal nature of the two-beam interference.

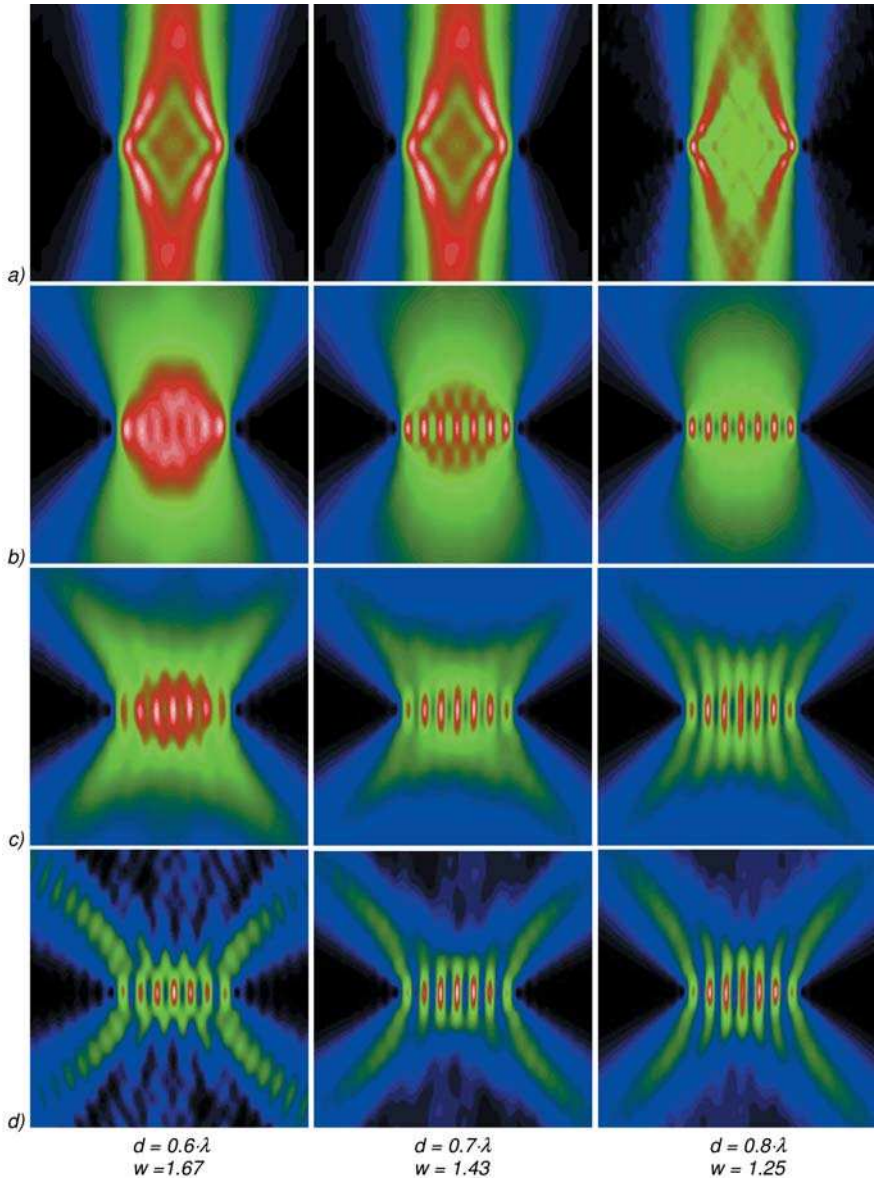


Figure 24-46: Image contrast for grating periods from $d = 0.6\lambda$ up to $d = 0.8\lambda$ for: a) conventional illumination with $\sigma = 0.3$; b) conventional illumination with $\sigma = 0.7$; c) annular illumination; d) Hopkins dipole.

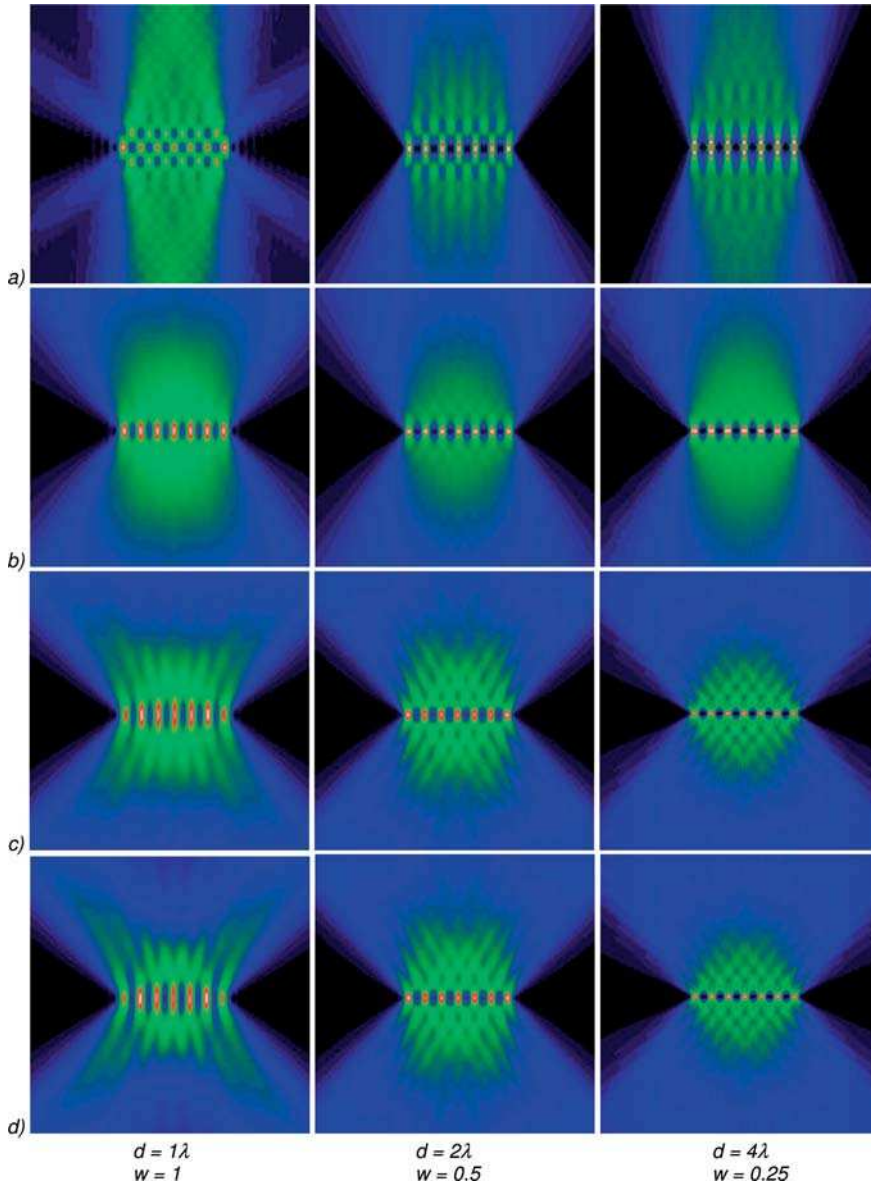


Figure 24-47: Image contrast for grating periods from $d = \lambda$ up to $d = 4\lambda$ for:
 a) conventional illumination with $\sigma = 0.3$; b) conventional illumination with $\sigma = 0.7$;
 c) annular illumination; d) Hopkins dipole.

24.5.5

Wigner Distribution Function for Gratings, Talbot Effect and Propagation-invariant Fields

For coherent illumination from the light-source point ν_q , the scalar field distribution behind an infinitely extended grating is given by

$$U_0(x) = \sum_m g_m e^{i2\pi(\frac{m}{d} + \nu_q)x}. \tag{24-41}$$

Thus the Wigner distribution function of a grating under coherent illumination is given by:

$$\begin{aligned} W(x, \nu) &= \int \sum_{m,n} g_m g_n^* e^{i2\pi[\frac{x}{d}(m-n) + \frac{\Delta x}{2d}(m+n)]} \cdot e^{-2\pi i \Delta x (\nu - \nu_q)} d\Delta x \\ &= \sum_{m,n} g_m g_n^* e^{i2\pi\frac{x}{d}(m-n)} \delta\left(\nu - \nu_q - \frac{m+n}{2d}\right). \end{aligned} \tag{24-42}$$

The diffraction orders become noticeable by discrete contributions of the Wigner distribution function parallel to the x -axis with [24-9]

$$W_k\left(x, \nu = \frac{k}{2d} + \nu_q\right) = \sum_m g_m g_{k-m}^* e^{i2\pi\frac{x}{d}(2m-k)}. \tag{24-43}$$

In table 24-2 the leading spatial periods $2m-k$ for the diffraction orders k are indicated, taking into account the diffraction orders g_m and g_{k-m} of a Ronchi grating. For such a Ronchi grating all even expansion coefficients, except the equal part, will vanish, i.e., $g_{2m} = 0$.

In the zeroth diffraction order for $k=0$, the Wigner distribution function, W_0 , takes on the function of a grating with doubled frequency $2/d$. In the first diffraction order W_1 with $k=1$ there are non-vanishing contributions only for $m=0$ and $m=1$, thus the basic frequency of W_1 is given by $|(2m-k)/d|=1/d$. For $k=2$ the leading basic frequency is for $m=1$ at 0 (constant) and for $m=-1$ at $-4/d$. For the higher diffraction orders, equivalent considerations can be undertaken.

Table 24-2: Leading periods $2m-k$ of the diffraction orders of the Wigner distribution function W .

$2m-k$	$k-m=0$	$k-m=1$	$k-m=2$	$k-m=3$	$k-m=4$	$k-m=5$
$k=0$	0	-2	-4	-6	-8	-10
$k=1$	1	-1	-3	-5	-7	-9
$k=2$	2	0	-2	-4	-6	-8
$k=3$	3	1	-1	-3	-5	-7
$k=4$	4	2	0	-2	-4	-6
$k=5$	5	3	1	-1	-3	-5
$k=6$	6	4	2	0	-2	-4

Figure 24-48 shows the Wigner distribution function for a finite grating with a period of 1.2λ ($NA = 1$) consisting of 11 periods under illumination by a perpendicular planar wave $\nu_q = 0$ and with incoherent illumination in figure 24-49. The structure described above according to eq. (24-43) and table 24-2 can be clearly recognized for coherent illumination. The 11 periods can be recognized in the -1^{st} and 1^{st} diffraction orders from the maxima with the 10 minima arranged in between at $\nu = 1/2d$. The diffraction orders W_k are faded because of the finite grating.

For incoherent illumination, the structure of the Wigner distribution function is faded in the frequency direction due to the superposition of all the W_k 's for all illumination directions ν_q (figure 24-49a). After a low-pass filtering by the imaging, the spectrum is cut back by the aperture and the distribution is blurred in the x -direction due to diffraction effects. The projection along the ν -axis delivers the intensity.

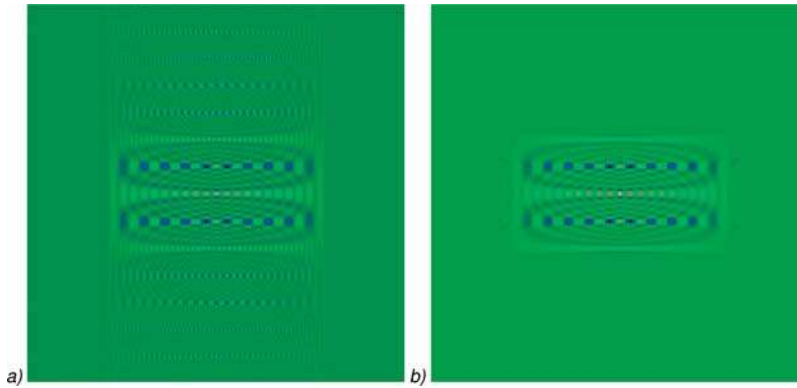


Figure 24-48: Wigner distribution functions for coherent illumination with $\sigma = 0.01$ of a grating structure in: a) the object and b) the image plane.

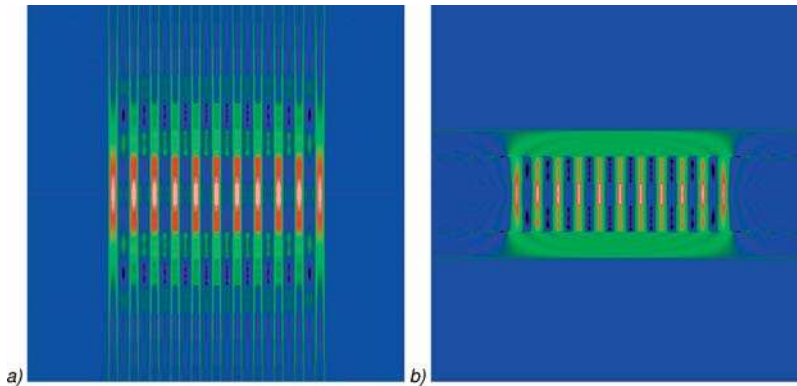


Figure 24-49: Wigner distribution functions for incoherent illumination with $\sigma = 0.01$ of a grating structure in: a) the object and b) the image plane.

For coherent illumination of a grating, additional intensity patterns similar to the original grating are formed in planes defocused to the image plane. This effect, named after Talbot, is shown in figure 24-50 for a grating with 13 periods as object under coherent axial illumination. The example is chosen at the coherent resolution limit with a $NA = 1$ and with a period of $d = \lambda$. In this case only two diffraction orders contribute to the image formation. If the Wigner distribution function of the grating is sheared by free-space propagation, the maxima of the two diffraction orders interfere constructively at certain distances due to projection in the ν -direction, at other distances they interfere destructively (see figure 24-51; showing destructive interference and thus contrast cancellation at $z = 3.5\lambda$ and the Talbot plane at 4λ). Accordingly there is contrast cancellation or secondary images due to the Talbot effect. Because of the symmetry, in addition to the actual Talbot planes, there are additional planes in which an intensity pattern similar to the object with inverse contrast, occurs (figure 24-51).

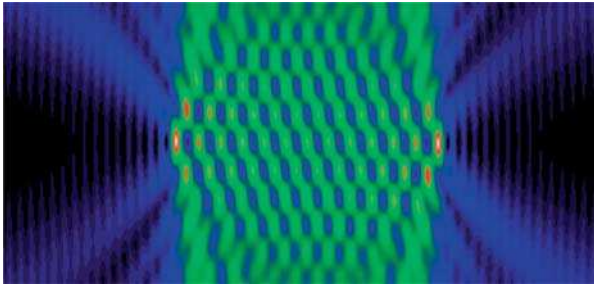


Figure 24-50: Aerial image of a grating for coherent illumination, to illustrate the Talbot effect.

An intensity distribution, which is invariant to propagation, and thus a large depth of focus DOF is achieved if the projection of the Wigner distribution function is invariant to propagation. In figure 24-51 the first diffraction orders for axial illumination are given for frequencies $\nu = 1/2d$. These horizontal patterns are shifted horizontally in the x -direction due to free-space leading to contrast cancellation, inverse contrast and Talbot planes. Only those contributions of the Wigner distribution function at $\nu = 0$, i.e. $W_1(x, 0)$ are invariant to propagation. By oblique illumination of under half the diffraction angle, a diffraction order can be shifted in the frequency direction over $\nu = 0$ so that a propagation-invariant Wigner distribution function W_k is generated. This illumination direction is given according to eq. (24-43) by

$$\nu_q = \pm \frac{k}{2d}. \quad (24-44)$$

Figure 24-53 shows the example above with oblique illumination according to eq. (24-44). Figures 24-54 and 24-55 illustrated the Wigner distribution functions in the image plane and defocused positions. Since the modulation carrying all information about the object is shifted to $\Delta\nu = 0$, free-space propagation has no impact on the image.

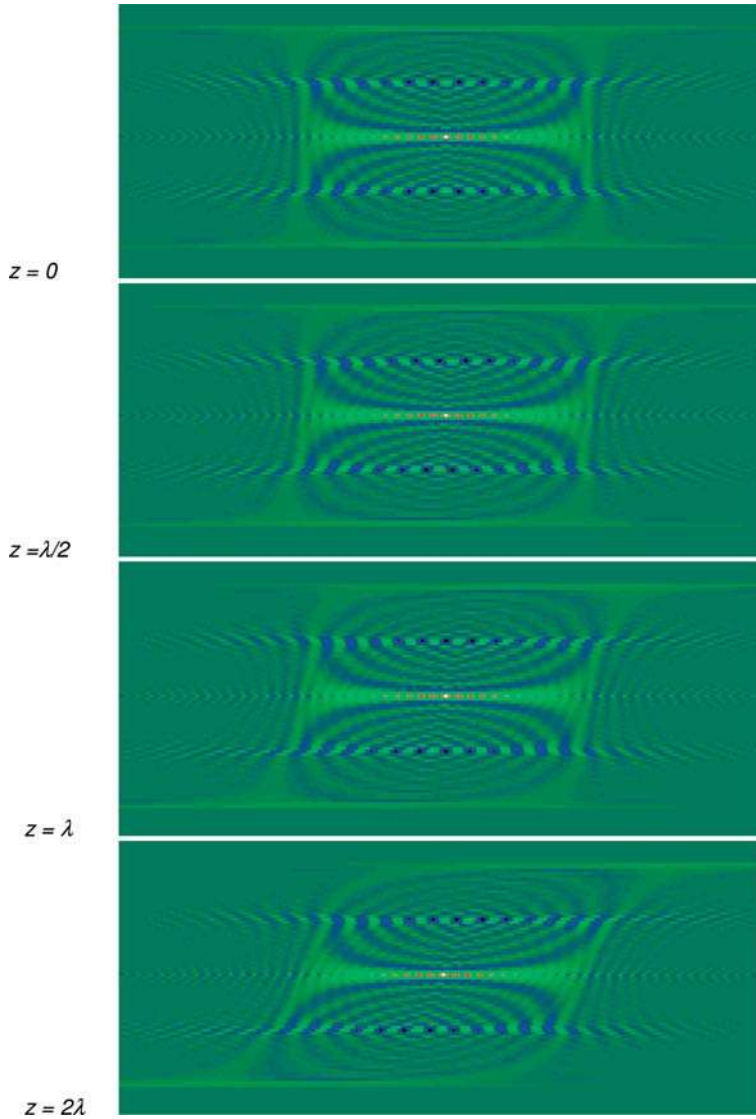


Figure 24-51: Free-space propagation of the Wigner distribution function of a grating for coherent illumination, to illustrate the Talbot effect.

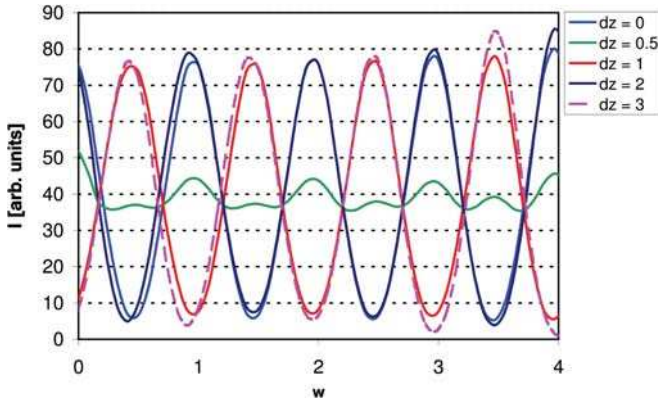


Figure 24-52: Intensity distributions in the Talbot planes using the projection of the Wigner distribution function; defocus dz in units of λ .

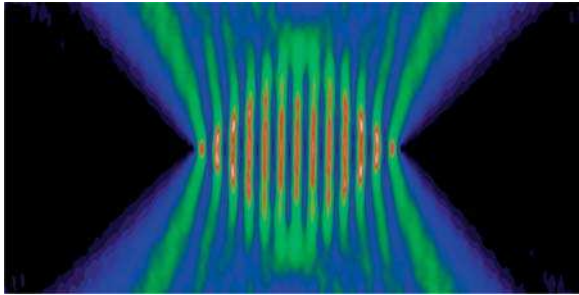


Figure 24-53: Aerial image of a grating at oblique coherent illumination to illustrate the propagation invariance of the Wigner distribution function, according to eq. (24-44).

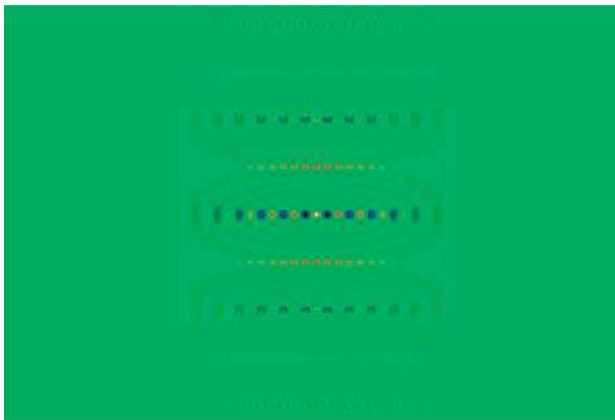


Figure 24-54: Wigner distribution function of a grating with 13 periods for oblique illumination.

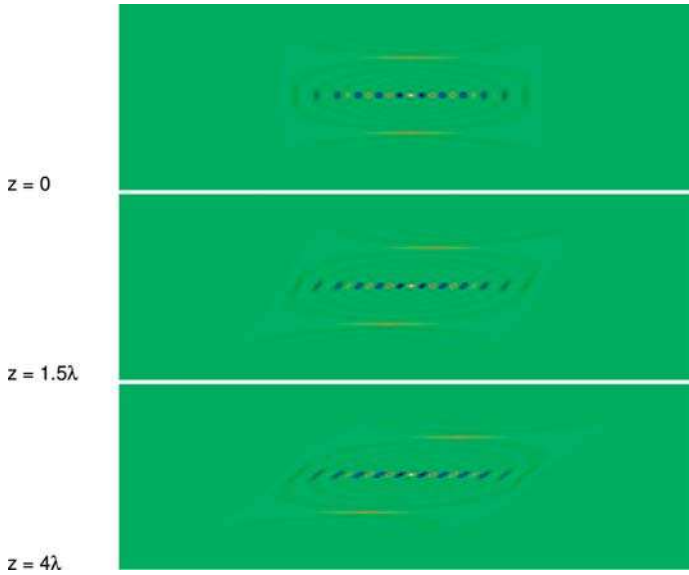


Figure 24-55: Free-space propagation of the Wigner distribution functions of a grating for oblique illumination after low-pass filtering to illustrate the propagation invariance of the Wigner distribution function.

24.6 Pinhole Imaging and Quasi-point Sources

24.6.1 Introduction

In practical setups, a point source is very often required for illumination purposes. Particularly in applications of measurement systems, this kind of illumination source generates an almost ideal wave front. Theoretically, the source can be realized by a very small pinhole, which is much smaller than the wavelength. However, from a practical viewpoint, there are several drawbacks with the use of such small pinholes. These are [24-9], [24-10]:

1. The manufacturing of a circular transmission pinhole of this size is expensive and quite difficult.
2. The measurement of the exact geometry and size is non-trivial, but this is necessary for control reasons.
3. The energy throughput of such a small pinhole is very low, this is a disadvantage for the signal-to-noise ratio. For small pinholes, the transmission scales as the fourth order of the diameter.
4. If the size of the pinhole is smaller than the wavelength, significant polarization effects occur, which may be unwanted. The scalar model for describing the emission of light is no longer applicable.

5. Normally, pinholes of this size are fabricated by thin chromium layers. But for sizes of this amount, the aspect ratio of diameter to depth of the pinhole opening becomes small and the channelling effects in this light waveguide generate an unwanted non-isotropic emission.
6. There are specific effects on the microscopic range depending on the material parameters such as plasmon excitation, resonances, etc.

Therefore, one tries to make the pinhole as large as possible. In this case, from the physical viewpoint, the source is a circular source with finite dimension. In this case, the illumination of the opening also plays a role in the description of the emitted light cone.

In this section, the modelling and limits of this kind of quasi point source are discussed.

24.6.2

Incoherent Image of a Circular Object

The ideal point-spread function of an aberration-free circular pupil with homogeneous illumination is given by the Airy function. If $\sin u$ is the sine of the aperture angle, it obeys the equation

$$I_{\text{PSF}}(r) = \left[\frac{2 \cdot J_1\left(\frac{2\pi}{\lambda} \cdot r \sin u\right)}{\frac{2\pi}{\lambda} \cdot r \sin u} \right]^2. \quad (24-45)$$

If a circular object with radius a is incoherently illuminated, the circle has to be convolved with this distribution to form the image. If the special rotational symmetry is taken into account, one gets the expression

$$I_{\text{pinhole}}(r) = \int_{\rho=0}^a \int_{\theta=0}^{2\pi} \left[\frac{2J_1(\pi\sqrt{r^2 + \rho^2 - 2r\rho \cdot \cos \theta})}{\pi\sqrt{r^2 + \rho^2 - 2r\rho \cdot \cos \theta}} \right]^2 \rho d\rho d\theta. \quad (24-46)$$

Figure 24-56 shows the image, calculated using this equation for different sizes of the pinhole. The radius of the pinhole is normalized in the form

$$d = a \cdot \frac{\sin u}{\lambda} \quad (24-47)$$

which corresponds nearly to the scaling of the radius of the Airy disc. It can be seen from the figure that, for pinhole sizes much smaller than the Airy radius $d \ll 1$, the image closely resembles the Airy distribution. But for radii larger than half of the Airy value, the difference becomes significant. To show this more clearly, figure 24-57 contains the same curves in a logarithmic scale.

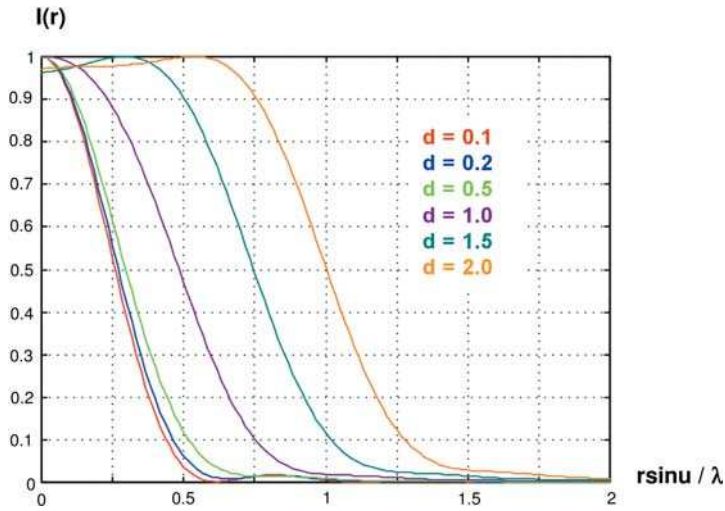


Figure 24-56: Image of a pinhole as a function of the relative size d .

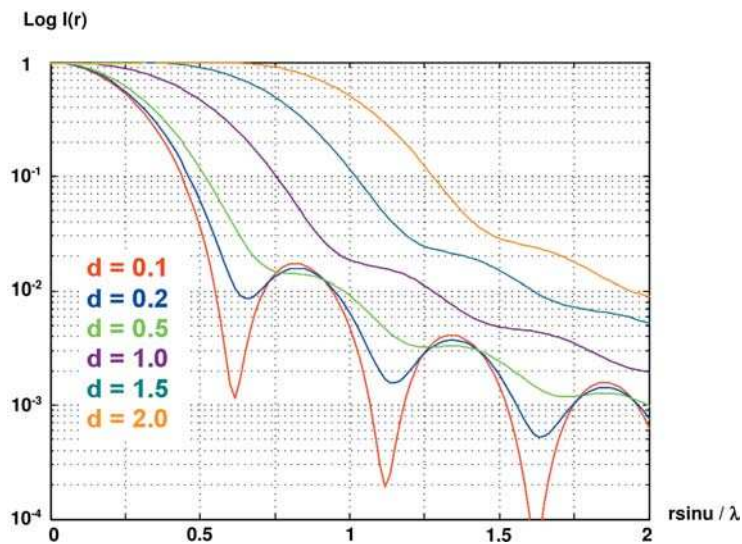


Figure 24-57: Image of a pinhole as a function of the relative size d on a logarithmic scale.

The appearance of the corresponding intensity distributions with the sizes of the circular pinhole are shown in figure 24-57. The vanishing diffraction fine structures with increasing size of the object can be seen clearly.

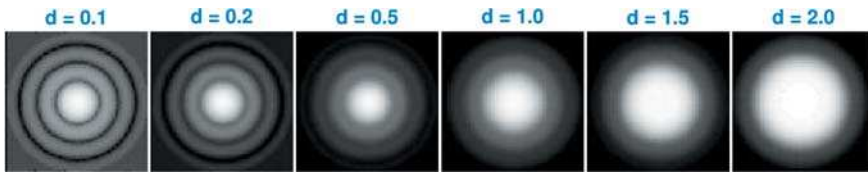


Figure 24-58: Images of a pinhole as a function of the relative size d .

24.6.3

Quasi-point Source

If the diameters of the resulting pinhole image for an intensity level of 50 % and 10 % are compared with the corresponding diameter of an ideal Airy pattern, one gets a measure for the largest pinhole size, which can be regarded as a quasi-point source. Figure 24-59 shows these ratios as a function of the relative size d of the object pinhole.

For diameters of the pinhole in the range

$$a < \frac{1}{8} \cdot r_{\text{airy}} \approx \frac{\lambda}{16 \cdot n \cdot \sin u} \quad (24-48)$$

the images are nearly identical to the ideal pattern and therefore, the point-source requirement is fulfilled to a high degree. If a 10% enlargement of the diameter is accepted, it follows from the numerical calculation as indicated in the figure, in the range

$$a < 0.4 \cdot r_{\text{airy}} \approx \frac{2\lambda}{4 \cdot n \cdot \sin u} \quad (24-49)$$

that the source can be approximately considered as a quasi-point source. In the range

$$a > 0.8 \cdot r_{\text{airy}} \quad (24-50)$$

the diameters are increasing almost linearly with the pinhole size, the image is dominated by the object size and not by the diffraction. Here, a point source assumption is definitely invalid.

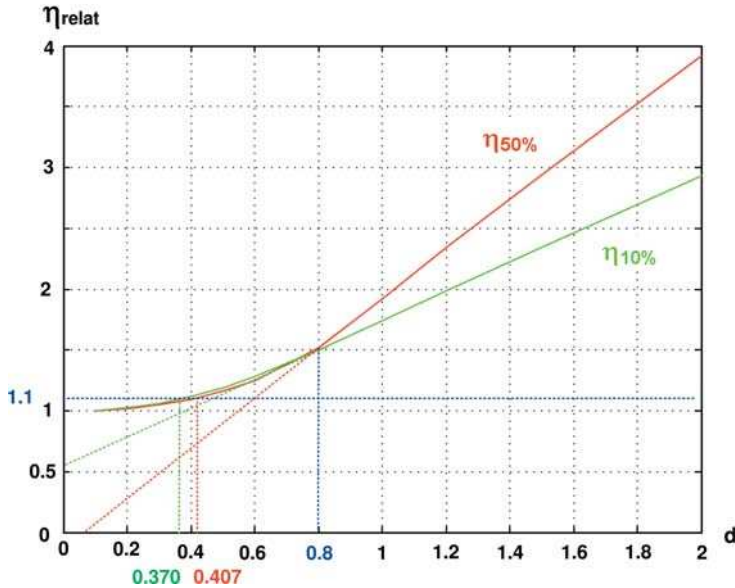


Figure 24-59: Diameter of a pinhole image at 10% and 50% relative to the Airy pattern.

To get an impression for the absolute sizes of the pinholes which are necessary for this purpose, figure 24-60 shows these values as a function of the numerical aperture and the wavelength. As can be seen, for most practical circumstances, the diameter of the pinhole has to be significantly smaller than 1 μm .

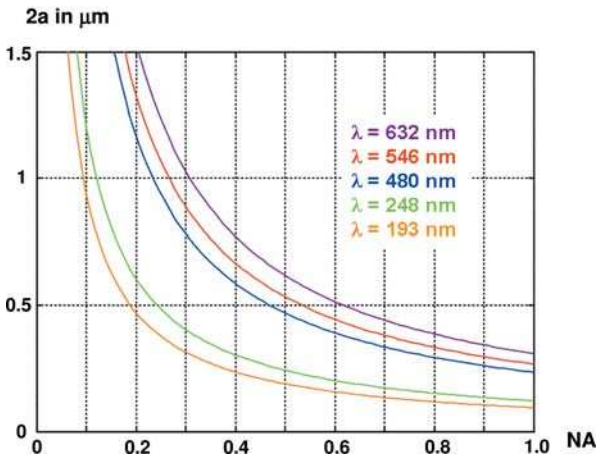


Figure 24-60: Limit of the pinhole size for a quasi-point-source model as a function of the numerical aperture and the wavelength.

24.6.4

Pinhole with Coherent Illumination

If the illumination of the pinhole object is coherent, the image shows more diffraction ring structures, as can be seen in figure 24-61 for different pinhole sizes. But the similarity to the ideal Airy point-spread pattern for small pinholes is sufficient up to a relative diameter of $d=0.8$. Therefore, the assumption of a quasi-point source is valid in a range of almost double the size than it is in the incoherent case. Figure 24-62 shows the section through the images in a logarithmic scale, where the differences can be seen much more easily.

In the case of pinhole imaging with coherent illumination, the pupil is usually not filled. Since the pupil contains the Fourier spectrum of the object, it shows the corresponding diffraction interferences. The greater the pinhole, the more structured is the pupil amplitude. As in the inverse case of a point-spread function for a homogeneous illuminated pupil, the distribution in the pupil is given by the Airy function

$$U(r_p) = \frac{2J_1\left(1.22 \cdot \pi \cdot \frac{a}{r_{\text{airy}}} \cdot \frac{r_p}{f \cdot \sin u}\right)}{1.22 \cdot \pi \cdot \frac{a}{r_{\text{airy}}} \cdot \frac{r_p}{f \cdot \sin u}}. \quad (24-51)$$

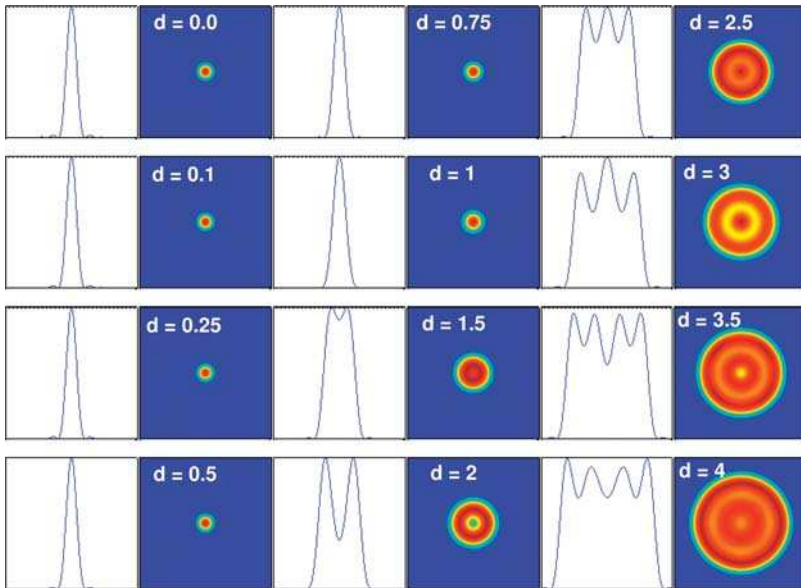


Figure 24-61: Image of a pinhole with different diameters for coherent illumination.

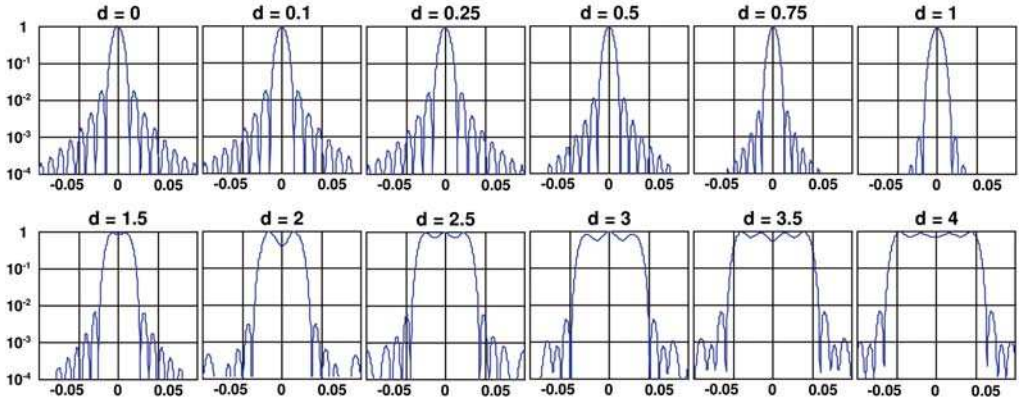


Figure 24-62: Image of a pinhole with different diameters for coherent illumination, section through the image in a logarithmic scale.

If the pinhole reaches the diameter of the Airy disc, the amplitude drops to zero at the rim of the pupil. Figure 24-63 shows the pupil illumination for various pinhole sizes for comparison.

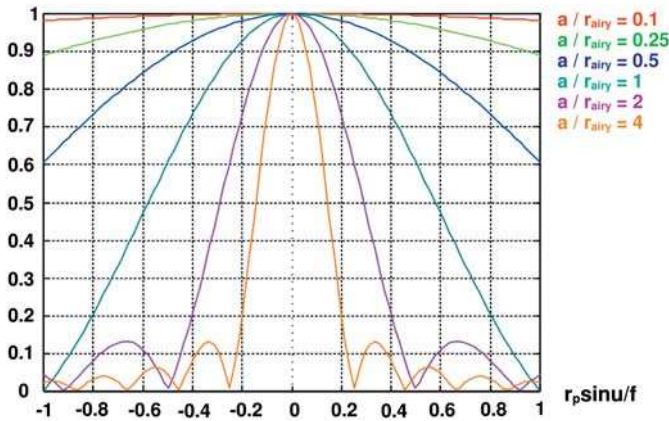


Figure 24-63: Pupil amplitude for the imaging of a pinhole with different diameters and coherent illumination.

24.6.5

Pinhole with Partial Coherent Illumination

If the pinhole is partially coherently illuminated with a coherence parameter σ , there is a transition between the two limiting cases discussed above. Especially in the case of an imaging setup with a high numerical aperture, it is almost impossible to generate a fully incoherent illumination. Figure 24-64 illustrates the image development with different coherence factors σ for three pinhole sizes.

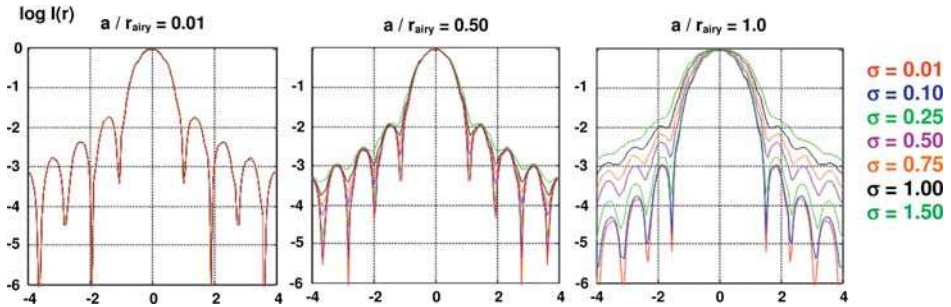


Figure 24-64: Image of a pinhole with different sizes a/r_{airy} and a partial coherent illumination with parameter σ .

In figure 24-65, a comparison between a pinhole image with partially coherent illumination and the coherent case is shown in a more quantitative manner. As can be seen, the rms-value of the compared images grows with the size of the pinhole and the σ -factor as expected. Below a coherence factor of approximately $\sigma=0.25$, the influence of the coherence is very low and in the most practical cases negligible.

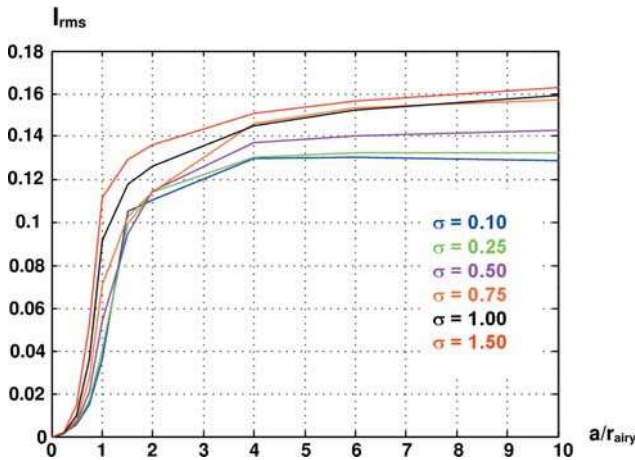


Figure 24-65: Rms value of the difference between the images of a coherent and a partially coherent illuminated pinhole as a function of the pinhole size and different coherence factors σ .

24.6.6

Defocusing Planes and Deconvolution

In the above discussion the pinhole image is always regarded directly in the image plane. As can be seen from figure 24-66, the effect of the finite pinhole size can be recognized much more easily in the defocusing planes of an image stack with variable z -values.

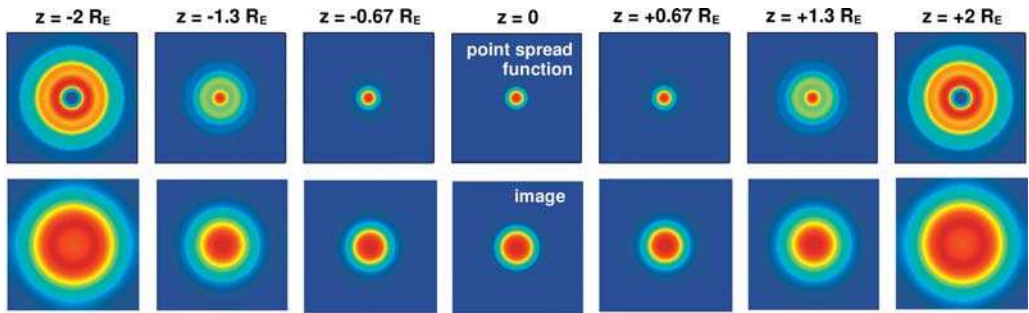


Figure 24-66: Ideal point-spread function and image stack for a pinhole with relative size $d = 1$ in different defocusing locations.

This observation leads to the concept of the pinhole deconvolution. If the effects of the finite pinhole size have to be removed for special applications, it is possible to retrieve the point-spread function out of the image with the help of the redundant information in several defocusing planes. As can be seen in the figure, the fine structure of the diffraction rings manifests more significantly far from the image plane.

24.7

Literature

- 24-1** Ernst Abbe, *Die Lehre von der Bildentstehung im Mikroskop*, edited by O. Lummer and F. Reiche, Vieweg, Braunschweig (Germany) (1910).
- 24-2** K. Michel, *Die Grundzüge der Theorie des Mikroskops* (Wissenschaftliche Verlagsgesellschaft M.B.H., Stuttgart, 1981).
- 24-3** M. D. Levenson et al., Improving resolution in photolithography with a phase shifting mask, *IEEE Transactions on Electron devices* **ED-29**, 1828–1836 (1982).
- 24-4** F.M. Schellenberg, *Resolution Enhancement Techniques in Optical Lithography*, SPIE Milestone Series Vol. MS 178 (2004).
- 24-5** A. Glindemann, New approximation for partially coherent imaging of straight edges, *J. Mod. Opt.* **36**, 659–668 (1989).
- 24-6** A. Glindemann and J. Kross, Symmetry in partially coherent imaging semi-transparent edges, *J. Mod. Opt.* **38**, 379–394 (1991).
- 24-7** C.J.R. Sheppard, A. Choudhury, Image formation in the scanning microscope, *Optica Acta* **24**, 1051–1073 (1977).
- 24-8** V. N. Mahajan, *Optical Imaging and Aberrations, Part II*, SPIE Press, Bellingham (2001).
- 24-9** M. Testorf and J. Ojeda Castaneda, Fractional Talbot Effect: analysis in phase space, *J. Opt. Soc. Am. A* **13**, 119–125 (1996).
- 24-9** F. J. Garcia de Abajo, *Opt. Express* **10**, 1475 (2002), Light Transmission through a single cylindrical hole in a metallic film.
- 24-10** N. Bonod, E. Popov and M. Neviere, *Opt. Commun.* **245**, 355 (2005), Light transmission through a subwavelength microstructure aperture: electromagnetic theory and applications.
- 24-11** A. Degiron, H.J. Lezec, N. Yamamoto and T. W. Ebbesen, *Opt. Commun.* **239**, 61 (2004), Optical transmission properties of a single subwavelength aperture in a real metal.
- 24-12** C. Obermüller and K. Karrai, *Appl. Phys. Lett.* **67**, 3408 (1995), Far field characterization of diffracting circular apertures.

25

Special System Examples and Applications

- 25.1 Introduction 410
- 25.2 Point-spread Functions for Annular Pupils 410
 - 25.2.1 Introduction 410
 - 25.2.2 Annular Pupils, Central Obscuration and Pupil Filters 411
- 25.3 Point-spread Functions of Non-uniform Illuminated Pupils 416
 - 25.3.1 Introduction 416
 - 25.3.2 General Gaussian Apodization 417
 - 25.3.3 Gaussian Profile with Truncation 418
- 25.4 Engineering of the Point-spread Function by Pupil Masks 423
 - 25.4.1 Introduction 423
 - 25.4.2 Characterization of the Three-dimensional Point-spread Function 423
 - 25.4.3 Characterization of Extended Depth of Focus 426
 - 25.4.4 Relation Between Axial and Transverse Resolution 427
 - 25.4.5 Ambiguity Function as Defocussed Transfer Function 429
 - 25.4.6 Image Multiplexing 430
 - 25.4.7 Fundamental Relationships 432
 - 25.4.8 Calculation of Masks 432
- 25.5 Special Pupil Masks 433
 - 25.5.1 Introduction 433
 - 25.5.2 Phase Masks According to Toraldo 434
 - 25.5.3 Logarithmic Phase Mask 435
 - 25.5.4 Chirped Ring Pupil 437
 - 25.5.5 Complex Filter Described by Zernike Expansions 439
 - 25.5.6 Cubic Phase Plates for Extended Depth of Focus 442
 - 25.5.7 Structured Illumination 447
- 25.6 Selected Practical Applications for Pupil Filtering Techniques 450
 - 25.6.1 Phase Contrast Filtering, Dark-field Illumination 450
 - 25.6.2 Frequency Doubling 453
 - 25.6.3 Defect Filtering 455
 - 25.6.4 Ronchi Test 456
- 25.7 Literature 463

25.1

Introduction

In this section, some examples and special applications are described, which illustrate the use of the theoretical results from the preceding chapters and which apply them to practical issues. Here, the effect of a modified pupil function on the generation of the point-spread function is one of the major topics. If only ring-shaped pupil forms are present, the point-spread function changes very significantly in the axial and transverse directions. Another case, which occurs very often in practice, is pupil apodization and inhomogeneous illumination. In particular, if laser sources are used for illumination, Gaussian pupil filling can be observed. In this case, the point-spread function is also modified. Here, the rim of the pupil is not the only parameter on which the performance of the system depends, the width of the illuminated area is also important.

Special types of pupil mask with the appropriate phase and transmission distribution can be used to modify the point-spread function in the desired way. These methods and some well known mask types are discussed in this chapter. To obtain the full benefit of this technique, a digital detection of the image combined with a image processing in the form of a deconvolution can be used. This is also treated briefly in this chapter.

Some practical methods of using these pupil-filtering techniques for measurement purposes, defect recognition, or phase imaging, are finally discussed.

25.2

Point-spread Functions for Annular Pupils

25.2.1

Introduction

The point-spread function considered so far has been restricted to conventional imaging optical systems with homogeneous transmission of the aperture. Real systems such as telescopes or mirror systems often have annular pupils. In addition to this, optical systems in general suffer from variations of the optical transmission over the pupil. The influence of inhomogeneous transmission on the point-spread function and the imaging characteristic will be outlined in this and section 25.3. As will be shown, with special annular pupil filters, the central maximum of the point-spread function can be reduced at the expense of larger secondary maxima. In special applications, inhomogeneous transmission of optical systems is thus desired. On the other hand, variable transmission of optical systems may be employed to reduce these diffraction rings of the Airy pattern. The latter, called apodization filters, are treated in section 25.3. The first effect, annular pupils and pupil filtering, is the content of this section. The special topic of enhancement of depth of focus by annular pupil filters is treated as an example of image multiplexing.

25.2.2

Annular Pupils, Central Obscuration and Pupil Filters

With the numerical aperture of the obscuration given by $\epsilon \cdot NA$ with $\epsilon = NA_{obs}/NA < 1$, in general the amplitude of an ideal image point in the image plane of an imaging system with annular pupil is obtained by

$$U(r) = 2\pi \int_{\rho=\epsilon}^1 J_0(2\pi r\rho)\rho d\rho = \frac{NA}{r \cdot \lambda} \left[J_1\left(2\pi r \cdot \frac{NA}{\lambda}\right) - \epsilon \cdot J_1\left(2\pi r \cdot \frac{\epsilon \cdot NA}{\lambda}\right) \right] \quad (25-1)$$

where $J_0(2\pi r\rho)$ and $J_1(x)$ are the Bessel functions. The radius r_n of the n^{th} dark minima of the resulting intensity pattern is determined according to

$$J_1\left(2\pi r_n \cdot \frac{NA}{\lambda}\right) = \epsilon \cdot J_1\left(2\pi r_n \cdot \frac{\epsilon \cdot NA}{\lambda}\right). \quad (25-2)$$

The point-spread function of an annular pupil in the limit $\epsilon \rightarrow 1$ is given by the Bessel function of 0. order $|J_0|^2$. For $\epsilon=0.5$, the first minimum is at $r_0=0.5\lambda/NA$, thus at a smaller radius as in the Airy pattern, while the next minima have larger radii. Table 25-1 compares radii and encircled energy for the first rings of the intensity pattern with $\epsilon = 0.5$.

Table 25-1: Radii and encircled energy of first rings of the Airy intensity pattern in comparison with central obscuration with $\epsilon = 0.5$.

	Central spot	first ring	second ring
r_n (Airy)	$r_0 = 0.611 \lambda/NA$	$r_1 = 1.1165 \lambda/NA$	$r_2 = 1.62 \lambda/NA$
Encircled energy (Airy)	83.8%	7.2%	2.8%
r_n ($\epsilon = 0.5$)	$r_0 = 0.5 \lambda/NA$	$r_1 = 1.14 \lambda/NA$	$r_2 = 1.75 \lambda/NA$
Encircled energy ($\epsilon = 0.5$)	47.9%	33.6%	7.3%

The normalized point-spread functions for a conventional and an ideal ring-shaped pupil with $\epsilon \rightarrow 1$ are compared in figure 25-1, with the Airy disk (25-1a), and the Bessel function $|J_0|^2$ (figure 25-1b). As can be seen, the central maximum is smaller for the Bessel Beam at the expense of the secondary maxima, which are much more distinctive than for the Airy disc. The resulting normalized point-spread functions for different obscuration parameters ϵ are compared in figure 25-2. The height of the ring intensities increases with increasing obscuration ϵ , while the FWHM diameter of the central maximum decreases slightly with obscuration parameter ϵ (figure 25-2a). The difference becomes more visible when comparing the encircled energy function, as illustrated in figure 25-2b, where the shift of the energy from the central maximum into the outer diffraction rings is observed.

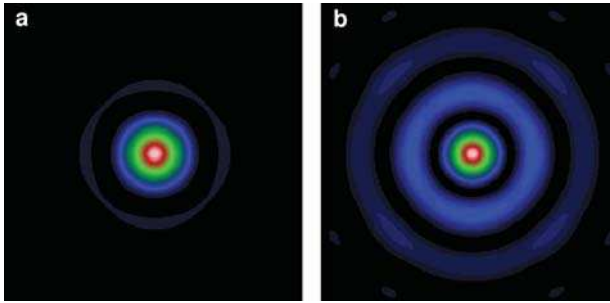


Figure 25-1: PSF for circular and annular pupils in comparison.
 a) Intensity pattern of the Airy disk.
 b) Intensity pattern of a circular ring pupil (Bessel function).

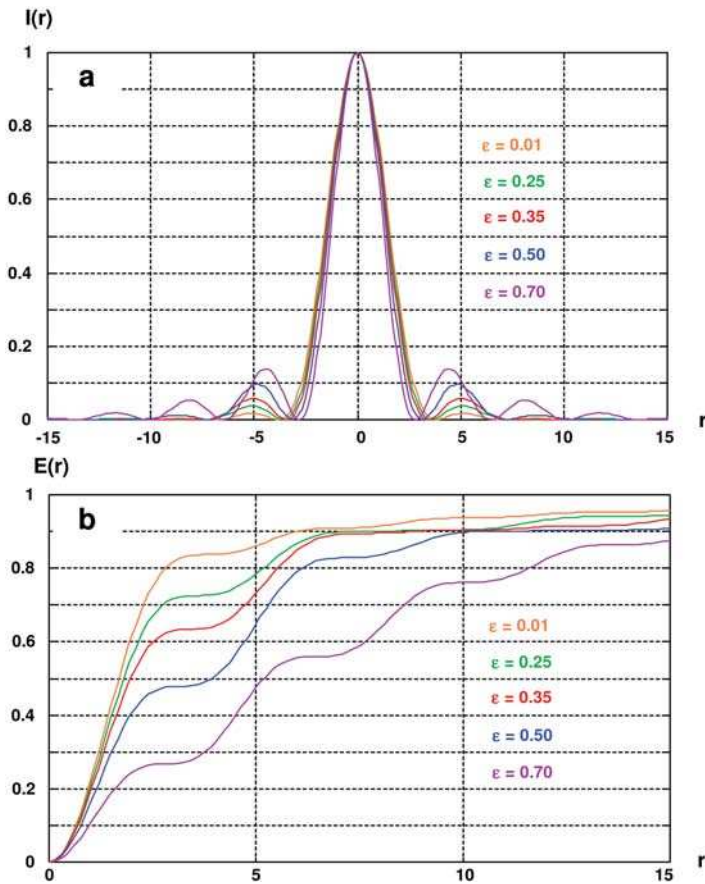


Figure 25-2: a) Intensity distribution $I(r)$ of the point-spread function of an annular pupil with different obscuration ratios ϵ ; b) encircled energy function of the different PSFs.

The effect of an annular pupil on the 3D point-spread function can be obtained by considering the generalized aperture (figure 25-3). The point-spread function in both the radial and longitudinal direction is analogously given by the Fourier transform of the projections of the three-dimensional transfer function. With $NA_{in} = \varepsilon \cdot NA_{out}$ it follows that

$$I_{PSF}(r, z = 0) = \left\{ \frac{1}{(1 - \varepsilon^2)\pi \frac{NA_{out}}{\lambda} r} \left[J_1 \left(2\pi \frac{NA_{out}}{\lambda} r \right) - \varepsilon \cdot J_1 \left(2\pi \frac{NA_{in}}{\lambda} r \right) \right] \right\}^2, \quad (25-3)$$

$$I(r = 0, z) = \left| \text{sinc} \left[\frac{(n \cos \alpha_{in} - n \cos \alpha_{out}) \cdot z}{\lambda} \right] \right|^2. \quad (25-4)$$

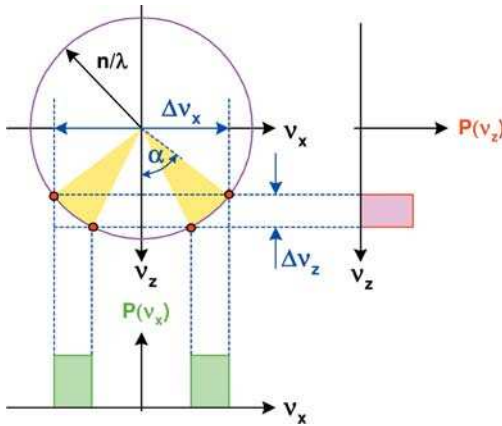


Figure 25-3: Generalized pupil $P(v_x, v_z)$ with central obscuration.

From the generalized aperture it follows that, for a lens systems with central obscuration or with an annular pupil, the maximum lateral resolution according to Abbe remains unchanged in the first order compared with a lens without central obscuration, but the depth of focus increases over the reduced v_z -extension of the generalized pupil. Therefore, the depth resolution is reduced in imaging with annular apertures, leading on the other hand to a larger depth of focus [25-1]. As illustrated in figure 25-4 for the example with $NA_{in} = 0.5$, the maximum depth of focus is obtained for apertures formed by small annular rings, while the point-spread function consists of many bright rings encircling the central maximum and thus having a diminishing effect on the image contrast [25-2]. The effect of “annular” pupils becomes especially visible for one-dimensional systems (cylindrical systems) with two off-axis segments of the “linear” pupil. For high obscuration, in the example with $NA_{in} = 0.4$, the image amplitude and intensity is, to a good approximation, given by the interference pattern produced by two plane waves (figure 25-5, last column).

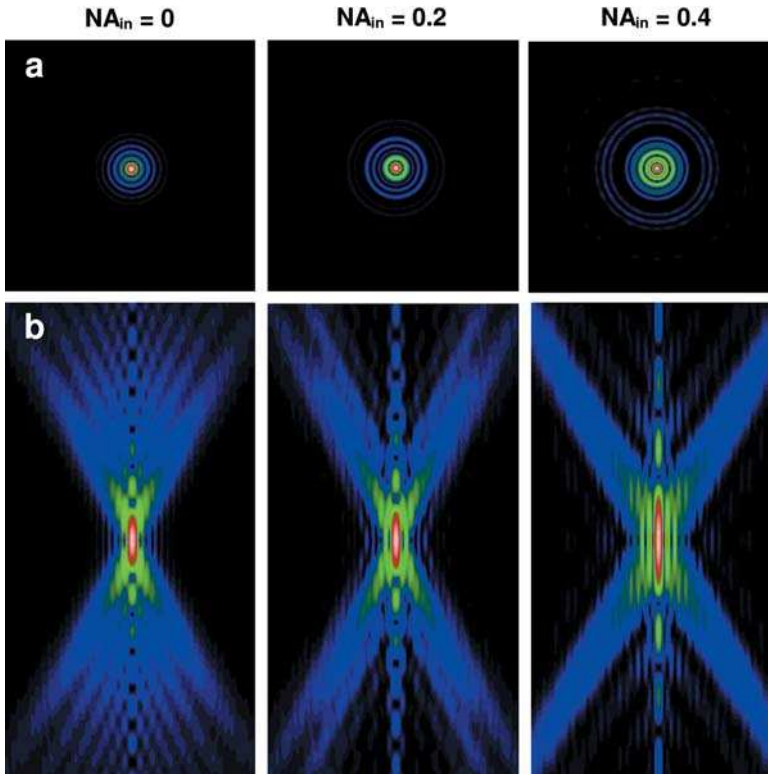


Figure 25-4: Intensity scans through three-dimensional point images of systems with central obscuration; $NA_{out} = 0.6$ with central obscuration of $NA_{in} = 0, 0.2$ and 0.4 , a) in x and y , b) in x and z (logarithmic scale).

The decreasing diameter of the central maximum of the point-spread function for annular pupils, stimulated the idea of optical super-resolution by annular pupils and also of pupil filtering. Annular apertures, or generally, apertures consisting of several annular rings, have been proposed in order to achieve super-resolution compared with the Rayleigh resolution [25-3], [25-4], [25-5], [25-6]. Since systems with annular pupils or transmission pupil filters have a limited transmission, it was also proposed to apply annular phase filters. With a set of binary phase filters of phase φ_n between radius ρ_n and ρ_{n-1} the amplitude of the point-spread function is given by

$$U(r) = \sum_n \frac{e^{i\varphi_n}}{r \cdot \lambda} \left[\rho_n \cdot J_1 \left(2\pi r \cdot \frac{\rho_n}{\lambda} \right) - \rho_{n-1} \cdot J_1 \left(2\pi r \cdot \frac{\rho_{n-1}}{\lambda} \right) \right]. \quad (25-5)$$

To summarize, various shapes of pupil filters have been investigated, with greater or lesser success. In every case, the width of the central maximum is reduced at the expense of the side orders of the point-spread function, which has very disturbing effects on the optical imaging of extended objects since the visibility is reduced by a

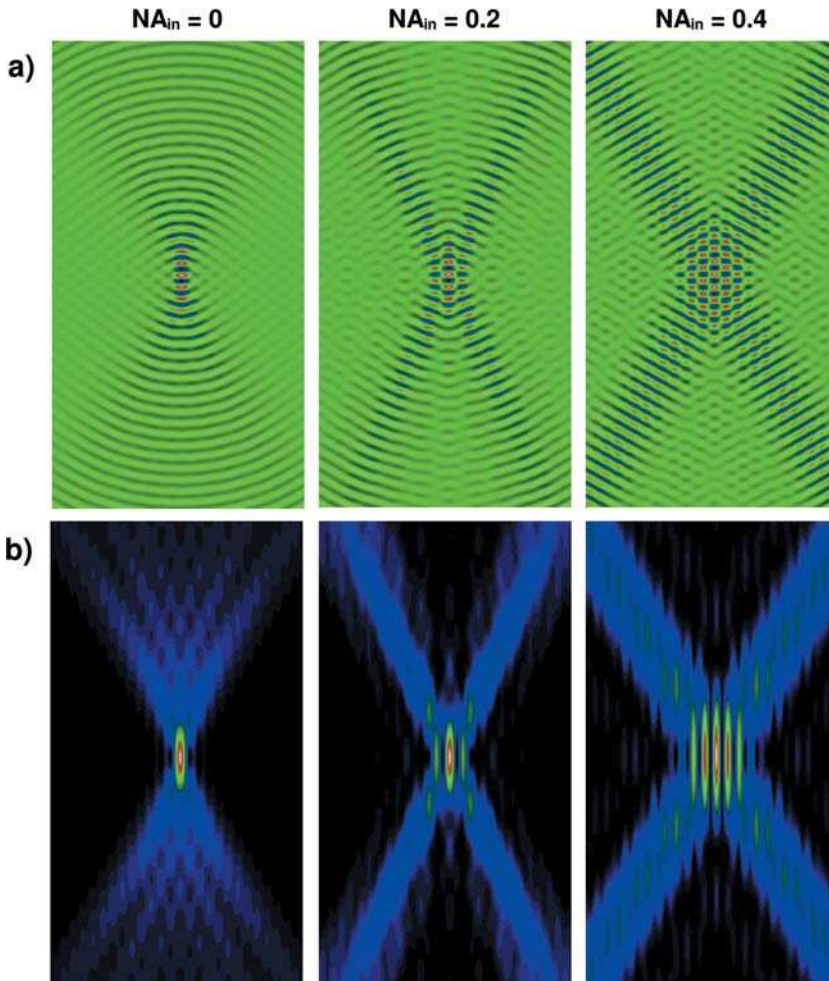


Figure 25-5: a) Amplitude and b) intensity scans through two-dimensional point images of cylindrical (1D) systems with central obscuration; $NA_{out} = 0.6$, central obscuration of $NA_{in} = 0, 0.2$, and 0.4 .

large background share due to the secondary maxima. The special shaping of the point image is mainly useful for scanning imaging methods, e.g., scanning microscopes or CD and DVD pickups [25-7], [25-8]. Due to the drawbacks, especially lower contrast, so far pupil filtering techniques only play a minor role within general optical imaging and the application of pupil filters is restricted only to special applications.

25.3

Point-spread Functions of Non-uniform Illuminated Pupils

25.3.1

Introduction

The point-spread functions of conventional or annular pupils suffer from large secondary maxima, which may have a diminishing effect on the optical image. On the other hand, it is possible to suppress the secondary maxima by an approximately Gaussian shape of the pupil function. The suppression of the secondary maxima is called apodization (from the Greek; it means 'removal of the feet' [25-9]). In figure 25-6, for comparison, the ideal case of a completely Gaussian-shaped pupil function is shown.

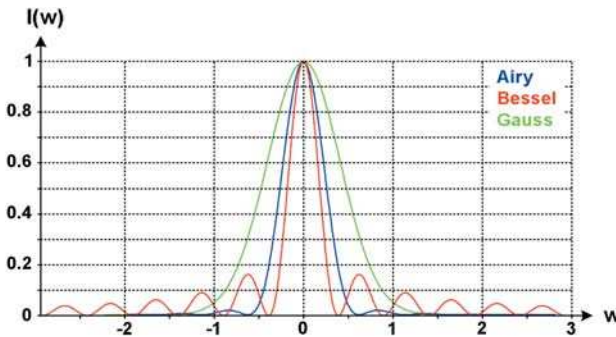


Figure 25-6: Comparison of the cross-sections of the Airy point image to the point image of a ring pupil with $\varepsilon \ll 1$ (Bessel function) and to a Gaussian-shaped pupil.

With the Gaussian apodization, the maximum encircled energy of 100% is reached at a smaller radius of the point-spread function compared with the Airy diffraction pattern, while with the latter, more energy is concentrated in the centre of the point image. The Bessel function of an annular pupil showing, in contrast, a smaller width of the central maximum (e.g., determined by the FWHM), distributes the energy over a larger area.

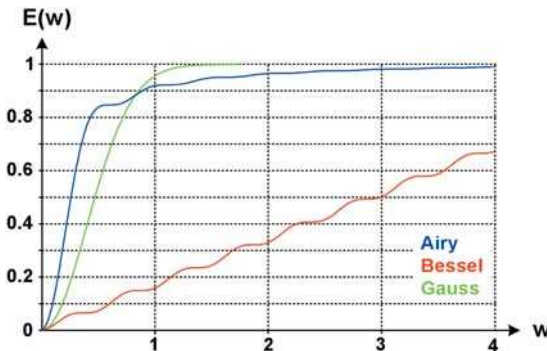


Figure 25-7: Encircled energy for different point-spread functions according to the profiles of figure 25-6.

In practice, the illumination of coherent optical systems is often realized with lasers operating in a fundamental mode, given by a Gaussian intensity distribution. Gaussian shapes of the point-spread functions are also obtained by inhomogeneous illumination with Gaussian beams. Particularly in scanning con-focal microscopy, when the object is scanned by focussed beams, Gaussian beam shapes may be applied instead of plane waves with corresponding Airy point-spread functions. Both, apodization pupil filters and inhomogeneous illumination by Gaussian beams are considered in this chapter.

25.3.2

General Gaussian Apodization

Apodization in general is expressed by inhomogeneous pupil functions of optical systems. Frequently so called super-Gaussian profile functions are applied. This function is given for rotational symmetry by the equation

$$P(r) = P_0 e^{-2(\frac{r}{a})^m} \tag{25-6}$$

The profile is determined by two parameters, given by radius w describing the lateral extent and the exponent parameter m , indicating the steepness of the gradient at the rim of the pupil. The special case $m=2$, results in the conventional Gauss profile. With increasing m , the distribution increasingly takes the shape of a top hat profile with a very large slope at the boundary. Figure 25-8 compares several super Gaussian profile functions for different values of the parameter m .

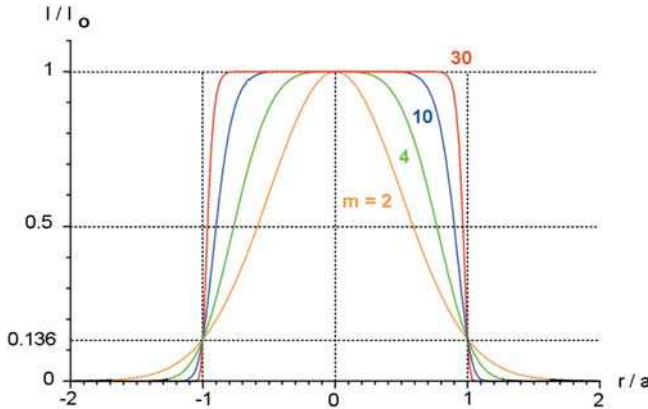


Figure 25-8: Super Gaussian profile for different exponential parameters m in one dimension.

Examples of point-spread function for pupils with super-Gaussian apodization and vanishing phase aberrations are compared in figure 25-9 on a logarithmic scale. It can be seen that the diffraction rings are absent only for the Gaussian profile with $m=2$, while for increasing power m the point-spread function becomes more and

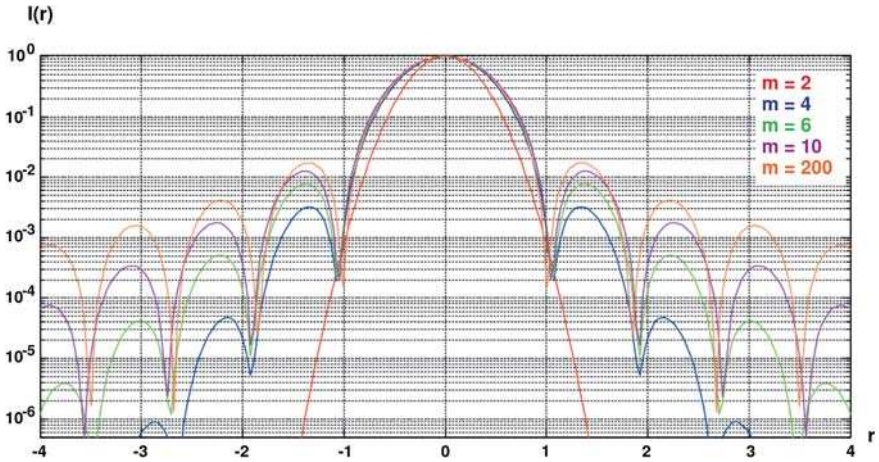


Figure 25-9: Intensity profiles of the focus of apodized pupils with super-Gaussian profiles and different exponential parameters m .

more similar to the Airy pattern. It is interesting to note that the radius of the dark rings or intensity minima is approximately constant. The value of m only has an effect on the height of the maxima.

25.3.3

Gaussian Profile with Truncation

Since the maximum aperture of optical systems is limited, the infinitely extended super-Gaussian function for the pupil function according to eq. (25-6) cannot be realized. The truncation by the aperture limit is of special interest for the collimation of Gaussian beams, such as laser beams, and is considered by the truncation ratio ε , given by

$$\varepsilon = \frac{D}{2w} \quad (25-7)$$

where the beam radius is w and the pupil diameter D [25-10], [25-11], [25-12]. Figure 25-10 illustrates truncated Gaussian profiles either by Gaussian illumination of an exit pupil or truncated apodization filters for different values of the truncation ratio ε . Truncation effects are mainly observed for truncation ratios $\varepsilon < 2$.

As the Airy pattern, the diameter of the point-spread function of a truncated Gaussian profile scales with λ/NA . The influence of the truncated apodization is considered by an empirical parameter K :

$$\delta r_{\text{spot}} = K \cdot \frac{\lambda}{NA} = K \cdot \frac{2 \cdot f \cdot \lambda}{D} \quad (25-8)$$

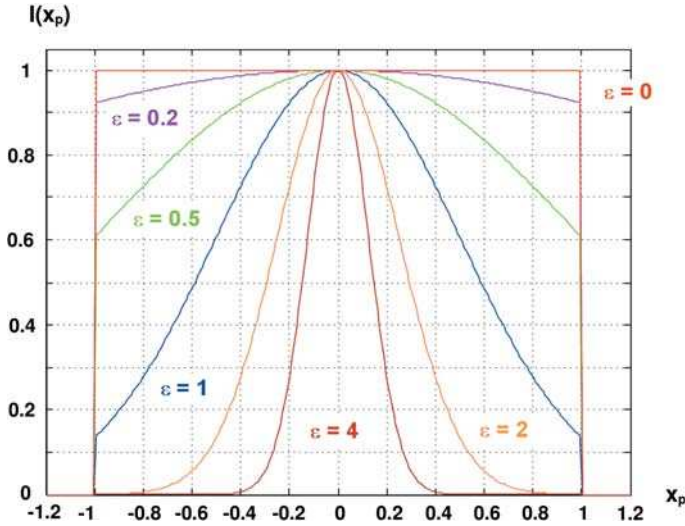


Figure 25-10: Intensity profiles of an apodized pupil with a truncated Gaussian profile for different truncation ratios.

where the pupil diameter is D and the focal length f . The parameter K depends on the apodization profile and the truncation ratio ϵ as well, and the focal spot size of a truncated beam cannot, in general, be evaluated analytically. Empirically, for the usual $1/e^2$ or the FWHM definitions, the factors can be approximated in the range $\epsilon < 2.5$ by [25-13]

$$K_{50\%} = 0.5145 + \frac{0.3562}{(1/\epsilon - 0.2161)^{2.176}} - \frac{0.3223}{(1/\epsilon - 0.2161)^{2.221}} \tag{25-9}$$

and

$$K_{13.5\%} = 0.8225 + \frac{0.3230}{(1/\epsilon - 0.2816)^{1.821}} - \frac{0.2660}{(1/\epsilon - 0.2816)^{1.891}} \tag{25-10}$$

Figure 25-11 compares the dependence of the spot sizes on the parameter ϵ according to exact values and the two definitions according to (25-9) and (25-10). Some typical intensity profiles are shown in figure 25-12.

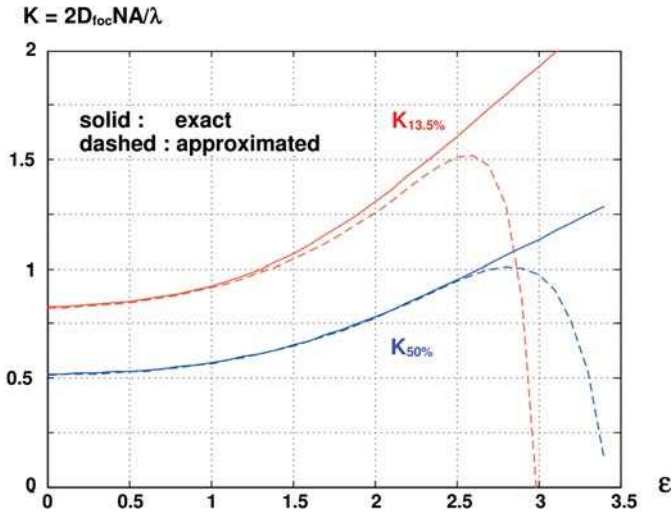


Figure 25-11: Correction factor K for the calculation of the focal spot diameter for an apodized pupil with a truncated Gaussian profile as a function of the truncation ratio ϵ .

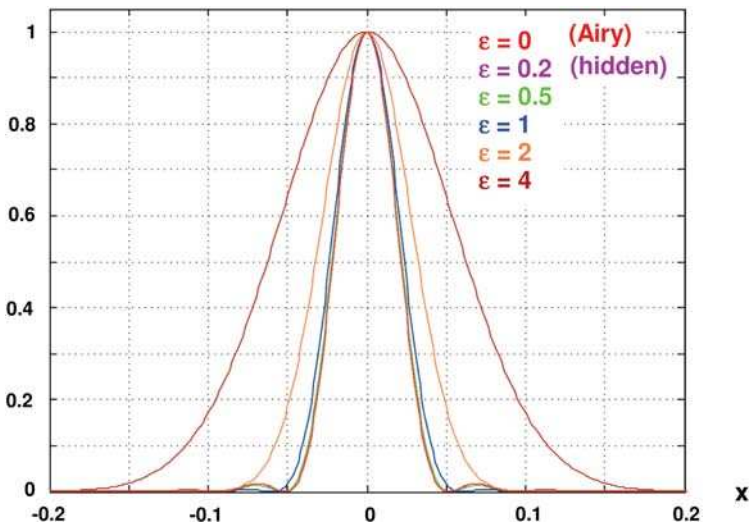


Figure 25-12: Intensity profiles of an apodized pupil with a truncated Gaussian profile for different truncation ratios.

In table 25-2, examples of focal spot sizes of pupils with truncation and apodization are compared, where different definitions of the spot size are applied. The definitions of the spot size by the first zero, as intensity thresholds (FWHM, $1/e^2$ radius) and by encircled energy (EncE) values are applied. As can be seen, the factor K varies by a factor of greater than 10, depending on the definition and shape of the pupil function.

Table 25-2: Comparison of focal spot sizes according to different definitions for different pupil functions.

Definition of focus criterion	1D	Gaussian	Super Gaussian	Circular (Airy)	Gaussian truncation $\varepsilon = 1$	Gaussian truncation $\varepsilon = 2/3$
m	0	2	6	0	2	2
ε	∞	∞	∞	∞	1	2/3
first zero	0.5	–	–	0.611	0.713	–
$I = 0.5$ (FWHM)	0.443	0.375	0.513	0.519	0.564	0.644
$I = 0.13534 = 1/e^2$	0.697	0.637	0.831	0.822	0.914	1.059
$I = 0.01$	0.908	0.966	1.122	1.092 (peak)	1.238	1.491
$I = 0.001$	0.969	1.183	2.109	1.174 (peak)	2.925	1.695
EncE = 0.86466	0.658	0.637	0.818	1.378	0.890	1.008
EncE = 0.95	1.989	0.779	1.040	3.915	1.104	1.201

In special applications, where it is important to overcome a threshold value, e.g., to trigger thermal or photo-chemical reactions, the main interest is not a small diameter of the focal spot but a very high energy density. An inhomogeneous illuminated or apodized pupil offers the advantage of influencing the focal intensity distribution at the expense of the total power. The absolute value of the focal point intensity is thus reduced. For a comparison of intensity peaks of different setups, it is thus necessary to take the integrated power into account. The amplitude of a homogeneous illuminated circular pupil of homogeneous transmission according to eq. (20-55) is given by

$$U(r) = \frac{NA}{r \cdot \lambda} J_1 \left(2\pi r \cdot \frac{NA}{\lambda} \right) = \pi \left(\frac{NA}{\lambda} \right)^2 \left[\frac{J_1 \left(2\pi r \cdot \frac{NA}{\lambda} \right)}{\pi \frac{NA}{\lambda} r} \right]. \quad (25-11)$$

With a constant Intensity I_p in the pupil, the peak intensity of the Airy distribution is thus given by [25-14]

$$I_{\text{airy}}(0) = |U(0)|^2 = \pi^2 \left(\frac{NA}{\lambda} \right)^4 I_p = \pi \left(\frac{NA}{\lambda} \right)^2 P \quad (25-12)$$

with the total incident Power P . For a Gaussian shape of width $w = D/\pi$ of the pupil transmission, corresponding to a truncation ratio of $\varepsilon = \pi/2$, the peak value of the intensity is given by

$$I_{GB}(0) = \frac{2\pi \cdot w^2}{\lambda^2 f^2} P = \frac{16w^4}{\pi^2 \lambda^2 f^2} I_p^{\text{Gauss}}(0) \quad (25-13)$$

where $I_p^{\text{Gauss}}(0)$ is the peak intensity at the centre of the pupil plane. For equal power P , however, the pupil intensity $I_p^{\text{Gauss}}(0)$ of the truncated Gaussian beam must be a factor of ~ 5 larger compared to the power I_p of the homogeneous pupil and Airy distribution. As a consequence, the Airy distribution of the homogeneous pupil has

a considerably higher peak intensity. As can be shown, a 30% loss of peak intensity in the Gaussian beam is observed.

In a general consideration of pupil transmissions with Gaussian profile and truncation ratio ϵ , the total power transmission of the pupil is

$$P = (1 - e^{-2\epsilon^2}) \cdot P_{in} . \tag{25-14}$$

For equal beam powers, the peak intensity of the truncated Gaussian beam at the pupil has to be increased in comparison to the top-hat profile by the relation

$$I_P^{Gauss}(0) = \frac{2\epsilon^2}{1 - e^{-2\epsilon^2}} \cdot I_P . \tag{25-15}$$

For equal total power, the peak value of the focal intensity is given by

$$I_{GB}(0) = \frac{2\pi a^2 P}{\epsilon^2 \lambda^2 f^2} \cdot (1 - e^{-\epsilon^2})^2 = \frac{2}{\epsilon^2} \cdot \frac{(1 - e^{-\epsilon^2})^2}{1 + e^{-2\epsilon^2}} \cdot I_{airy}(0) . \tag{25-16}$$

Figure 25-13 illustrates the peak intensity for constant input power P according to eq. (25-16). Uniform transmission of the pupil and homogeneous illumination gives the highest peak value. After a normalization to the total transmitted power, the maximum value of the peak intensity, eq. (25-16), is obtained at a truncation ratio of $\epsilon = 1.121$ or the beam radius $w = 0.892 D/2$. It is given by $\sim 81.45\%$ of the peak intensity of the Airy distribution.

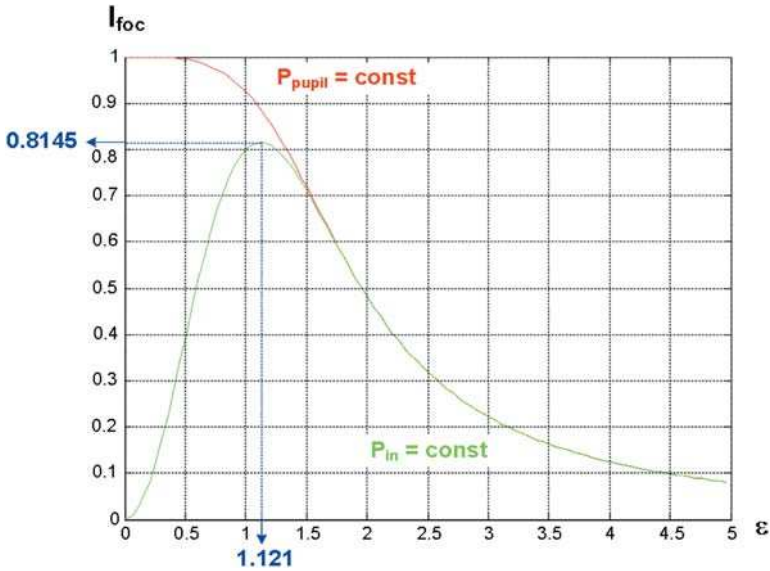


Figure 25-13: Intensity peak on the optical axis for focussed truncated Gaussian transmission profile of the pupil with and without renormalization of the total power.

25.4

Engineering of the Point-spread Function by Pupil Masks

25.4.1

Introduction

The complex pupil function $P(x_p, y_p)$ has a strong influence on the form of the focal intensity distribution $I(x, y, z)$. In particular, the phase distribution affects the behaviour very sensitively. In general there are three aspects, which are important for the function $I(x, y, z)$ [25-15]:

1. The geometry of the pupil area boundary.
2. The amplitude distribution inside the pupil area.
3. The phase inside the pupil area.

These dependencies can be used on the other side to form or reshape the focal caustic by introducing an appropriate complex filtering mask in the pupil in order to obtain some desired effects. The most important claims in the context of this point-spread function engineering are:

1. An improved transverse resolution.
2. An improved axial resolution.
3. An improved axial depth of focus.

If the illumination is fully coherent, the generation of the intensity distribution in the focal region can be understood by the interference of the Huygens elementary waves, emitted by the exit pupil of the system. Therefore, the different aims described above are not independent. The caustic has to be formed according to a practical application. In general, some drawbacks have to be accepted. First of all, it is necessary to characterize the caustic distribution, in order to have a quantitative measure which has to be optimized.

There are already two different situations, which have to be distinguished. In the first case, the pupil mask works as a wavefront coding device and changes the point-spread function in a passive way. In the second case, the primary detected image cannot be used directly. But if the mask and therefore the optical transfer function are known, with the help of a digital calculation, the generation of a final image can be performed in the computer.

25.4.2

Characterization of the Three-dimensional Point-spread Function

If enhancement of the resolution in either the longitudinal z -direction or the transverse x -direction is desired, the definition of the so called gain-factors is a possible way of generating a quantitative measure in order to describe the shaping of the caustic distribution [25-16], [25-17].

For simplification, in the following, the considered system is assumed to have rotational symmetry. Therefore, the description of the transverse direction is

reduced to the radial coordinate r . If the pupil function is denoted by $P(r_p)$ and the boundary of the pupil is at $r_p = a$, with the help of the coordinates in the image space, normalized in the usual way with aperture angle θ according to [25-14] is in the lateral direction

$$v = \frac{2\pi}{\lambda} \cdot r \cdot \sin \theta \quad (25-17)$$

and in the axial direction

$$u = \frac{2\pi}{\lambda} \cdot z \cdot \sin^2 \theta. \quad (25-18)$$

Then the field amplitude can be described in Fresnel approximation along the optical axis as

$$U(0, u) = 2 \int_0^1 P(r_p) \cdot e^{\frac{1}{2}iu \cdot r_p^2} r_p dr_p \quad (25-19)$$

and perpendicular to the axis in the ideal image plane as

$$U(v, 0) = 2 \int_0^1 P(r_p) \cdot J_0(v \cdot r_p) r_p dr_p. \quad (25-20)$$

With the help of the auxiliary parameter

$$t_p = \left(\frac{r_p}{a} \right)^2, \quad (25-21)$$

the following moments of the pupil function can be defined:

$$M_0 = \int_0^1 P(t_p) dt_p, \quad (25-22)$$

$$M_1 = \int_0^1 t_p \cdot P(t_p) dt_p, \quad (25-23)$$

$$M_2 = \int_0^1 t_p^2 \cdot P(t_p) dt_p. \quad (25-24)$$

For small values of the coordinates u and v , by a Taylor expansion of the Fresnel integral, the intensity near the axis in the vicinity of the image plane can be expressed in the form

$$\begin{aligned}
 I(u, v) = \frac{1}{|M_0|^2} & \left[|M_0|^2 - \frac{i u}{2} \cdot (M_0 M_1^* - M_0^* M_1) + \frac{u^2}{8} \cdot (2|M_1|^2 - M_0 M_2^* - M_0^* M_2) \right] \\
 & \cdot \left[|M_0|^2 - \frac{v^2}{4} \cdot (M_0 M_1^* + M_0^* M_1) + \frac{v^2}{64} \cdot (M_0 M_2^* + M_0^* M_2 + 4|M_1|^2) \right].
 \end{aligned} \tag{25-25}$$

This equation is a decoupled expansion of the intensity in the two directions up to second order in the variables u and v . In the special case of a real pupil function, this expression simplifies to

$$I(u, v) = |M_0|^2 \left[1 + \frac{u^2}{4} \cdot \left(\frac{M_1^2}{M_0^2} - \frac{M_2}{M_0} \right) \right] \cdot \left[1 - \frac{v^2}{2} \cdot \frac{M_1}{M_0} \right]. \tag{25-26}$$

If the focal caustic is manipulated by a complex pupil mask, this quadratic form can be used in a first-order approximation to describe the changes of the form quantitatively. The transverse gain factor describes an improved transverse resolution and is defined as

$$G_T = 2 \frac{M_1}{M_0}. \tag{25-27}$$

If the value of $G_T > 1$, the resolution is increased in comparison with the case of the ideal point-spread function of Airy. Similarly, the axial gain factor is defined in the equation

$$G_A = 12 \cdot \left[\frac{M_2}{M_0} - \left(\frac{M_1}{M_0} \right)^2 \right]. \tag{25-28}$$

If $G_A > 1$, the system shows an increased axial resolution, $G_T < 1$ means that the system has an enlarged depth of focus. Using these two gain parameters, in the first approximation, the intensity distribution can be expressed in quadratic form as

$$I(v, u) = I_0 \cdot \left(1 - \frac{1}{4} \cdot v^2 \cdot G_T \right) \cdot \left(1 - \frac{1}{48} \cdot u^2 \cdot G_A \right). \tag{25-29}$$

It should be noted that, in addition to the stretching of the caustic intensity volume, depending on the moments M_j , the centre of the volume suffers a shift along the optical axis of amount

$$u_{\text{peak}} = \frac{-2 \cdot \text{Im}(M_0^* M_1)}{\text{Re}(M_2^* M_0) - |M_1|^2}. \tag{25-30}$$

Figure 25-14 shows, for the special case of the ideal point-spread function, the quadratic approximation of the intensity in the two directions.

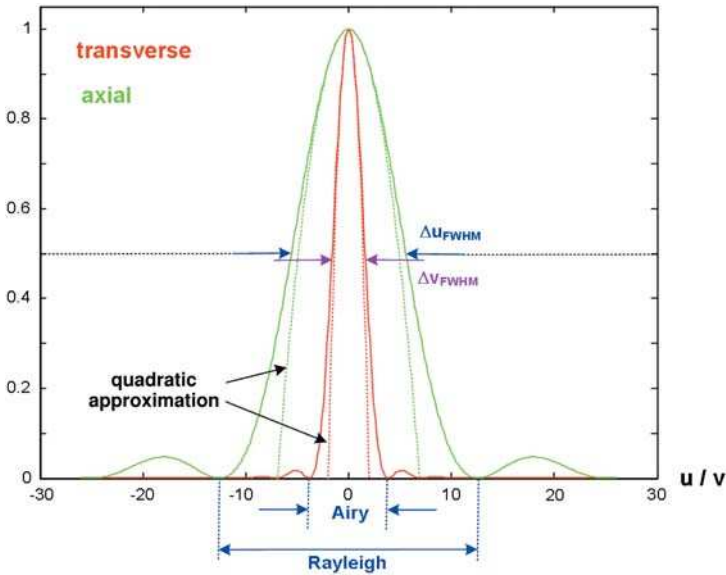


Figure 25-14: Ideal point-spread function in the transverse and axial direction and quadratic approximation of the peak width.

For the assessment of the imaging quality, both the relative size of the focus volume and also the absolute height of the peak intensity, measured with the help of the Strehl number D_S , are relevant. The value of D_S can be expressed with the help of the moments as

$$D_S = |M_0|^2 - u_{\text{peak}} \cdot \text{Im}(M_0^* M_1). \quad (25-31)$$

In addition, for practical reasons, the energy throughput of the system is an important measure. If the pupil function is not only imaginary, a finite amplitude transmission reduces the throughput of the system and therefore deteriorates the signal-to-noise ratio.

Furthermore, the contrast of the imaging is influenced by enlarged side lobes of the central focal peak.

In the case of a system with extended depth of focus, the uniformity of the intensity distribution along the axis is a particularly important criterion for the application.

25.4.3

Characterization of Extended Depth of Focus

If the amplitude in the image plane at $z = 0$ serves as a reference, a correlation of the defocused amplitude with this reference may serve as a quality measure [25-15]

$$\varepsilon^2(z) = 1 - \frac{\left| \iint U_{\text{ref}}^*(x, y, 0) \cdot U(x, y, z) dx dy \right|^2}{\left(\iint |U_{\text{ref}}(x, y, 0)|^2 dx dy \right)^2 \cdot \left(\iint |U(x, y, z)|^2 dx dy \right)^2}. \quad (25-32)$$

If the peak value of the focal intensity is shifted along the optical axis, the reference profile has to be chosen at another z position.

As a special form of scaling this correlation, the so called Hilbert space angle θ_H can be used to characterize the extended depth of focus. It is defined with the help of the intensity correlation as

$$\cos \theta_H(z) = \frac{\int I(x, z) \cdot I(x, 0) dx}{\sqrt{\int I^2(x, z) dx} \cdot \sqrt{\int I^2(x, 0) dx}}. \quad (25-33)$$

If the angle has a small value, the point-spread function profiles agree well and the uniformity during defocusing is good.

From the viewpoint of information theory, it is usual to use the defocus criterion according to Fisher. It describes a point of stationarity of the incoherent transfer function as a function of the defocus and is defined by the equation

$$J(z) = \int \left| \frac{\partial H_{OTF}(v, z)}{\partial z} \right|^2 dv. \quad (25-34)$$

25.4.4

Relation Between Axial and Transverse Resolution

For the ideal point-spread function, the lateral resolution is usually described by the Airy radius

$$\Delta x_{\text{airy}} = \frac{0.61 \cdot \lambda}{n \cdot \sin \theta}. \quad (25-35)$$

The half-width at half-maximum is given numerically by

$$\Delta x_{\text{fwhm}} = 0.258 \cdot \frac{\lambda}{n \sin u}. \quad (25-36)$$

In the quadratic approximation of the central peak

$$I(x) = I_0 \cdot \left[1 - \left(\frac{x}{\Delta x} \right)^2 + \dots \right] \quad (25-37)$$

it follows that

$$\Delta x_{\text{quad}} = \frac{\lambda}{\pi \cdot \sin u}. \quad (25-38)$$

Similarly, in the longitudinal direction, the depth is classically defined by the Rayleigh range

$$\Delta z_{\text{rayleigh}} = 2R_E = \frac{2 \cdot \lambda}{n \cdot \sin^2 \theta}. \quad (25-39)$$

The half-width at half-maximum results in

$$\Delta z_{\text{fwhm}} = 0.886 \cdot R_E = \frac{0.886 \cdot \lambda}{n \cdot \sin^2 \theta} \quad (25-40)$$

and the quadratic approximation gives the value

$$\Delta z_{\text{quad}} = \frac{2 \cdot \sqrt{3} \cdot \lambda}{\pi \cdot n \cdot \sin^2 \theta} = \frac{1.103 \cdot \lambda}{n \cdot \sin^2 \theta}. \quad (25-41)$$

Depending on these definitions, the ratio of the longitudinal to the transverse resolution takes the values given in table 25-3.

Table 25-3: Ratio of the axial to the lateral resolution.

Definition	Ratio $\eta = \frac{\Delta z}{\Delta x}$
Airy/Rayleigh, zero points	$\eta_{\text{ideal}} = \frac{3.28}{n \cdot \sin \theta}$
FWHM	$\eta_{\text{fwhm}} = \frac{3.43}{n \cdot \sin \theta}$
Quadratic approximation	$\eta_{\text{quad}} = \frac{3.47}{n \cdot \sin \theta}$

As can be seen from the table, the ratios between the resolutions in the axial and lateral direction only depend on the numerical aperture of the setup. As a consequence, it can be seen that an isotropic focal volume is only possible with an extremely high numerical aperture of 3.4. This can only be realized in media with a very high refractive index for an aperture angle near to 90°. This means that, in reality, the focal volume always has a prolate elliptical form.

If the numerical aperture is eliminated in the above equations, it follows the simple relationship

$$\frac{\Delta z_{\text{rayleigh}}}{\lambda} = 4n \cdot \left(\frac{\Delta x_{\text{airy}}}{\lambda} \right)^2 \quad (25-42)$$

if the factor in the Airy diameter is approximated as one. This means that only the refractive index can influence this ratio of the axial resolution to the square of the lateral resolution, if the resolution values are scaled to the wavelength.

25.4.5

Ambiguity Function as Defocussed Transfer Function

According to Duffieux, the incoherent transfer function can be written as an auto-correlation of the pupil function P in the form

$$H_{OTF}(v) = \int P\left(x_p + \frac{\lambda \cdot f \cdot v}{2}\right) \cdot P^*\left(x_p - \frac{\lambda \cdot f \cdot v}{2}\right) dx_p. \quad (25-43)$$

If in the pupil function here a defocusing term, z , is introduced,

$$P(x_p) = P(x_p, 0) \cdot e^{i\pi \cdot \lambda \cdot z \cdot v^2} \quad (25-44)$$

the transfer function then follows as

$$H_{OTF}(v, z) = \int P\left(x_p + \frac{\lambda \cdot f \cdot v}{2}\right) \cdot P^*\left(x_p - \frac{\lambda \cdot f \cdot v}{2}\right) \cdot e^{2\pi \cdot i \cdot \frac{z}{\lambda} \cdot x_p \cdot v} dx_p. \quad (25-45)$$

This corresponds to the definition of the ambiguity function A [25-18]

$$H_{OTF}(v, z) = A(v, v \cdot \tan \varphi) \quad (25-46)$$

if the rotation angle φ is introduced by

$$\tan \varphi = \frac{z}{\lambda} \cdot n \cdot \sin^2 \theta. \quad (25-47)$$

The interpretation of this equation shows that the ambiguity function is a representation of the transfer function for defocussed systems, the defocussing is given by the rotation angle φ . Figure 25-15 shows this in one dimension. A section through the function distribution under the angle φ gives the transfer function for the corresponding value of defocus. This is shown in figure 25-16.

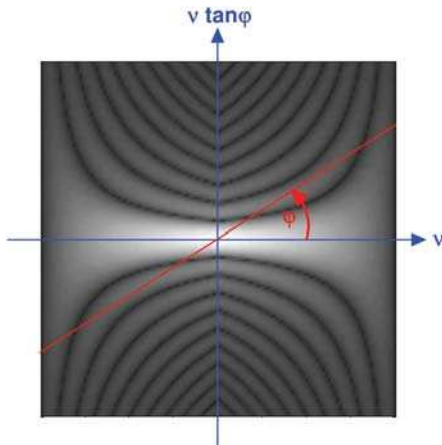


Figure 25-15: Ambiguity function as a transfer function for defocussed systems. The rotation angle φ corresponds to the value of defocus.

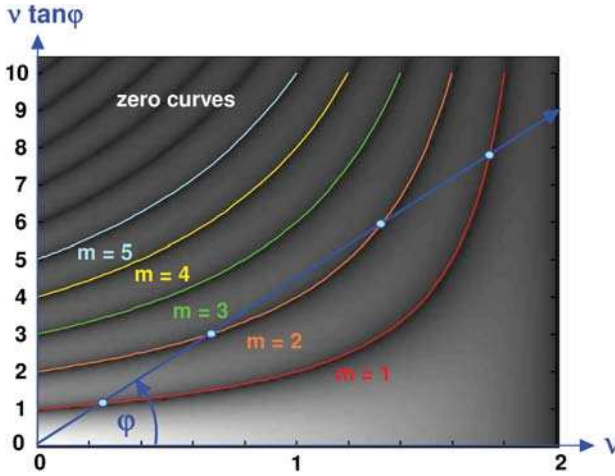


Figure 25-16: Ambiguity function as a transfer function for defocused systems. With increasing defocusing, an increasing number of zeros are crossed in the transfer function.

25.4.6

Image Multiplexing

Application of a linear grating in the pupil plane causes multiple reproduction of the object pattern in the image plane (figure 25-17a). Multiplication of the object spectrum by a grating of period d and shape function $g(v)$ for the profile of each period,

$$u_1(v) = u_0(v) \cdot \text{comb}\left(\frac{v}{d}\right) \otimes g(v) \tag{25-48}$$

leads to an image amplitude given by

$$U_1(x) = |d| \cdot U_0(x) \otimes \text{comb}(d \cdot x) \cdot G(x) . \tag{25-49}$$

The image pattern is thus a repeated function of the object with intensities $\sim G^2(x)$, the Fourier transform of the shape function $g(v)$ of the grating. To obtain identical image intensities for each copy, special gratings have been developed, now frequently called Dammann gratings [25-19].

As shown in chapter 23 in more detail, high-NA imaging systems, in particular, suffer from small depth of focus. It was proposed therefore to apply pupil filters for enhancing the depth of focus. One method proved more popular than the others because it is simple to implement. The depth of focus is enhanced by multiple focus positions z and $-z$, centred at the ideal focus $z = 0$. With quadratic phase change due to defocus z_4 and the application of eq. (20-39) the coherent image amplitude is obtained by

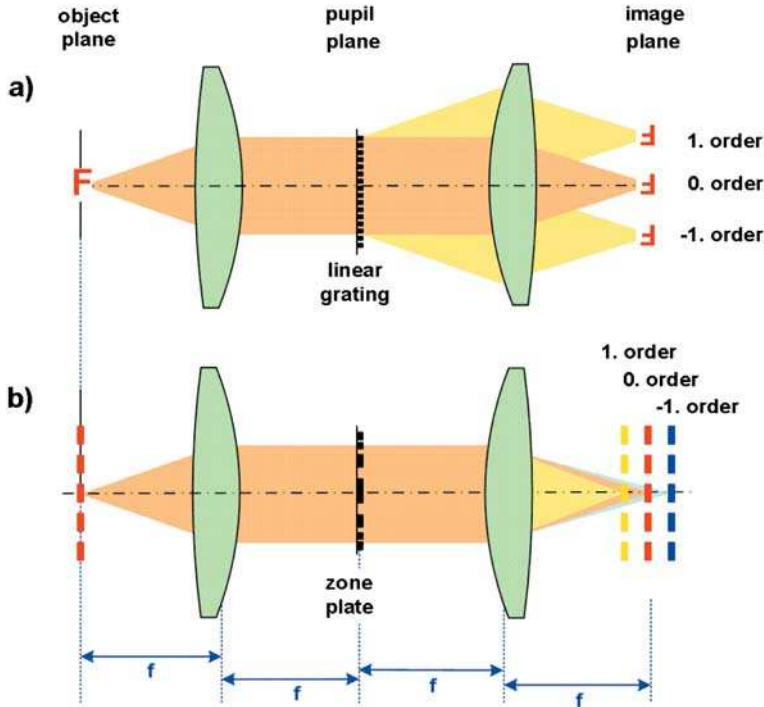


Figure 25-17: a) In-line image multiplexing by application of a linear grating in the pupil plane. b) Imaging with extended depth-of-focus by application of a Fresnel Zone Plate in the pupil plane.

$$U(x, y, z) = \iint t(v_x, v_y) \cdot h(v_x, v_y) \cdot e^{i2\pi\frac{NA^2}{4}z \cdot (2\rho^2 - 1)} \cdot e^{2\pi i(v_x x + v_y y)} dv_x dv_y \quad (25-50)$$

with $\rho = \sqrt{v_x^2 + v_y^2}$ [25-20]. The pupil function for two image positions with defocus $\pm \delta z$ can be written as

$$P(\rho) = \frac{1}{2} \left[e^{i\pi\frac{NA^2}{2}z \cdot (2\rho^2 - 1)} + e^{-i\pi\frac{NA^2}{2}z \cdot (2\rho^2 - 1)} \right] = \cos \left(\pi NA^2 z \cdot \rho^2 - \frac{\pi NA^2}{2} z \right). \quad (25-51)$$

Figure 25-18 illustrates the phase filter for $NA = 1$ for several z . Such phase filters have been approximated by binary phase filters [25-21]. As can easily be seen, pupil filters according to eq. (25-51) are equivalent to binary phase Fresnel lenses of low power, producing $\pm 1^{\text{st}}$ diffraction orders with small defocus (figure 25-17b). The interpretation of pupil filters for longitudinal image multiplexing as Fresnel diffraction gratings offers a wide variety of special filters for depth of focus enhancement. Similar filters have been described by polynomial expansions of eq. (25-51) [25-22]. However, pupil filters of this type are not required since the image positions can also be stepped through [25-21] and integrated over time, or, for scanning imaging, by tilting the detector plane.

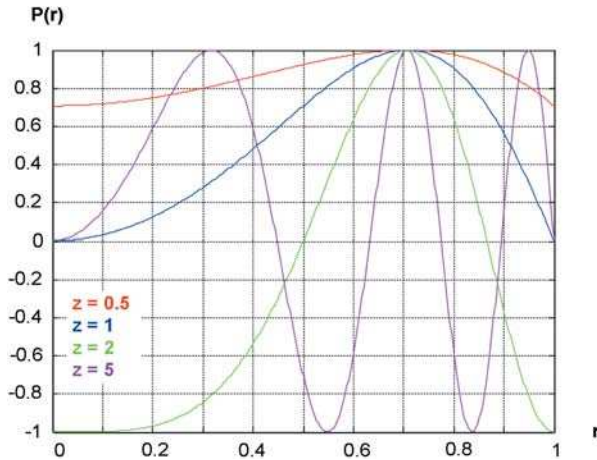


Figure 25-18: Pupil filters for depth of focus enhancement by focus plane doubling.

25.4.7

Fundamental Relationships

There are some fundamental relationships between the limits of resolution and the caustic parameters in general, which can be deduced from simple analytical treatment of the diffraction integral and Fourier optics [25-23], [25-24]. These are:

1. In the case of a transverse super resolution, the Strehl number must necessarily be smaller than one. This decrease in the Strehl ratio and therefore the contrast of the imaging setup is independent of the functional realization of the complex mask.
2. The largest Strehl ratios are obtained for pure phase masks.
3. There is no principal limit known for the improvement of the lateral resolution. But the decrease in the Strehl number for increasing resolution, results in a limited practical benefit of this type of arrangement.
4. The improvement in the axial resolution has a factor of two for amplitude masks and 2.5 for phase masks as a principal limit.
5. Pure phase masks always work asymmetrically around the ideal image plane, while transmission masks do not.

25.4.8

Calculation of Masks

There are several possible ways to determine pupil phase masks for special purposes concerned with the modification of the point-spread function in three dimensions [25-25].

For some simple special cases, an optimization criterion can be evaluated by first principles. If the uniformity of the axial distribution of the caustic is required, then from the stationary phase method, a radial symmetric phase mask with a logarithmic form can be obtained. If, on the other hand, it is assumed that the modulation transfer function has a stationary point, a cubic curve of the phase is obtained in one dimension.

These solutions are usually restricted to some special circumstances such as a one-dimensional system, rotational symmetry, paraxial imaging, a small amount of defocus, considering only the full width at half maximum of the caustic intensity function, or something else.

For practical applications, it is often recommended that solutions of this kind should be improved by numerical methods to take some real conditions into account. But the use of the simple solutions often gives a good starting point for general optimization algorithms.

In reality, many conditions and assumptions of these basic calculations are not fulfilled. Some of these occurring in practice are, for example:

1. There are several wavelengths in use.
2. The pupil is not exactly circular.
3. It is not possible to locate the mask exactly in the pupil plane for technical reasons.
4. There are pupil aberrations which have to be taken into account.

To get the best solution over all the conditions valid in the considered system, a numerical least-squares fit optimization can help to obtain a mask with the best performance. On one hand, it is possible to describe the pupil by a complex discretized function, which is calculated with the help of an algorithm of the Gerchberg-Saxton type [25-25], [25-26] or an equivalent numerical computation scheme.

Other methods describe the pupil function by simple appropriate functional forms such as polynomial expansions and then optimize a few parameters.

25.5

Special Pupil Masks

25.5.1

Introduction

Apart from the desired effect of a complex pupil mask, there are several different possible ways to perform point-spread function engineering. The first attempt is to introduce a pure imaginary phase filter in the pupil of an optical system. The advantage of this kind of mask is the transmittance, since there are no absorption effects to attenuate the energy of the imaging.

The second form is a purely absorbing filter mask in the pupil. This offers some advantages; in particular, an unwanted shift of the image plane can be avoided.

In the most general case, the pupil filter is constructed as a complex mask with phase and absorbing effects. In this case, the greatest degrees of freedom can be used to shape the focal caustic. The structuring of the pupil function to modify the intensity distribution in the image space is sometimes called wavefront coding.

In principle, there are two different ways to attempt this. In the first case, the shaped caustic is used directly for image formation. In the second and more complex case, the primary image is detected with a digital sensor and the final image formation is calculated in post-processing algorithms with the help of the knowledge of the pupil mask. This more sophisticated form of digital image formation offers the possibility of generating quite new imaging aspects.

25.5.2

Phase Masks According to Toraldo

One of the oldest methods of introducing pupil masks follows the proposal of Toraldo di Francia [25-3], [25-27], [25-28]. The pupil is modified using a system of digital phase rings with phase steps of π . Depending on the number of the radii and the values of the ring boundaries, different effects can be achieved.

The simplest system of this kind is the well known Fresnel zone plate, which produces a focussing effect in different orders, according to the principle of grating diffraction. These radii correspond to the radii of the Fresnel zones, so the sign of the destructive interference zones are inverted by the phase plate. Figure 25-19 illustrates this simple example.

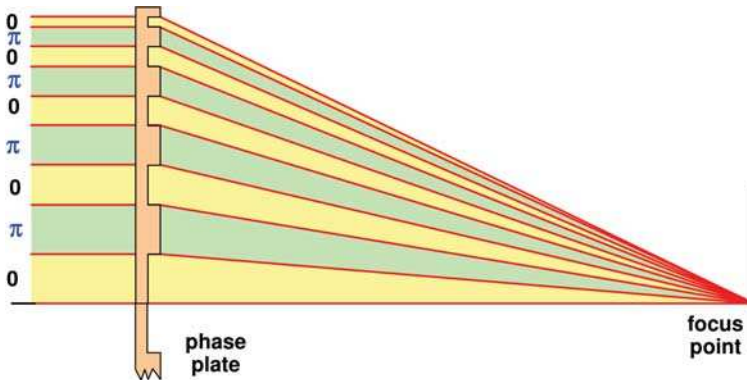


Figure 25-19: Fresnel zone plate as a simple phase mask according to the principle of Toraldo di Francia.

If, in the more general case, the ring radii are given by ε_j as the relative values in respect to the outer radius of the pupil circle with the phase values Φ_j , one gets a system, as illustrated in figure 25-20.

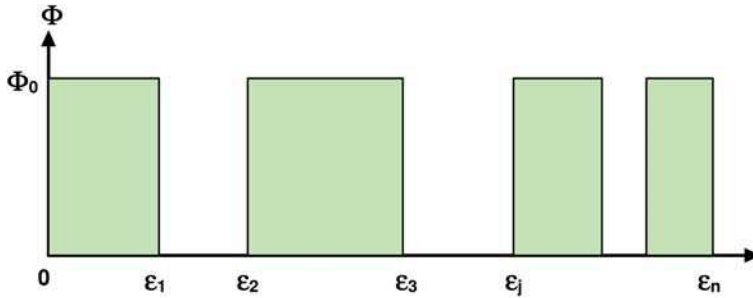


Figure 25-20: General phase mask according to the principle of Toraldo di Francia.

In this case, the field in the ideal image plane is given by the equation

$$U(r) = \sum_{j=1}^n e^{2\pi i \Phi_j} \cdot \left[\varepsilon_j^2 \cdot \frac{2J_1(kr \sin u' \varepsilon_j)}{kr \sin u' \varepsilon_j} - \varepsilon_{j-1}^2 \cdot \frac{2J_1(kr \sin u' \varepsilon_{j-1})}{kr \sin u' \varepsilon_{j-1}} \right] \quad (25-52)$$

where u' is the aperture angle in the image space.

As an example, if three rings are chosen with relative separating radii $\varepsilon_1=0.35$ and $\varepsilon_2=0.62$ and a phase step π between neighboring zones [25-29], [25-30], an extended depth of focus is achieved, as indicated in figure 25-21.

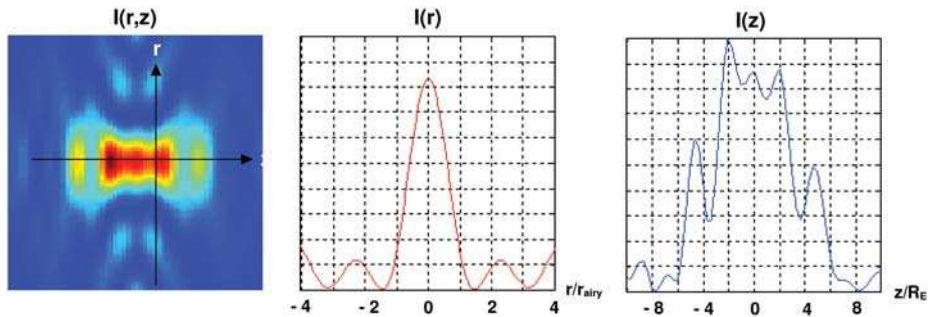


Figure 25-21: Phase mask with three zones according to the principle of Toraldo di Francia, with extended depth of focus.

25.5.3

Logarithmic Phase Mask

If the condition of a symmetric extended depth of focus for a constant intensity along the optical axis with the help of the method of stationary phase is assumed, one gets the following equation, where A , B and u_{image} are parameters describing the demands [25-31]

$$P(x_p) = A \cdot \text{sign}(x_p) \cdot a^2 \cdot x_p^2 \cdot (\log|x_p| + B) - \frac{a \cdot x_p}{f} \cdot u_{\text{image}} \cdot \quad (25-53)$$

If a depth of focus Δz is required, an explicit form of the phase mask is given by

$$\begin{aligned} \Phi_{\text{Log}}(r) = & -\frac{\sqrt{r^2 + f^2}}{\lambda} + \frac{r^2}{2\lambda} \cdot \left(\frac{1}{f + \Delta z/2} - \frac{1}{f} \right) \\ & + \frac{a^2}{2\lambda\Delta z} \cdot \ln \left[1 - \frac{2f\Delta z}{a^2} + \frac{2r^2f \cdot \Delta z^2}{a^4} + \frac{2\Delta z}{a^2} \cdot \sqrt{r^2 + \left(f - \frac{fr^2}{a^2} \right)^2} \right]. \end{aligned} \quad (25-54)$$

Figure 25-22 shows the distribution of the phase for this kind of mask. The resulting intensity distribution in the image space is shown in figure 25-23 for a value of $\Delta z = 10R_E$. As can be seen from the curves, the uniformity of the intensity along the axis is far from being perfect. In particular, it can be stated that the form of the caustic focal volume has a form like a cone with a transverse broadening in the intra-focal direction towards the system.

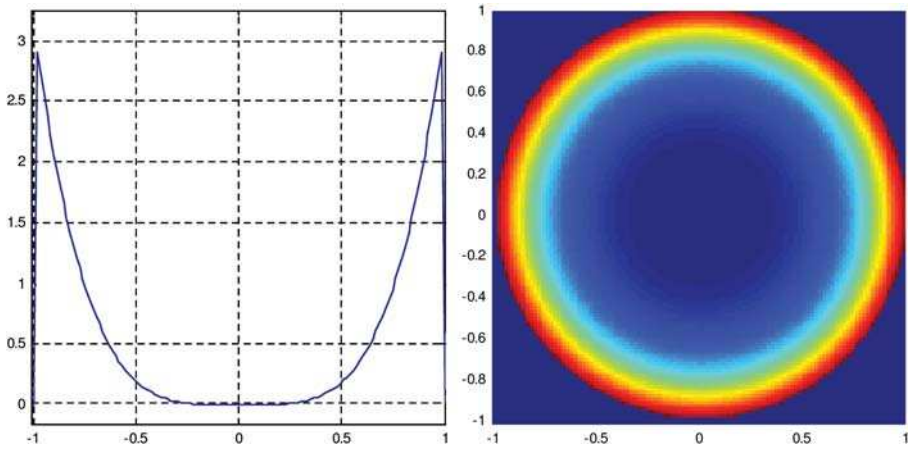


Figure 25-22: Phase mask with a logarithmic distribution of the phase.

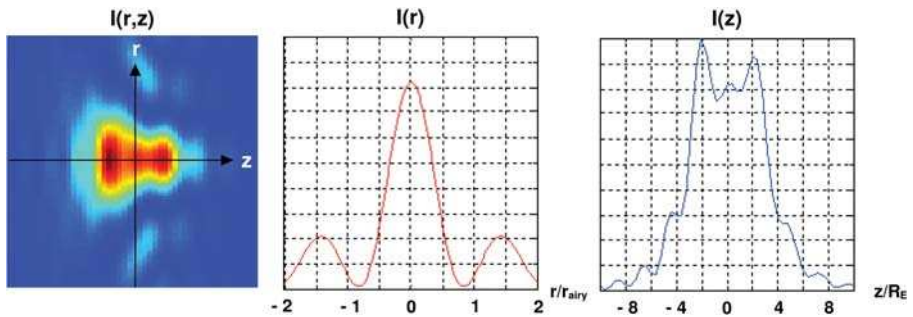


Figure 25-23: Focal intensity profile for a phase mask with a logarithmic distribution of the phase and a desired depth of focus of $\Delta z = 10 R_E$.

25.5.4

Chirped Ring Pupil

It can be demonstrated in a very elementary way, as shown in chapter 25.4.6 in form of the image multiplexing principle, that the axial intensity distribution is equal to the Fourier transform of the pupil function, expressed by the squared coordinate r_p^2 . If the pupil is divided into single ring zones, every zone can be apodized by a purely transmission function with a functional dependence proportional to $\cos(ar_p^2)$ to generate a sinc-formed intensity distribution in the image space, shifted along the axis in a controlled manner [25-32]. If the various sinc functions are arranged at an appropriate distance, the superposition of all these contributions forms a caustic with an extended depth of focus. This idea is illustrated in figure 25-24.

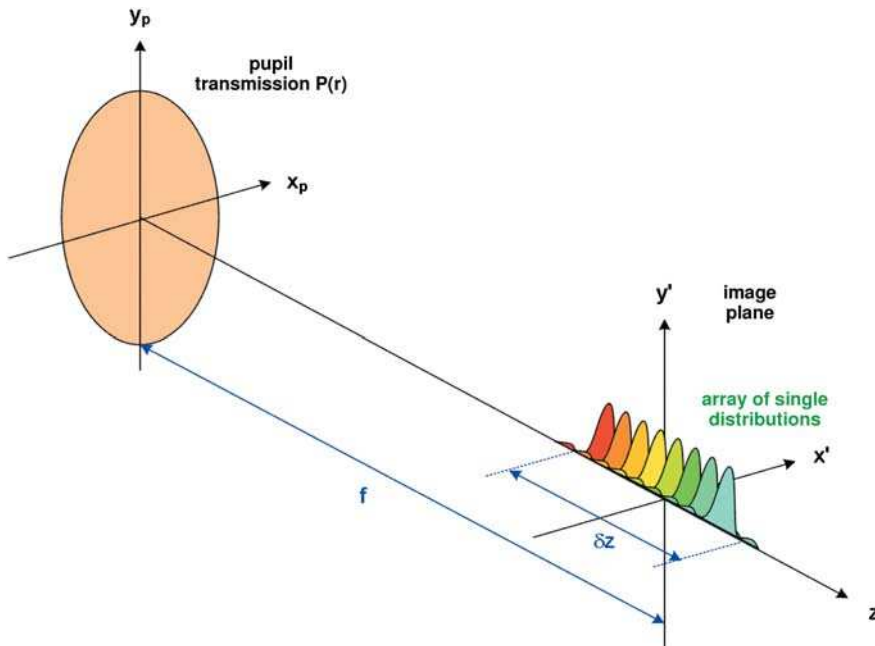


Figure 25-24: Chirped amplitude pupil for extended depth of focus.

The amplitude can be described by a convolution in the form

$$U(z) = \sum_{n=-m}^m \delta(z - n \cdot \delta z) \cdot \text{sinc}(z) \tag{25-55}$$

if it is assumed, that $2m+1$ single sinc functions are superimposed. From this expression, the pupil function is derived by the chirp function

$$P(r_p) = \frac{1}{2m + 1} \cdot \left[1 + 2 \cdot \sum_{n=1}^m (-1)^n \cdot \cos \left(2\pi \cdot n \cdot r_p^2 \right) \right]. \tag{25-56}$$

This distribution shows a sharp peak at the relative pupil radius $1/\sqrt{2}$. Figure 25-25 shows the typical behaviour of this function for $m=8$. One of the major drawbacks of this attempt is the fact that, for a sufficiently high depth of focus, the value of m must also be sufficiently large and then the integral transmission of the pupil reduces to

$$T = \frac{1}{(2m + 1)^2} \cdot \tag{25-57}$$

For the above example, with $m=8$ and a corresponding extension of the depth of focus of approximately a factor of 8, a transmission of $T = 0.34\%$ follows, which is rather low. But, as figure 25-26 shows, the performance of the mask concerning the focus shaping is excellent in both the axial and the transverse directions. As every ring zone acts at another position, the lateral resolution is given by the typical properties of a ring aperture. This can be seen very clearly in the picture in the middle of figure 25-26, the low convergence of the outer diffraction rings decreases the contrast of the imaging system significantly.

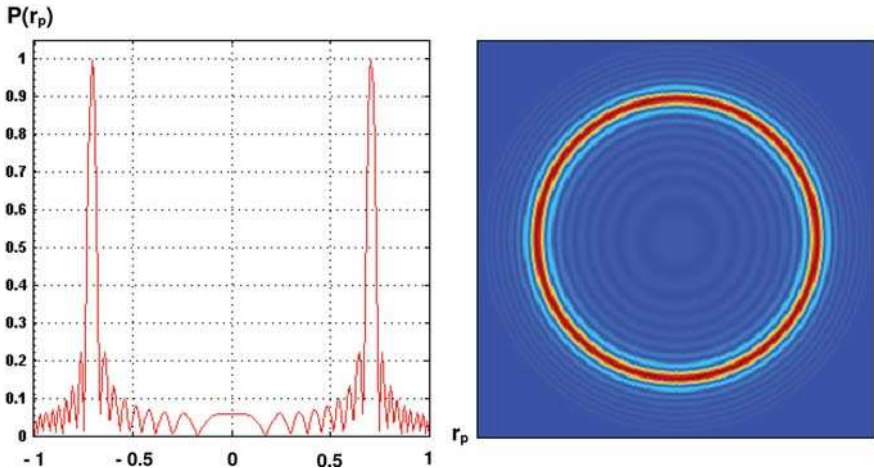


Figure 25-25: Chirped cos-amplitude mask for extended depth of focus generation.

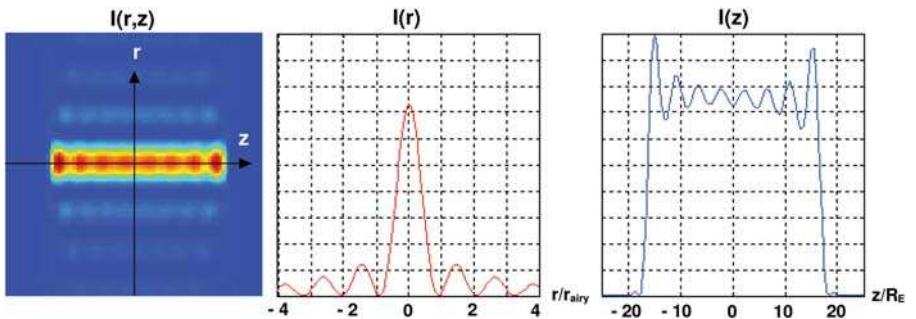


Figure 25-26: Intensity distribution of a chirped cos-amplitude mask for extended depth of focus generation.

25.5.5

Complex Filter Described by Zernike Expansions

In a very general attempt, the pupil mask can be described by Zernike expansions, to allow arbitrary profiles for the transmission and the phase function.

Figure 25-27 shows one example of a complex pupil filter of this kind, which generates a caustic in the image space with an improved axial resolution of approximately 40 %. Figure 25-28 shows the corresponding intensity distribution. As can be seen in the picture in the middle of the figure, the transverse resolution is hardly affected by the mask.

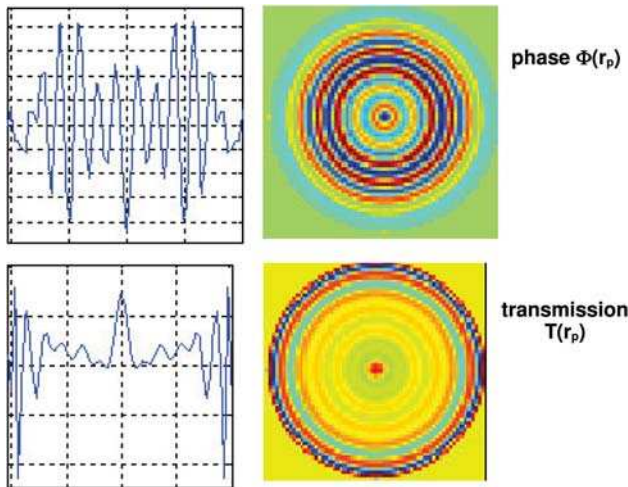


Figure 25-27: Complex decomposed Zernike mask for high axial resolution.

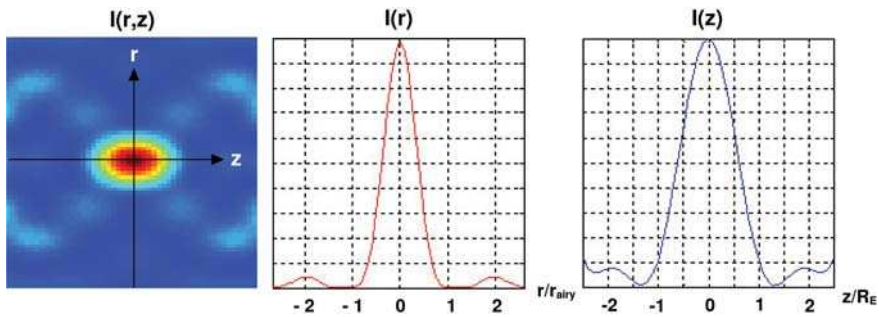


Figure 25-28: Intensity distribution of the complex decomposed Zernike mask for high axial resolution of figure 25-27.

As a second example, the mask in figure 25-29 exhibits a super resolution effect in the transverse direction, with an improved lateral resolution of 44 %. This is shown in figure 25-30.

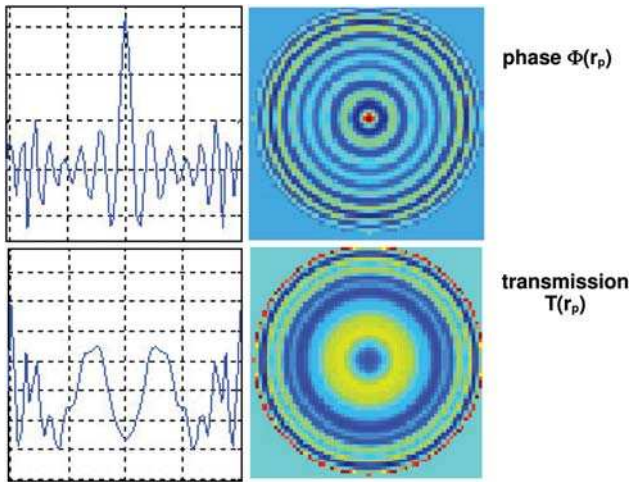


Figure 25-29: Complex decomposed Zernike mask for high lateral resolution.

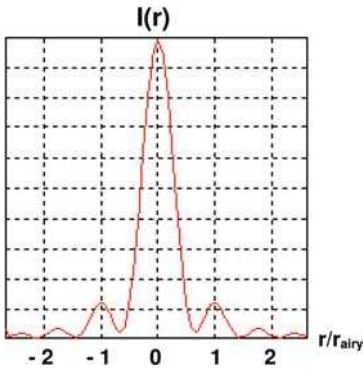


Figure 25-30: Intensity distribution of the complex decomposed Zernike mask for high lateral resolution of figure 25-29.

As a third example, figures 25-31 and 25-32 show a complex mask, described by a Zernike expansion for the phase and the amplitude function, which delivers an extended depth of focus with a factor of approximately 7.

Figure 25-33 shows an example of the modulation transfer function as a function of the defocussing z for the above system with and without the mask. It can be seen that, for the normal system as shown in figure a, the transfer function decreases very slowly for quite low defocussing values. In the case of the system with the mask, the slow-down of the transfer function with the spatial frequency is quicker, but it demonstrates a rather good uniformity in the defocus range.

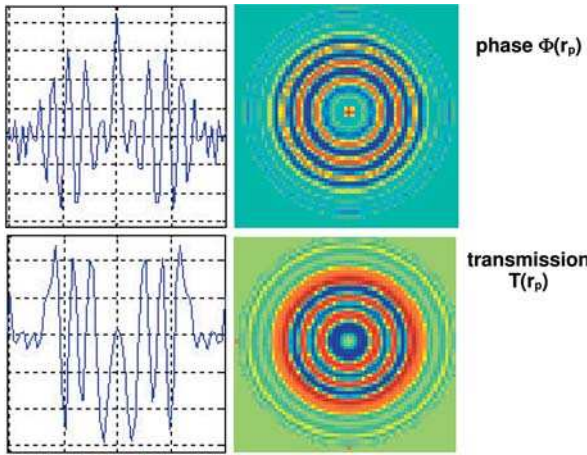


Figure 25-31: Complex decomposed Zernike mask for an enlarged depth of focus.

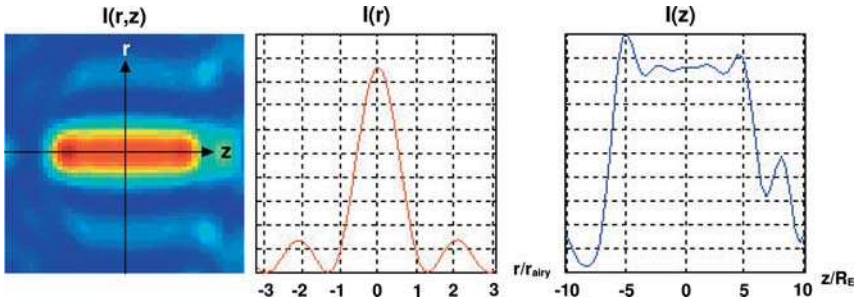


Figure 25-32: Intensity distribution of the complex decomposed Zernike mask for an enlarged depth of focus of figure 25-31.

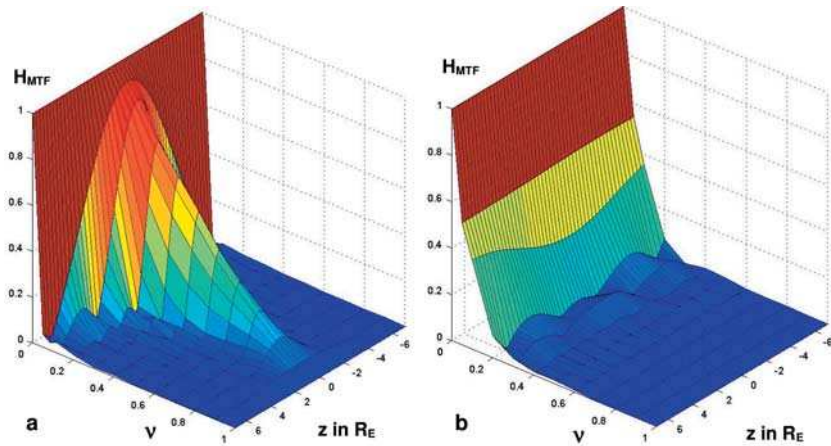


Figure 25-33: Modulation transfer function as a function of the defocussing z of a conventional system and a system with an extended depth of focus, for comparison.

25.5.6

Cubic Phase Plates for Extended Depth of Focus

One of the best known examples for obtaining a system with an enlarged extended depth of focus is the insertion of a cubic phase plate in the pupil of an optical system [25-33], [25-34], [25-15]. The pupil mask generates a primary image of poorer quality, but after deconvolution of the digital detected image with the known transfer function, an image of quite good quality can be obtained, which is in first approximation independent of the defocussing in a certain range.

If x and y are the normalized pupil coordinates and α is a parameter which scales the strength of the pupil filter, the system works with the pure imaginary pupil mask of the form

$$P_{\text{mask}}(x_p, y_p) = e^{i\alpha(x_p^3 + y_p^3)}. \quad (25-58)$$

In the easiest implementation, the pupil boundary is formed as a quadratic opening, so the coordinate directions x and y are fully decoupled. Figure 25-34 shows the corresponding phase surface.

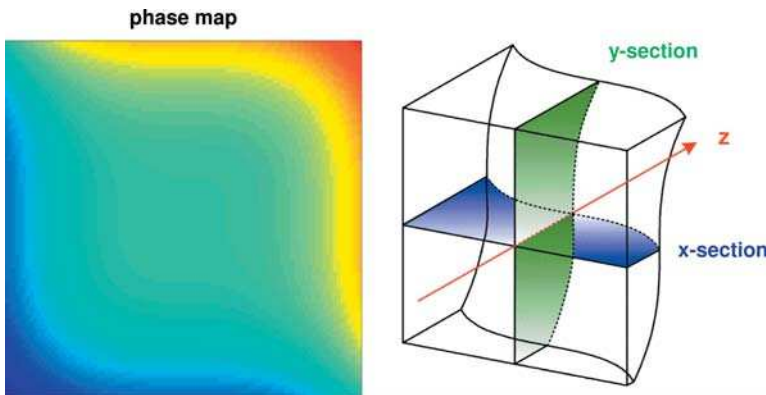


Figure 25-34: Phase function of a cubic phase plate as a contour representation and in form of a perspective sketch.

The parameter α defines the effect of the filter and determines the depth of focus, which is reached with the mask. Typical values of this parameter are in the range $\alpha > 20$.

Figure 25-35 shows the form of the point-spread function of a system with a cubic phase plate. As can be seen, the phase distortion generates a point-spread function with a very large extent, which resembles a coma-type point-spread function in the x and y directions. In reality, the extension of the depth of focus is large enough, if the parameter α is large. This then causes a very broad point-spread function, so the primary image is of very poor quality.

Figure 25-36 illustrates the form of the corresponding modulation transfer function. The transfer is very high along the coordinate directions, but decreases very

quickly in the diagonal directions. The long range of a nearly constant modulation along the axes is typical. If the system should be deconvolved over a large range of defocus, the transfer function is not allowed to have zero points. Therefore, a low but almost constant value is desired for the curve of the modulation as a function of the defocus.

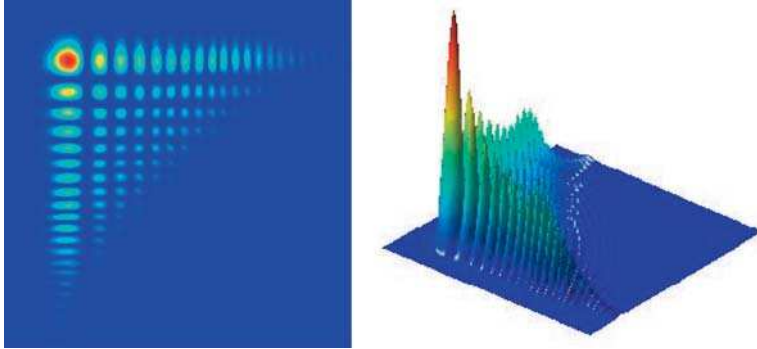


Figure 25-35: Point-spread function and modulation transfer function of a system with a cubic phase plate.

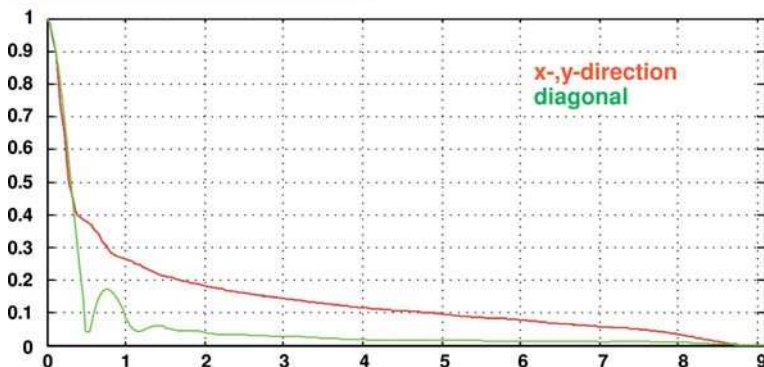
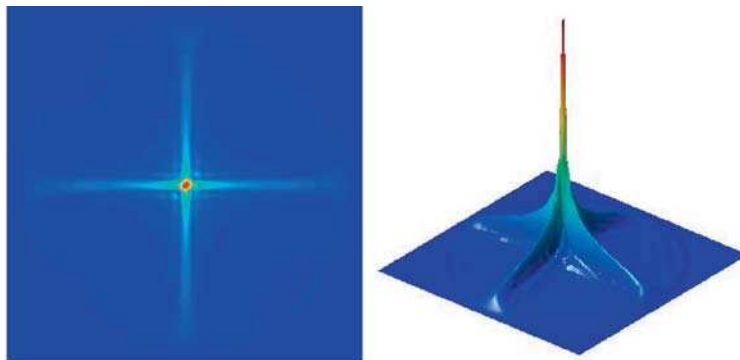


Figure 25-36: Point-spread function and modulation transfer function of a system with a cubic phase plate.

If the phase function in only one dimension is described by only one simple polynomial term in the form

$$\Phi(x_p) = B \cdot x_p^\alpha \tag{25-59}$$

with the coefficient α , then using the method of stationary phases, one gets the ambiguity function

$$A(v, v \cdot \tan \varphi) = \sqrt{\frac{\pi}{12|\alpha \cdot v|}} \cdot e^{i \left(\frac{\alpha \cdot v^3}{4} - \frac{\pi^2 v^2 \tan^2 \varphi}{3\alpha v} \right)}. \tag{25-60}$$

It can be seen that, for large values of the parameter α , this function can be approximated by

$$A(v, v \cdot \tan \varphi) \approx \sqrt{\frac{\pi}{12|\alpha \cdot v|}} \cdot e^{i \frac{\alpha \cdot v^3}{4}} \tag{25-61}$$

and therefore it has no dependence on the defocus parameter φ . Figure 25-37 shows the change in the form of the absolute value of the ambiguity function with increasing parameter, α . It can be seen that, for large α , the form of the distribution supports sufficiently large values of the frequency and large defocussing, which is proportional to a rotation angle around the origin, see figure 25-16 for an explanation.

The analytical calculation for this simple one-dimensional case shows that the invariance of the point-spread function on the defocus is only approximately valid and that, in reality, there are some effects which deteriorate the expected ideal performance. But the advantage of the method in real imaging can be seen in figure 25-38. There, a system is shown with an object, having a large depth so that in the conventional image only some details are seen sharply. If the cubic phase mask is inserted, the primary image is much worse, but the deconvolution delivers an image, which shows the details of the object very sharp in nearly all planes. Figure 25-39 shows the same effect with the help of a simulation. Here, the advantage of the system with the mask in comparison to conventional imaging can be seen very clearly. In the right-hand side picture of the first row of this figure, the detected primary picture with the mask shows a significant loss of symmetry. This is one drawback of the cubic mask, with its odd phase function. The point-spread function is broadened extremely anisotropically. This corresponds to an offset in the image and causes a jump in the image location. It should be noticed that, for this example cal-

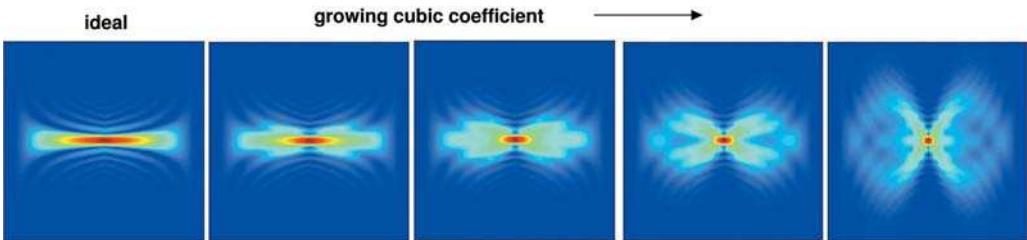


Figure 25-37: Development of the form of the ambiguity function of a system with cubic phase plate and increasing value of the parameter α .

culution, a circular pupil is assumed. This causes some significant differences in the form of the distribution of the point-spread function in the centre region, in comparison with the square pupil of figure 25-35.

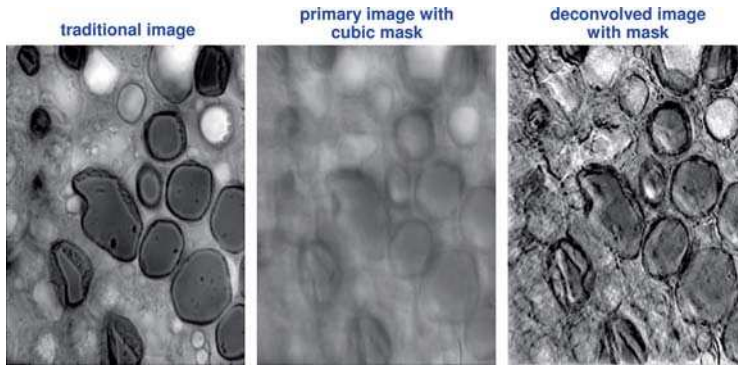


Figure 25-38: Experimental demonstration of the influence of the cubic phase plate with and without deconvolution in comparison with traditional image formation [25-35].

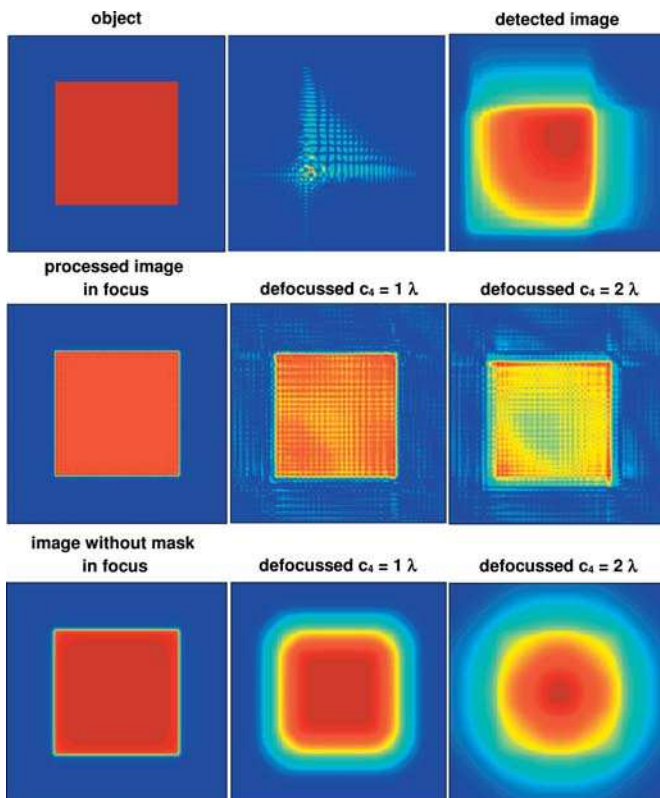


Figure 25-39: Simulated demonstration of the influence of the cubic phase plate with and without deconvolution in comparison with traditional image formation.

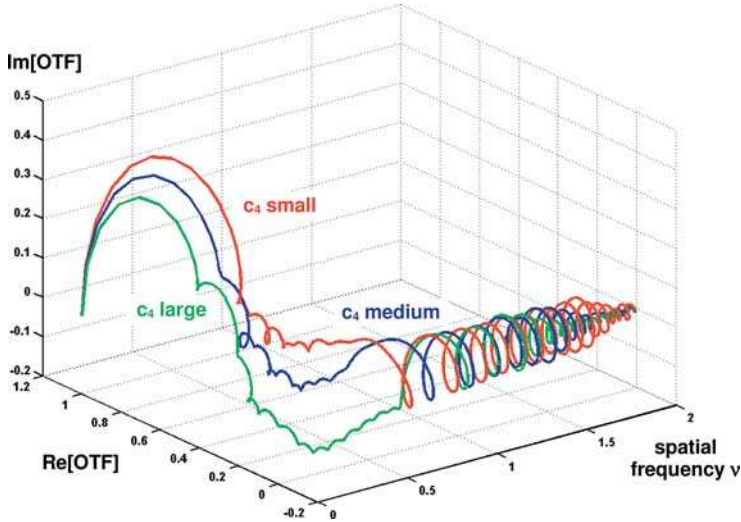


Figure 25-40: Real and imaginary part of the optical transfer function for a system with a cubic phase mask of constant strength and different defocussing coefficients c_4 .

If an optical system containing a cubic phase plate is defocussed, the phase function is composed of the two terms

$$\Phi(x_p) = \alpha \cdot x_p^3 - c_4 \cdot (2x_p^2 - 1) \quad (25-62)$$

where the second term describes the defocussing and is formulated in the usual Zernike convention. In an interesting case, these two function compensate each other. The modulation takes a nearly constant value as a function of the defocussing, but the complex optical transfer function oscillates very quickly. This corresponds to a large variation in the phase transfer function and causes the lateral offset discussed above. Figure 25-40 shows a representation of the complete complex behaviour of the transfer function in a 3D-diagram. The complex vector of the optical transfer function rotates around the axis, its absolute length remains nearly constant, which corresponds to an invariant contrast. Figure 25-41 shows another representation, where on the left side, the modulation is shown with a growing defocussing parameter c_4 and on the right side, only the real part is projected onto the axis. This discussion shows that the phase of the optical transfer depends on the value of the defocus. Therefore, the object details with larger frequency components cannot be reconstructed correctly. This causes a loss of resolution and imaging artefacts, which occur for larger defocussing values. For systems which are not diffraction-limited, this may not be a serious problem, but for microscopical applications, the imaging shows these disadvantages.

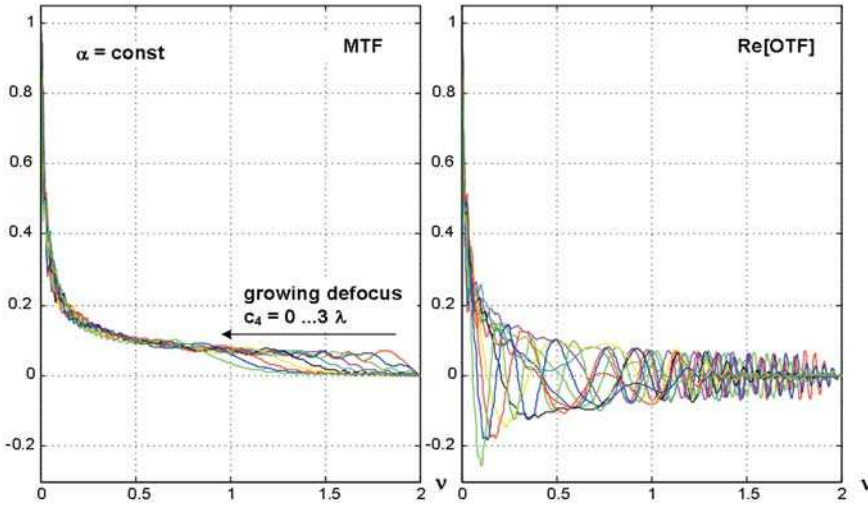


Figure 25-41: Absolute value and real part of the optical transfer function for a system with a cubic phase mask of constant strength and growing defocussing coefficient c_4 .

25.5.7

Structured Illumination

The principle of multiplexing in the frequency domain can be used to enlarge the resolution of an imaging system. If the object is illuminated with a grating-typed structured illumination of the form [25-36], [25-37], [25-38]

$$I_L(x, y) = 1 + \Delta I \cdot \cos(\nu_o \cdot x + \varphi_o) \tag{25-63}$$

with a grating frequency ν , this carrier frequency transforms the object spectrum in the Fourier space to both frequency arms $+\nu_o$ and $-\nu_o$. This is shown in figure 25-42.

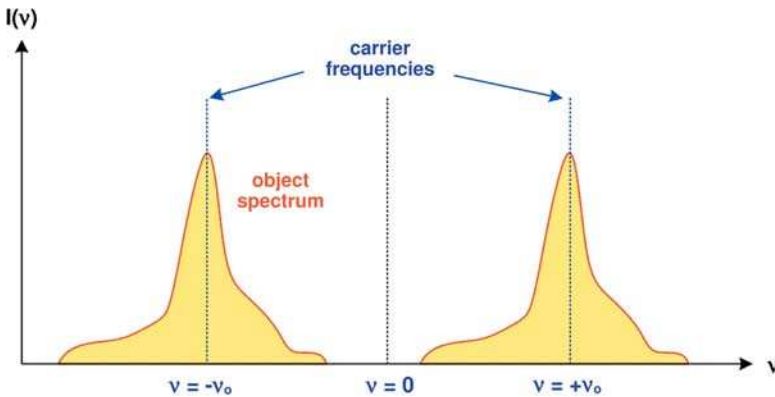


Figure 25-42: Principle of the carrier frequencies and the shift of the object spectrum.

If the detection of the image is filtered, the carrier frequency can be removed and the object is reconstructed. If the projection of the grating is coherent and the imaging is incoherent, the imaging process can be described in one dimension by the equation

$$I(x') = \int I_L(x_L) \cdot \left| \int g_L(x_L + x_o) \cdot t(x_o) \cdot g_{psf}(x + x_o) dx_o \right|^2 dx_L \quad (25-64)$$

where t is the object transmission function, g_L is the coherent transfer function of the illumination system and g_{psf} is the amplitude point-spread function of the imaging system. If a sinusoidal grating is assumed, this equation gives the image signal

$$I(x') = I_o(x') + I_s(x') \cdot \sin \varphi_o + I_c(x') \cdot \cos \varphi_o. \quad (25-65)$$

The information of the object is coded in this distribution and is contained in the expression

$$I_{\text{image}} = \sqrt{I_s^2 + I_c^2}. \quad (25-66)$$

The extraction of the desired information is possible, if the imaging process is performed for several values of the illumination phase angle φ_o . A post-processing then eliminates the angle. If, for example, the angle takes the three values

$$\varphi_o = 0 / \frac{2\pi}{3} / \frac{4\pi}{3} \quad (25-67)$$

which correspond to the generation of three images with a slightly shifted grating, the three images can be used to calculate the true image with the help of the equation

$$I_{\text{image}} = \sqrt{(I_1 - I_2)^2 + (I_1 - I_3)^2 + (I_2 - I_3)^2}. \quad (25-68)$$

Figure 25-43 shows a calculated example for resolution enhancement using a structured illumination in the described manner. A single image for one concrete value of φ_o is already coded by the illumination. The imaging without the grating illumination cannot resolve the object grating, which is rotated by a small amount.

The grating illumination causes a multiplexing of the frequencies in the Fourier space. In the extreme case, the resolution can be enlarged by a factor of two. This is shown in figure 25-44 and occurs if the directions of the frequency vectors of the grating and the biggest object frequency are opposed.

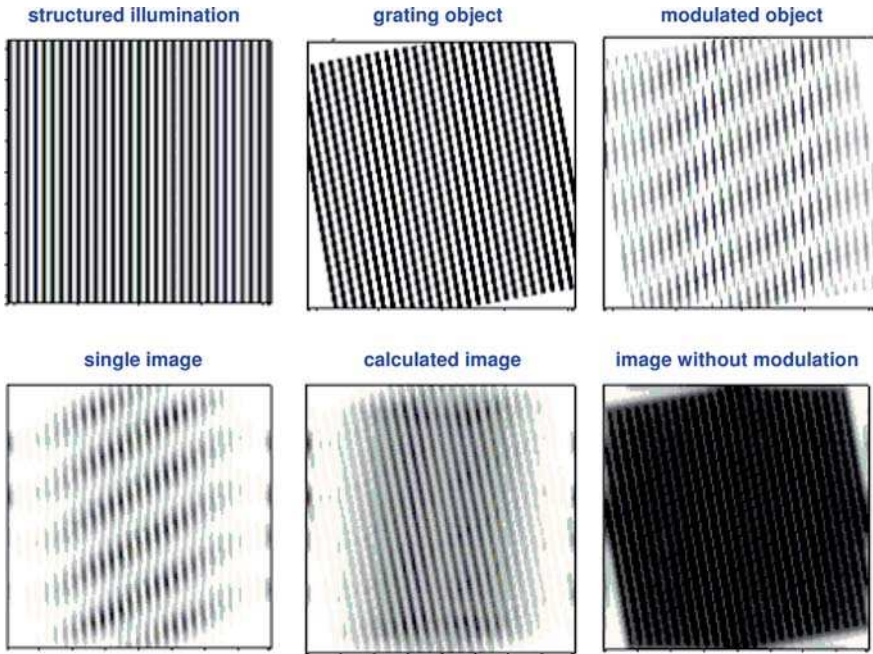


Figure 25-43: Illustration of the gain in resolution for the imaging of a grating with a period near the limiting frequency, by structured illumination, and post-processing by calculation.

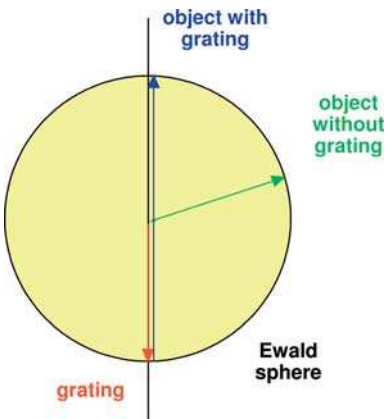


Figure 25-44: Extreme case of the grating projection with a doubling of the resolution.

To understand the method in the spatial domain, figure 25-45 shows the generation of the diffraction orders in the object plane. The principle resembles the resolution enhancement in microscopy due to a skew illumination (see the next chapter). Here the skew direction comes from the first diffraction order of the illumination grating.

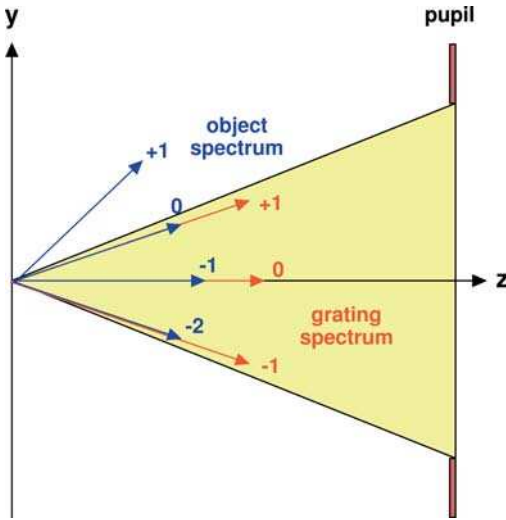


Figure 25-45: Illustration of the enhanced resolution of the grating projection in the spatial domain.

25.6 Selected Practical Applications for Pupil Filtering Techniques

25.6.1 Phase Contrast Filtering, Dark-field Illumination

The images of phase objects are typically hardly visible in optical imaging. With the square of the transmission function

$$|T(x)|^2 = |Ae^{i\varphi(x)}|^2 = A^2 \tag{25-69}$$

in the incoherent image, according to eq. (21-56), phase structures cannot be resolved at all. However, even weak phase objects can be made visible by pupil filtering. The transmission function weak phase object can be expanded into a power series:

$$T(x) = Ae^{i\varphi(x)} = A + iA\varphi(x) + \dots \tag{25-70}$$

The object spectrum is then given by

$$t(\nu) = A\delta(0) + iAF[\varphi](\nu) . \tag{25-71}$$

The spectrum thus consists of a dominating real part, given by the “zero order”, and an imaginary part, given by the Fourier transform of the phase $\varphi(x)$. The phase function can thus be made visible after suppressing the zero order by a pupil filter. After convolution with the effective light source, the first term leads to an identical

image of the light source in the entrance pupil – as illustrated in figure 25-46 for illumination with a small axial light source. As a consequence, complementary source-pupil masks are applied as phase-contrast pupil filters [25-39], [25-40]. The direct part from the light source can thus be effectively filtered out by the application of a pupil filter of the same shape as the light source. The image intensity is then given by

$$I(x) = A^2|\varphi(x)|^2. \quad (25-72)$$

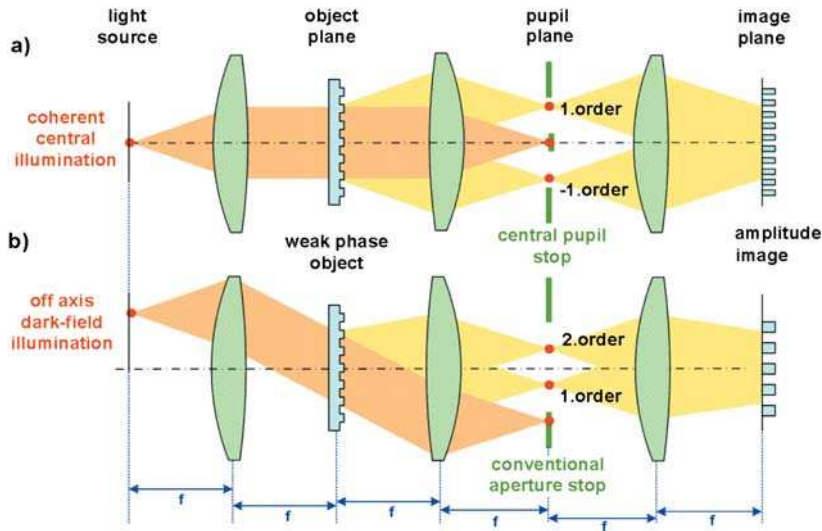


Figure 25-46: Phase contrast imaging for imaging of weak phase objects: a) with central pupil obscuration; and b) for dark-field illumination.

Since the contrast enhancement is achieved by filtering the zero diffraction order or image of the effective light source, the same effect can alternatively be achieved by dark-field illumination, i.e., when the effective light source image is outside of the entrance pupil (figure 25-46b) [25-41]. As consequence, the zeroth diffraction order cannot pass the lens and only higher diffraction orders contribute to the image. Dark-field illumination in combination with pupil filters was proposed in order to produce colour effects in microscopic images when, e.g., the background light passes a different colour section of a pupil filter than does the scattered light of the object (Rheinberg Differential Colour Illumination [25-42]).

For illustration, figure 25-47 shows an x - z scan through the amplitude distribution of the image of a rectangular phase object of 3λ diameter with strong phase ($\varphi = \pi$) and weak phase ($\varphi = \pi/10$), obtained in the focal region of a ideal system of $NA = 0.4$. The strong phase shift is easy to be recognized, while the weak phase shift causes only a weak disturbance in the wavefronts.

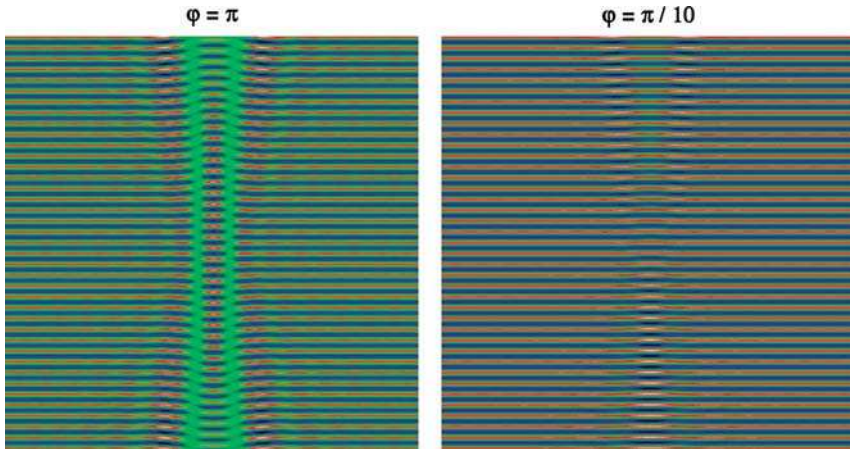


Figure 25-47: Amplitude distribution (x - z scan) of the image of phase structures of phase φ .

In figure 25-48, the simulation results for the image formation of the phase structure are compared for coherent illumination, conventional bright-field illumination with $\sigma=0.7$ and annular dark-field illumination with $\sigma_{\text{in}}=1.1$ and $\sigma_{\text{out}}=1.3$ (at $NA=0.4$). While, for the strong phase shifts, the contrast for dark-field and bright-field is equal in both cases $V \sim 90\%$, the weak phase shifter is not resolved in the conventional illumination settings with $\sigma=0.05$ and $\sigma=0.7$, but imaged with again a high contrast of $V \sim 90\%$ by dark-field illumination.

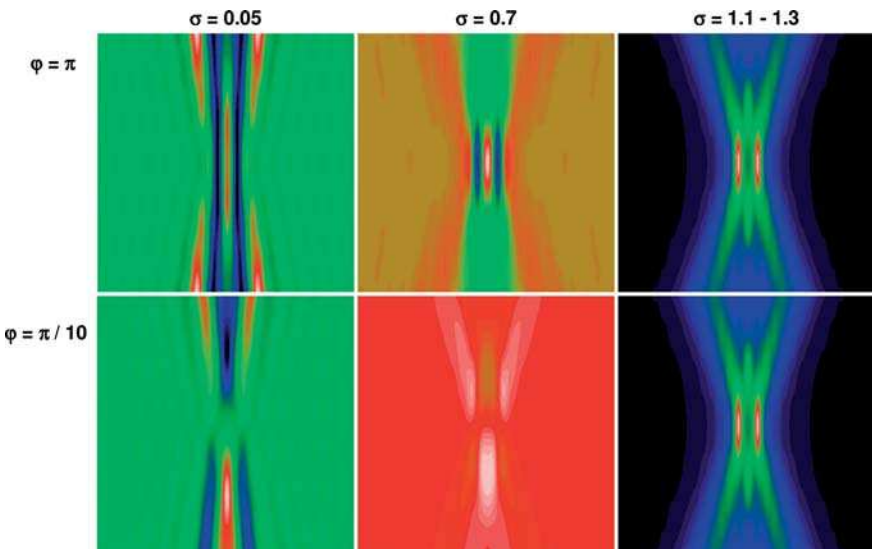


Figure 25-48: Comparison of the images (x - z scans) of a strong and a weak phase shifter with coherent, incoherent conventional and annular dark-field illumination.

According to Zernike, instead of an opaque filter to block the zero order, a phase filter of the shape of the light source and phase given by an odd multiple of $\pi/2$ is applied [25-43]. One obtains finally for the transmitted spectrum

$$t'(v) = h(v) \cdot t(v) = A\delta(0)e^{i\frac{(2n+1)}{2}\pi} + AF[e^{i\varphi(x)}]. \quad (25-73)$$

For the intensity it follows

$$I(x) = \left| Ae^{i\left[\frac{\pi}{2} + \varphi(x)\right]} \right|^2 \approx A^2 |i[1 + \varphi(x)]|^2 \approx 1 + 2 \cdot \varphi(x). \quad (25-74)$$

In contrast to eq. (25-72), the phase function now influences the intensity linear.

25.6.2

Frequency Doubling

Diffraction gratings show typically strong 0th and 1st diffraction orders, while, e.g., the 2nd diffraction orders vanish. With the image period given by the minimum distance of diffraction orders in the entrance pupil, application of a filter to block the zero diffraction order may lead to a grating image of doubled frequency:

$$I(x) \sim \cos\left(2\pi \frac{2}{d}x\right). \quad (25-75)$$

Consequently, a central pupil obscuration may lead to a doubled image frequency, as long as both of the 1st diffraction orders pass the lens aperture. For frequency doubling, phase gratings can also be applied, when the phase of the grating is selected appropriately to cancel the zero diffraction order and only the two 1st diffraction orders contribute to the image. Frequency doubling by phase masks or pupil filtering has been proposed for optical lithography in order to reduce the expenses for lithography masks [25-44], [25-45], [25-46].

In general, images of phase objects have low contrast for incoherent illumination. With phase contrast imaging, the contrast can be enhanced [25-41]. The effect of phase contrast filtering is illustrated for the example of annular dark-field illumination for the imaging of a phase grating of four periods with period $d = 3\lambda$. Figure 25-49 illustrates, for the amplitude distribution through the focus, the phase shift according to a strong phase grating with $\varphi = \pi$ and a weak phase grating with phase $\varphi = \pi/5$. Figure 25-50 illustrates the images under different illumination conditions. All resolved images exhibit frequency doubling. Both gratings are imaged with identical contrast under annular dark-field illumination ($V \sim 59\%$). Figure 25-51 illustrates the 3D-transfer function for the illumination setting and the transmitted grating spectrum. With dark-field illumination, the spectrum read-out is laterally almost invariant and thus for thin objects many images, which are almost identical, are superposed.

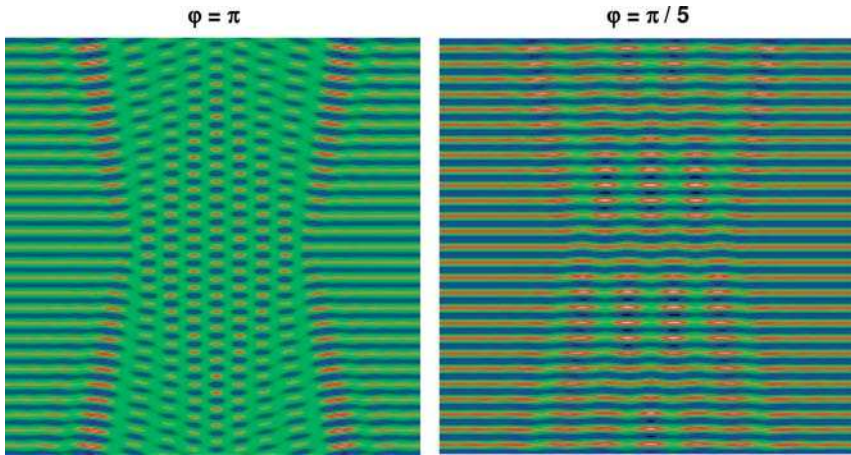


Figure 25-49: Amplitude distribution of on-axis illuminated phase grating of period $d = 3\lambda$ ($NA = 0.8$).

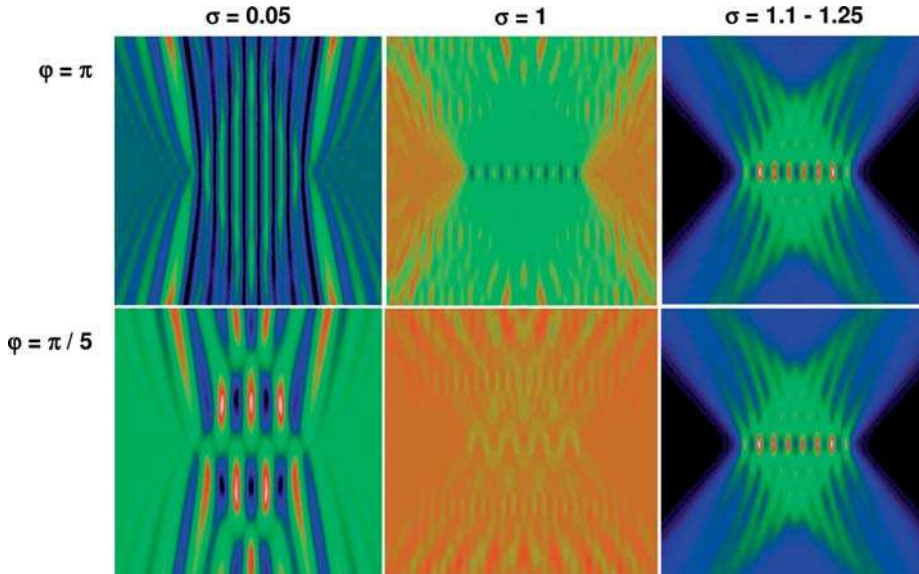


Figure 25-50: Intensity distribution of a phase grating with four periods of period $d = 3\lambda$ ($NA = 0.8$).

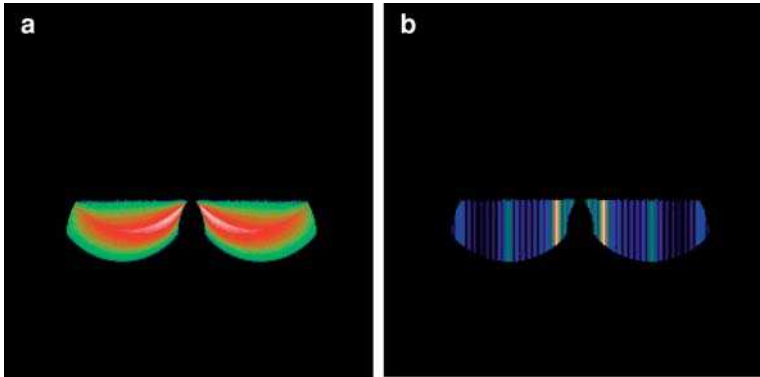


Figure 25-51: a) Dark-field 3D MTF for $NA = 0.8$ and $\sigma_{in} = 1.1 - \sigma_{out} = 1.25$; b) transmitted frequency spectrum of a phase grating with period 3λ .

25.6.3

Defect Filtering

The transmission function of a deteriorated periodic object can be given as a bilinear superposition of the ideal object transmission and the transmission function of the error:

$$T(x) = T_{ideal}(x) + T_{error}(x). \tag{25-76}$$

To detect the errors, it is advantageous to block the diffraction orders of the ideal object by a pupil filter, such that

$$u_1(v) = h(v) \cdot [t_{ideal}(v) + t_{error}(v)] = t_{error}(v). \tag{25-77}$$

with

$$h(v) = \begin{cases} 1 & t_{ideal}(v) = 0 \\ 0 & t_{ideal}(v) \neq 0. \end{cases} \tag{25-78}$$

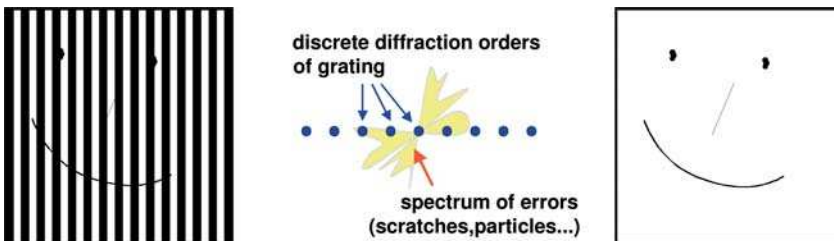


Figure 25-52: Defect detection by spatial filtering technique: Grating with errors, frequency spectrum and filtered image.

This method is especially applicable for defect analysis in periodic objects like gratings (see figure 25-52 for an illustration) [25-47]. As pupil filters, spatial modulators and non-linear crystals have been proposed. In general, difference filters in the pupil plane can be applied to suppress or enhance certain object details.

25.6.4

Ronchi Test

The Ronchi test was introduced for the metrology of the aberrations of optical systems and is also applied to the metrology of optical surfaces [25-48], [25-49]. Here it is described as a special example of image multiplexing for the imaging of phase objects [25-50]. To simplify the discussion, an ideal system is assumed with transmission function $h(\nu) = 1$.

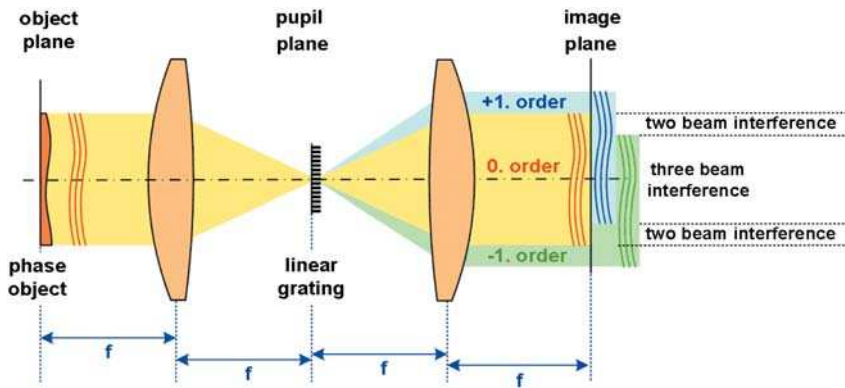


Figure 25-53: Setup of a Ronchi Test for an imaging optical system.

The transmission function of a phase object is given by

$$T(x) = A(x) \cdot e^{i\phi(x)} \tag{25-79}$$

with $A(x) \approx 1$. After multiplication to the illumination distribution $S(x)$ and Fourier transformation, the distribution in the entrance pupil is obtained:

$$u_0(\nu) = t(\nu) \otimes s(\nu). \tag{25-80}$$

The field distribution is then multiplied to the filter function of the Ronchi grating:

$$u_1(\nu) = u_0(\nu) \cdot \text{ronchi}\left(\frac{\nu}{d}\right). \tag{25-81}$$

In the image plane, we finally obtain

$$U_1(x) = \frac{d}{2} \cdot U_0(x) \otimes \text{comb}(d \cdot x) \cdot \text{sinc}\left(\frac{d}{2} \cdot x\right). \tag{25-82}$$

After expansion of the weighted comb function we obtain

$$U_1(x) = \sum_m c_m \cdot U_0\left(x - \frac{m}{d}\right) \quad (25-83)$$

with the diffraction amplitudes $c_m = 0.5 \cdot \text{sinc}(m/2)$. The first diffraction amplitudes c_m and efficiencies η_m are summarized in table 25-4.

Table 25-4: Diffraction amplitudes and efficiencies of a Ronchi grating.

	Order m	0	1	2	3
Amplitude	Amplitude c_m	0.5	$1/\pi=0.32$	0.0	$-1/3\pi=-0.11$
	Efficiency η_m	25 %	10 %	0 %	1 %

The image intensity is thus given by a series expansion:

$$I(x) = \left| \sum_m c_m \cdot U_0\left(x - \frac{m}{d}\right) \right|^2 = \sum_{m,n} c_m c_n^* \cdot U_0\left(x - \frac{m}{d}\right) \cdot U_0^*\left(x - \frac{n}{d}\right). \quad (25-84)$$

With $k = m - n$ we obtain

$$I(x) = \sum_{k=0}^N i_k(x). \quad (25-85)$$

After insertion of the transmission function and the coherence function, Γ_S , $i_k(x)$ can be written as

$$i_0(x) = \sum_m \left| c_m \cdot T\left(x - \frac{m}{d}\right) \right|^2 \cdot \Gamma_S\left(x - \frac{m}{d}, x - \frac{m}{d}\right), \quad (25-86)$$

$$i_k(x) = 2 \sum_m \text{Re} \left\{ c_m c_{m-k}^* \cdot \Gamma_S\left(x - \frac{m}{d}, x - \frac{m-k}{d}\right) \cdot T\left(x - \frac{m}{d}\right) \cdot T^*\left(x - \frac{m-k}{d}\right) \right\}. \quad (25-87)$$

For an incoherent source, the coherence function Γ_S according to eq. (21-37) is given by the Fourier transform of the intensity of the effective light source:

$$\Gamma_S(x_1, x_2) = S(x_1) \cdot S^*(x_2) = \int |s(\nu_q)|^2 \cdot e^{i2\pi\nu_q(x_1-x_2)} d\nu_q \quad (25-88)$$

As a consequence, the coherence function $\Gamma_S(x_1, x_2)$ is constant for $x_1 = x_2$ and eqs. (25-86), (25-87) can be simplified to

$$i_0(x) = \Gamma_S(0, 0) \cdot \sum_m \left| c_m \cdot T\left(x - \frac{m}{d}\right) \right|^2, \quad (25-89)$$

$$i_k(x) = 2 \cdot \Gamma_S\left(0, \frac{k}{d}\right) \cdot \sum_m \operatorname{Re} \left\{ c_m c_{m-k}^* \cdot T\left(x - \frac{m}{d}\right) \cdot T^*\left(x - \frac{m-k}{d}\right) \right\}. \quad (25-90)$$

For phase objects, the image intensity is thus given by a constant background intensity i_0 superposed by multiple interference patterns formed by shifted images of the object. With amplitude $A = 1$ and real diffraction efficiencies c_m of the grating, we obtain

$$\begin{aligned} i_k(x) &= 2 \cdot \Gamma_S\left(0, \frac{k}{d}\right) \\ &\cdot \sum_m \operatorname{Re} \left\{ c_m c_{m-k}^* \cdot A\left(x - \frac{m}{d}\right) \cdot A^*\left(x - \frac{m-k}{d}\right) \cdot e^{i[\varphi(x-\frac{m}{d}) - \varphi(x-\frac{m-k}{d})]} \right\} \\ &= 2 \cdot \Gamma_S\left(0, \frac{k}{d}\right) \cdot \sum_m c_m c_{m-k} \cdot \cos \left[\varphi\left(x - \frac{m}{d}\right) - \varphi\left(x - \frac{m-k}{d}\right) \right]. \end{aligned} \quad (25-91)$$

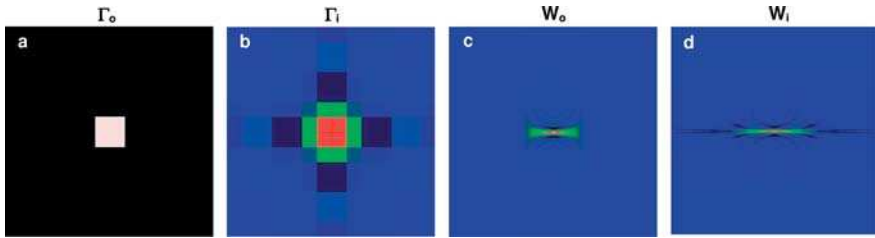


Figure 25-54: Coherence and Wigner distribution function of a Ronchi Test with grating period $d = \lambda/NA$ for an ideal optical system for coherent illumination. Coherence function in: a) the object; and b) the image plane, and Wigner distribution function in: c) the object plane; and d) the image plane.

The intensity i_k is thus given by a superposition of interference patterns equivalent to a shearing interferometer with a lateral shear of k/d . Figure 25-54 illustrates the Ronchi test for the coherence transfer function and the Wigner distribution function for the example of a constant phase object of finite width. The image intensity is given by the diagonal elements of the coherence function, $I(x) = \Gamma_i(x, x)$. The coherence function in the image plane $\Gamma_i(x_1, x_2)$, shown in figure 25-54b, is formed by a superposition of several copies of the coherence function of the object, shown in figure 25-54a. The copies of the object coherence function are shifted in accordance with the Ronchi grating period. As can be seen in figure 25-54b, with the Ronchi grating period $d = \lambda/NA$ the first diffraction order overlaps 50% with the zero order. For a phase object with constant phase $\varphi(x) = \varphi_0$, the interference pattern i_k is also constant and the image intensity $I(x)$ depends only on the number of diffraction orders. The Wigner distribution function WDF of the amplitude distribu-

tion in the image plane, shown in 25-54d, is given by a periodic repetition of the WDF in the object plane, shown in 22-54c. According to eq. (25-91), the interference pattern in the image plane is formed by a superposition of the phase functions $\varphi(x)$ of the object transmission function, shifted by multiples of $1/d$. As a consequence, object frequencies given by multiples of the frequency $1/d$ are not resolved. The Wigner distribution function in the image plane, shown in figure 25-54d, vanishes for multiples of the shear frequency $1/d$.

Figure 25-55 compares the Ronchi test with the Ronchi grating period $d = 8\lambda/NA$ for the example of an ideal phase object and a phase object given by a phase distribution $\sim x^3$ as described by coma aberration (Z_7).

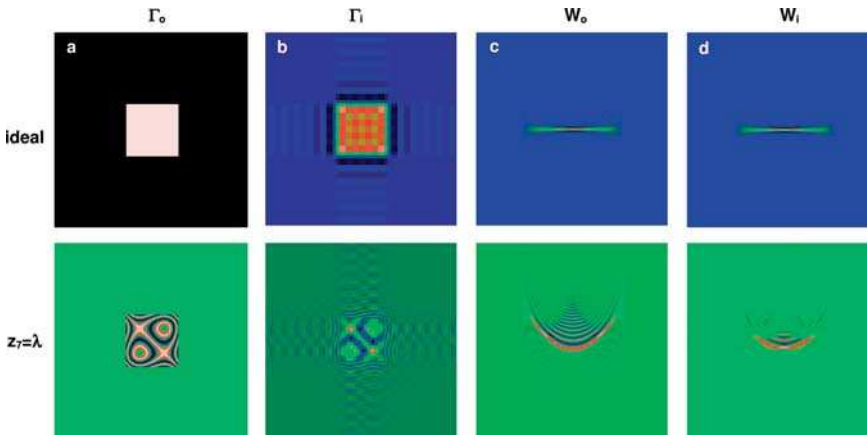


Figure 25-55: Coherence and Wigner distribution function of a Ronchi Test with grating period $d = 8\lambda/NA$ for an ideal optical system and with coma ($Z_7 = \lambda$) for coherent illumination. Coherence function in: (a) the object; and (b) the image plane, and Wigner distribution function in: (c) the object; and (d) the image plane.

The visibility of the interference i_k of order k depends on the coherence function $\Gamma_S(0, k/d)$. After consideration of the zeroth and first diffraction orders only, we obtain

$$i_0(x) = \Gamma_S(0, 0) \cdot \left[c_0^2 I_0(x) + c_1^2 I_0\left(x - \frac{1}{d}\right) + c_{-1}^2 I_0\left(x + \frac{1}{d}\right) \right], \quad (25-92)$$

$$i_1(x) = 2 \cdot \Gamma_S\left(0, \frac{1}{d}\right) \cdot \operatorname{Re} \left\{ c_1 c_0^* \cdot T\left(x - \frac{1}{d}\right) \cdot T^*(x) + c_0 c_{-1}^* \cdot T(x) \cdot T^*\left(x + \frac{1}{d}\right) \right\}, \quad (25-93)$$

$$i_2(x) = 2 \cdot \Gamma_S\left(0, \frac{2}{d}\right) \cdot \operatorname{Re} \left\{ c_1 c_{-1}^* \cdot T\left(x - \frac{1}{d}\right) \cdot T^*\left(x + \frac{1}{d}\right) \right\}. \quad (25-94)$$

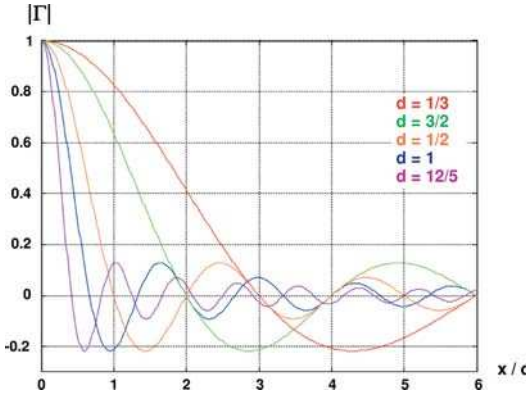


Figure 25-56: Coherence function with a source $s(\nu)$ given by rect functions with different widths.

The visibility of the interference patterns i_k is thus determined by the coherence function at k/d . As shown in figure 25-56, with different sizes, and incoherent source distributions described by a rect function, different interference patterns of order k can be selected. For an effective light source given by a rect function of width $d/2$ with

$$s(\nu) = \text{rect}\left(\frac{2 \cdot \nu}{d}\right) \tag{25-95}$$

the normalized coherence function is given by

$$\Gamma_s(0, x) = \text{sinc}\left(\frac{d}{2}x\right). \tag{25-96}$$

As a consequence, $\Gamma_s(0, k/d)$ vanishes for even $k > 0$. Since the coherence between point distances $2/d$ vanishes, and the interference $i_2(x)$ is dampened. After insertion of the transmission function and for a source distribution of width $d/2$ it follows for the interference pattern $i_1(x)$ in the image plane that

$$i_1(x) = C \cdot \left\{ \cos \left[\varphi(x) - \varphi\left(x - \frac{1}{d}\right) \right] + \cos \left[\varphi\left(x + \frac{1}{d}\right) - \varphi(x) \right] \right\}. \tag{25-97}$$

The interference pattern of first order is thus equivalent to a shearing interferometer with lateral shear of $\pm 1/d$. Examples of interference patterns for phase functions given by different Zernike aberrations are illustrated in figure 16-49. Figure 25-57 illustrates the impact of an incoherent source with extension $d_i = d/2$ in comparison to coherent illumination for the example illustrated in figure 25-55.

As can be shown, the second-order interference pattern is also vanishing for effective light sources equivalent to the Ronchi grating applied as a pupil filter in eq. (25-81) [25-51].

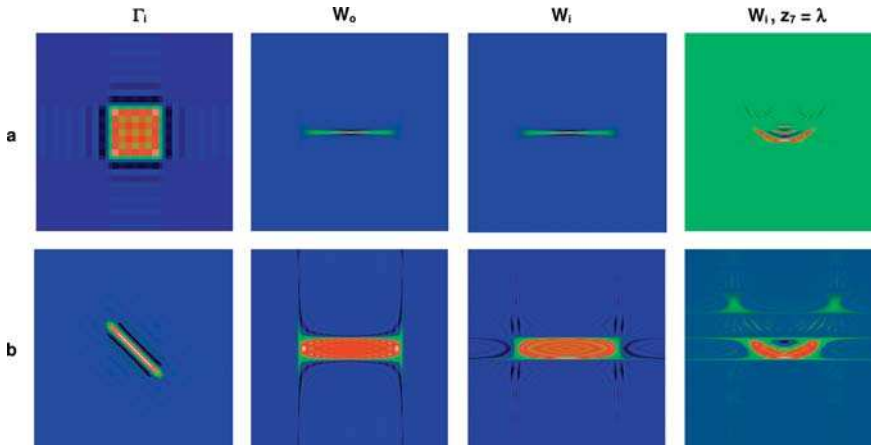


Figure 25-57: Coherence and Wigner distribution function of a Ronchi Test with grating period $d = 8\lambda/NA$ for an ideal optical system and with coma ($Z_7 = \lambda$) for different extensions d_i of the light source: a) coherent illumination; and b) $d_i = d/2$.

The Ronchi test is frequently applied to the metrology of optical systems. The source is placed in the object plane of the optical system and the grating in the image plane (see figure 25-58). With an additional lens, e.g. that of a microscope, the pupil plane is image on a CCD camera, where the interferogram is recorded.

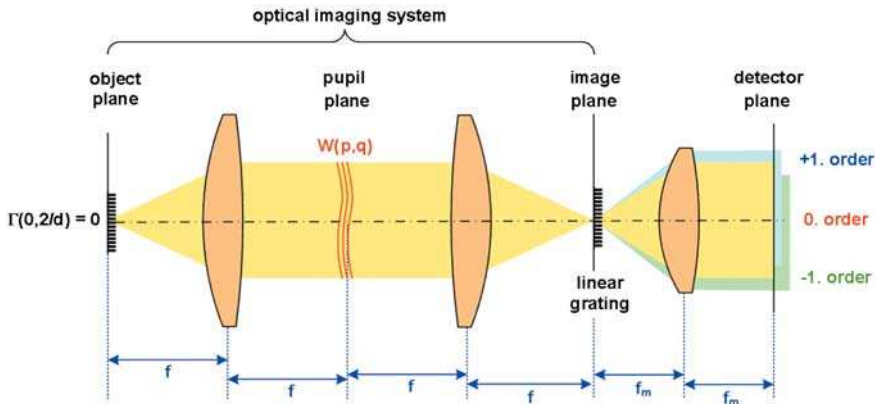


Figure 25-58: Ronchi aberration test of an optical system with grating as source distribution.

Figures 25-59 and 25-60 illustrate the Ronchi test of a test setup according figure 25-58 for the Wigner distribution function. The Wigner distribution function in the exit pupil W_e is compared with the W_d in the detector plane for the example of distortion (Z_2) and coma (Z_7) of different magnitudes, respectively.

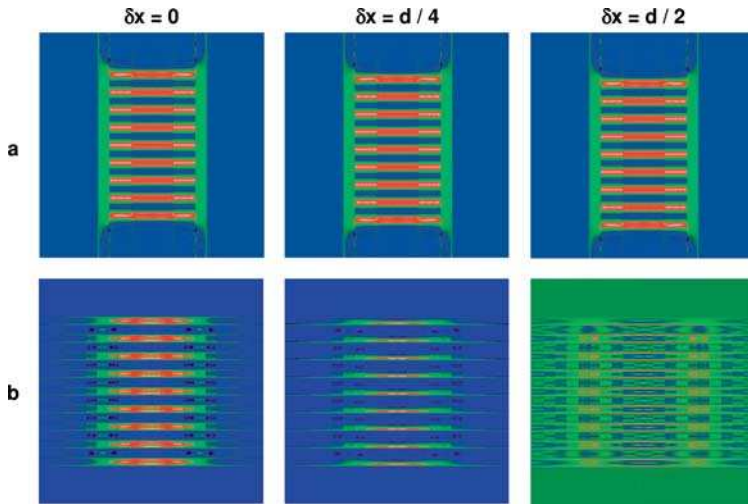


Figure 25-59: Wigner distribution function in the exit pupil (a) and the detector plane (b) for different amounts of distortion $\delta x \sim Z_2/NA$.

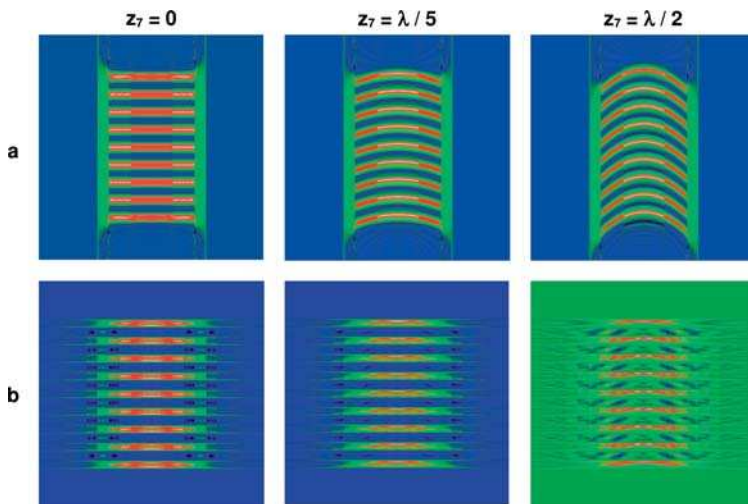


Figure 25-60: Wigner distribution function in the exit pupil a) and the detector plane b) for different amounts of distortion Z_7 .

In figure 25-61 Ronchigrams for three simple aberration types are shown for illustration.

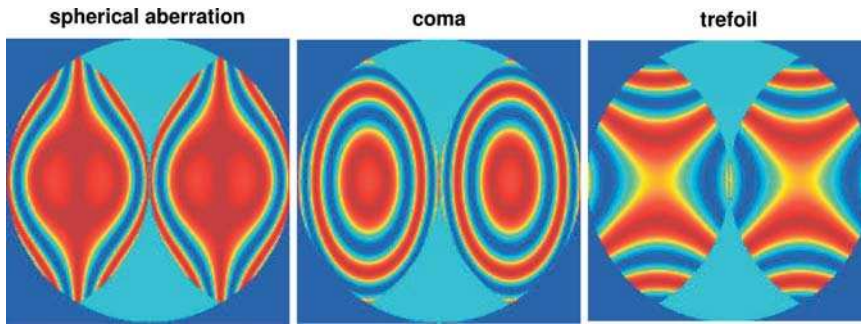


Figure 25-61: Ronchigrams for some simple aberrations.

25.7

Literature

- 25-1 W.T. Welford, *J. Opt. Soc. Am.* **50**, 749 (1960), Use of annular apertures to increase depth of focus.
- 25-2 S.T. Yang, R.L. Hsieh, Y.H. Lee, R.F.W. Pease, G. Owen, *SPIE* **1264**, 477 (1990), Effect of central obscuration on image formation in projection lithography.
- 25-3 G.Toraldo di Francia, *Nuovo Cimento* **9**, Suppl., 426 (1952), Super-gain antennas and optical resolving power.
- 25-4 S. Fujiwara, *J. Opt. Soc. Am.* **52**, 287 (1962), Optical properties of conic surfaces.I. Reflecting cone.
- 25-5 T. Asakura and S. Nagai, *Jpn. J. Appl. Phys.* **10**, 879 (1971), Further studies of far-field diffraction by modified annular and annulus apertures.
- 25-6 H.F.A. Tschunko, *Appl. Opt.* **18**, 3770 (1979), Imaging performance of annular apertures. 3: Apodization and modulation transfer functions.
- 25-7 Z.S. Hegedus, V. Sarafis, *J. Opt. Soc. Am.* **A 3**, 1892 (1986), Super resolving filters in confocally scanned imaging systems.
- 25-8 R. Arimoto et al., *Appl. Opt.* **31**, 6653 (1992), Imaging properties of axicon in a scanning optical system
- 25-9 Ch. Hofmann, *Die optische Abbildung* (Akademische Verlagsgesellschaft, Leipzig, 1980)
- 25-10 D. A. Holmes, J. E. Korka and P. V. Avizonis, *Appl. Opt.* **11**, 565 (1972), Parametric study of apertured focused Gaussian beams.
- 25-11 E. Marom, B. Chen and O. G. Ramer, *Opt. Eng.* **18**, 79 (1979), Spot size of focused truncated Gaussian beams.
- 25-12 R. Jozwicki, *Optica Acta* **30**, 1011 (1983), Parametric analysis of truncated Gaussian beams.
- 25-13 Melles Griot Optics Guide, www.mellesgriot.com
- 25-14 M. Born and E. Wolf, *Principles of Optics* (Cambridge University Press, Cambridge, 1999).
- 25-15 P. Török and F.-J. Kao, *Optical Imaging and Microscopy* (Springer, Berlin, 2003), chapters 4 and 5.
- 25-16 M. Yun, L. Liu, J. Sun and De'an Liu, *J. Opt. Soc. Am.* **A 22**, 272 (2005), Three-dimensional super resolution by three-zone complex pupil filters.
- 25-17 D. M. de Juana, J. E. Oti, V. F. Canales and M. P. Cagigal, *Opt. Lett* **28**, 607 (2003), Design of super resolving continuous phase filters
- 25-18 K.-H. Brenner, A W. Lohmann and J. Ojeda-Castaneda, *Opt. Commun.* **44**, 323(1983), The ambiguity function as a polar display of the OTF.
- 25-19 H. Dammann and K. Görtler, *Opt. Commun.* **3**, 312 (1971), High-efficient in-line multiple imaging by means of multiple phase holograms.
- 25-20 G. Häusler, *Opt. Commun.* **6**, 38 (1972), A method to increase the depth of focus by two-step image processing.

- 25-21 H. Fukuda, T. Terasawa and S. Okazaki, *J. Vac. Sci. Technol.* **B 9**, 3113 (1991), Spatial filtering for depth of focus and resolution enhancement in optical lithography.
- 25-22 R. Hild, M.J. Yzuel, J.C. Escalera and J. Campos, *Opt. Eng.* **37**, 1353 (1998), Influence of non-uniform pupils in imaging periodical structures by photolithographic systems.
- 25-23 T. R. Sales and G. M. Morris, *Opt. Lett.* **22**, 582 (1997), Fundamental limits of optical super resolution.
- 25-24 T. R. Sales, *Phys. Rev. Lett.* **81**, 3844 (1998), Smallest focal spot.
- 25-25 H. J. Caulfield (Ed.), *Optical Information Processing: A Tribute to Adolf Lohmann* (SPIE Press, Bellingham, 2002), chapter 13
- 25-26 J. R. Fienup, *Appl. Opt.* **21**, 2758 (1982), Phase retrieval algorithms: a comparison.
- 25-27 M. Martinez-Corral, M. T. Caballero, E. H. Stelzer and J. Swoger, *Optics Express* **10**, 98 (2002), Tailoring the axial shape of the point-spread function using the Toraldo concept.
- 25-28 C. J. Shepard, G. Calvert and M. Wheatland, *J. Opt. Soc. Am. A* **15**, 849 (1998), Focal distribution for super resolving Toraldo filters.
- 25-29 H. Wang and F. Gan, *Appl. Opt.* **41**, 5263 (2002), Phase-shifting apodizers for increasing focal depth.
- 25-30 H. Wang and F. Gan, *Appl. Opt.* **40**, 5658 (2001), High focal depth with a pure phase apodizer.
- 25-31 X. Liu, X. Cai and C. P. Grover, *Proc. SPIE* **5174**, 51 (2003), Extending the focal depth of an optical tracking system using a phase pupil plate.
- 25-32 J. Ojeda-Castaneda and L. R. Berriel-Valdos, *Appl. Opt.* **29**, 994 (1990), Zone plate for arbitrarily high focal depth.
- 25-33 S. S. Sherif, W. T. Cathey and E. R. Dowski, *Appl. Opt.* **43**, 2709 (2004), Phase plate to extend the depth of field of incoherent hybrid imaging systems.
- 25-34 W. T. Cathey and E. R. Dowski, *Appl. Opt.* **41**, 6080 (2002), New paradigm for imaging systems.
- 25-35 E. R. Dowski, publication with permission.
- 25-36 V. Solomon, Z. Zalevsky and D. Mendlovic, *Appl. Opt.* **42**, 1451 (2003), Super resolution by use of code division multiplexing.
- 25-37 M. G. Gustafsson, D. A. Agard and J. W. Sedat, *Proc. SPIE* **3919**, 141 (2000), Doubling the lateral resolution of wide-field fluorescence microscopy using structured illumination
- 25-38 W. Lukosz, *J. Opt. Soc. Am.* **56**, 1463 (1966), Optical systems with resolving powers exceeding the classical limit.
- 25-39 Rhodes, M. Koizumi, *Proc. IEEE of the Tenth optical Computing Conference 1983*, 32 (1983), Image enhancement by partially coherent imaging.
- 25-40 G. Indebetouw and Ch. Varamit, *J. Opt. Soc. Am. A* **2**, 794 (1985), Spatial filtering with complementary source-pupil masks.
- 25-41 T. Noda, S. Kawata and S. Minami, *Appl. Opt.* **29**, 3810 (1990), Three-dimensional phase contrast imaging by an annular illumination microscope.
- 25-42 G.H. Needham, *The Practical Use of the Microscope*, (Thomas Publications, 1958).
- 25-43 F. Zernike, *Z. Tech. Phys.* **16**, 454 (1935), Das phasenkontrastverfahren bei der mikroskopischen Beobachtung.
- 25-44 H. Hänsel and W. Polack, *German Democratic Republic Patent No. 126.361* (1976), Verfahren zur Herstellung einer Phasemaske mit Amplitudenstruktur.
- 25-45 M. D. Levenson et al., *IEEE Trans. Electron Devices* **ED-29**, 1828 (1982), Improving resolution in photolithography with a phase-shifting mask.
- 25-46 T. Jewell and D. L. White, *J. Lightwave Technol.* **7** (9), 1386 (1989), Spatial frequency doubling lithography SFDL.
- 25-47 N.N. Axelrod, *US* 3,658,420 (1969), Photomask inspection by spatial filtering.
- 25-48 V. Ronchi, *Appl. Opt.* **3**, 437 (1963), Forty years of history of a grating interferometer.
- 25-49 A. Cornejo, *Appl. Opt.* **9**, 1897 (1970), D. Malacara, Ronchi test of aspherical surfaces, analysis and accuracy.
- 25-50 R. Barakat, *J. Opt. Soc. Am.* **59**, 1432 (1969), General diffraction theory of optical aberration tests, from the point of view of spatial filtering.
- 25-51 A. Cornejo-Rodriguez, *Ronchi Test*, in *Optical Shop Testing*, ed. D. Malacara, p. 283 (Wiley, New York, 1978).

26

Polarization

- 26.12 Literature 520
- 26.1 Introduction 467
- 26.2 Polarization States 467
 - 26.2.1 Representation of Polarization States 468
 - 26.2.2 Jones Vector 468
 - 26.2.3 Ellipse of Polarization 470
 - 26.2.4 Orthogonal Jones Vectors 471
 - 26.2.5 Jones Vectors in Different Bases 472
 - 26.2.6 Unpolarized Light 472
 - 26.2.7 Partial Polarization 473
 - 26.2.8 Polarization Matrix 473
 - 26.2.9 Stokes Vector 475
 - 26.2.10 Poincaré Sphere 478
- 26.3 Jones Matrix 479
 - 26.3.1 Definition 479
 - 26.3.2 Jones Matrix Acting on a Jones Vector 480
 - 26.3.3 Succession of Jones Matrices 480
 - 26.3.4 Jones Matrix Acting on a Polarization Matrix 481
 - 26.3.5 Examples of Jones Matrices 481
 - 26.3.6 Rotated and Mirrored Jones Matrix 482
 - 26.3.7 Jones Matrix for Different Basis Polarization States 483
 - 26.3.8 Eigenpolarizations of a Jones Matrix 483
 - 26.3.9 Jones Matrix of a Retarder 484
 - 26.3.10 Jones Matrix of a Partial Polarizer 487
 - 26.3.11 Pauli's Spin Matrices 489
 - 26.3.12 Jones Matrix Decomposition 489
- 26.4 Müller Matrix 491
 - 26.4.1 Definition 491
 - 26.4.2 Examples 492
- 26.5 Müller–Jones Matrix 493
- 26.6 Light in Anisotropic Media 494
 - 26.6.1 Anisotropic Media 494

- 26.6.2 Principal Refractive Indices of an Anisotropic Medium Without Spatial Dispersion and Optical Activity 495
- 26.6.3 Fresnel Ellipsoid 496
- 26.6.4 Index Ellipsoid 497
- 26.6.5 Types of Birefringent Media 497
- 26.7 Eigenwaves in Anisotropic Media 501
- 26.7.1 Plane Waves in Anisotropic Media 501
- 26.7.2 Eigenwaves and their Polarization 502
- 26.7.3 Properties of the Eigenpolarizations 506
- 26.7.4 The Intersection Ellipse 506
- 26.8 Jones Matrix of Propagation 507
- 26.9 Jones Matrices of Propagation for Common Media 508
- 26.9.1 Eigenpolarizations and n -values 508
- 26.9.2 Coordinate Systems 509
- 26.9.3 Uniaxial Crystal 509
- 26.9.4 Biaxial Crystal 510
- 26.9.5 CaF_2 with Spatial Dispersion at $\lambda = 193 \text{ nm}$ 511
- 26.10 Beam-splitting in an Anisotropic Medium 511
- 26.11 Examples of Polarization-optical Elements 516
- 26.11.1 Quarter-wave and Half-wave Retarder 516
- 26.11.2 Babinet–Soleil Compensator 516
- 26.11.3 Faraday Rotator 518
- 26.11.4 Brewster Plate 519
- 26.12 Literature 520

26.1

Introduction

The originators of wave optics, Huygens, Young and Fresnel, began with the assumption that light is a scalar wave, which was in perfect agreement with the physical knowledge of their time. In 1808, however, Malus discovered that light, which had been reflected from transparent media, lost its symmetry around the propagation direction: It was “polarized” [26-1]. 130 years before him, Huygens was close to the discovery of polarization as he was investigating the birefringence of calcite [26-2].

Light is a transverse electromagnetic wave – in the classical limit – and obeys Maxwell’s equations. For this reason it can be polarized. The resulting mode of vibration of the electric field vector is called “polarization”. The impact of the polarization on optical imaging has two sources: a) The contrast of interference depends on the mutual polarization of the contributing waves. b) The interaction with media and interfaces is governed by the electromagnetic boundary conditions.

The first statement means that “polarization matters”. We can neglect it only under special conditions. The second statement has the consequence that interaction with media and interfaces may change the state of the polarization. Reflection and transmission at interfaces and coatings and birefringence are the most important causes for polarization effects in common optical systems.

26.2

Polarization States

During propagation, the electric field vector changes its orientation and magnitude and describes a certain trajectory in 3D-space (figure 26-1). A propagating wave is transverse, i.e. the electric field vector, at each instant in time, is perpendicular to the propagation direction. The trajectory is periodic with the free space wavelength λ , i.e., it repeats itself every wavelength.

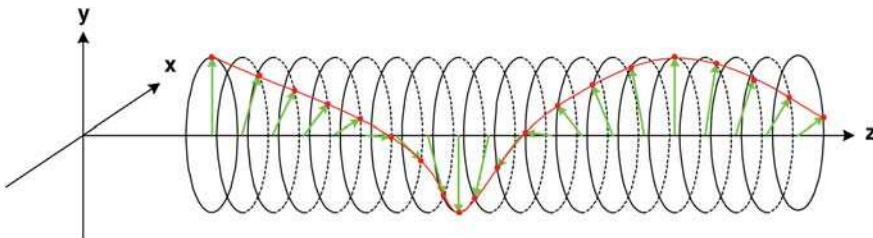


Figure 26-1: Example of the curve of the electric field vector during propagation.

The polarization state of a plane electromagnetic wave is given by the curve which the electric field vector follows in a plane which is transverse to the direction of propagation (the xy -plane in figure 26-1). Polarization states may be completely or

partially polarized ranging from completely polarized to completely unpolarized. The polarized part is in general elliptically polarized. This section is concerned with the mathematical description of the polarization states.

26.2.1

Representation of Polarization States

	Parameter	Properties
1	Ellipse of polarization Ellipticity e , orientation ψ	Complete polarization only
2	Complex parameter parameter ζ	Complete polarization only
3	Jones vector Components of \vec{E}	Complete polarization and phase
4	Stokes vector Stokes parameter S_0-S_4	Partial polarization
5	Poincare sphere Points on and inside unit sphere	Graphical representation only
6	Coherency matrix 2×2 -matrix C	Partial polarization

26.2.2

Jones Vector

Jones vectors [26-3] can describe completely polarized fields only. They denote the polarization state of light in isotropic media by a complex two-component vector.

$$\vec{E} = \left(\begin{matrix} E_x \\ E_y \end{matrix} \right) \left. \begin{matrix} \bullet \text{---} \bullet A_x e^{i\varphi_x} \\ \bullet \text{---} \bullet A_y e^{i\varphi_y} \end{matrix} \right\} \text{Complex scalar quantities.} \tag{26-1}$$

Two components are sufficient because of the transverse nature of the fields. Because the components are complex, Jones vectors are, strictly speaking, not vectors but phasors.

The Jones vector contains the state of polarization, the amplitude, and the phase of the wave. If we are interested in the polarization state only, a global phase can be extracted by defining a phase-reduced Jones vector

$$\vec{E}_{\text{red}} = e^{-i\varphi_g} \vec{E}. \tag{26-2}$$

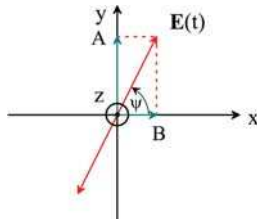
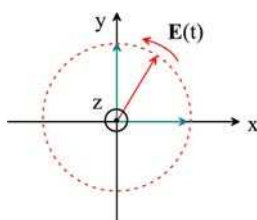
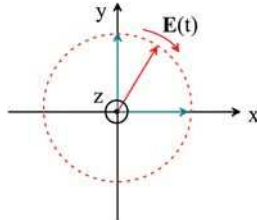
A common definition of a phase-reduced Jones vector is obtained by extracting the mean of the x - and y -polarized phase

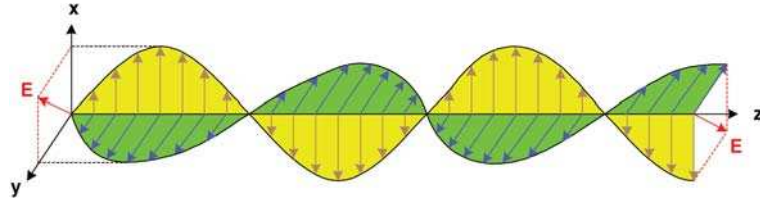
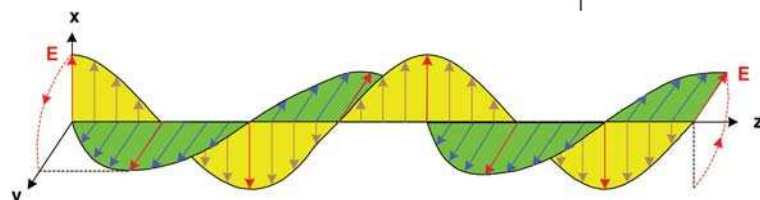
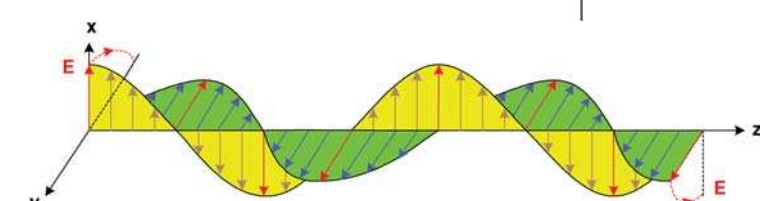
$$\varphi_g = \frac{\varphi_x + \varphi_y}{2}. \tag{26-3}$$

yielding

$$\vec{E}_{\text{red}} = \left(\begin{matrix} A_x e^{i\frac{\varphi_x - \varphi_y}{2}} \\ A_y e^{-i\frac{\varphi_x - \varphi_y}{2}} \end{matrix} \right). \tag{26-4}$$

Table 1: Simple polarization states

Polarization state	Jones vector	E-vector
Linear	$\vec{E}_{red} = \begin{pmatrix} A \\ B \end{pmatrix}, \quad A, B \in \text{Re}$	
Left-handed circular	$\begin{pmatrix} 1 \\ i \end{pmatrix}, \quad \text{i.e. } \varphi_y - \varphi_x = \frac{\pi}{2}$	
Right-handed circular	$\begin{pmatrix} 1 \\ -i \end{pmatrix}, \quad \text{i.e. } \varphi_y - \varphi_x = -\frac{\pi}{2}$	

It should be kept in mind that this is only a special case. It makes sense if both components have at least nearly identical amplitudes. If the amplitudes are extremely different, the phase value of the smaller amplitude will, in general, be known with much less accuracy than the phase of the higher amplitude and the above definition of the global phase is correspondingly inaccurate. In general, the global phase of a phase-reduced Jones vector is unknown.

The intensity corresponding to a Jones-vector is

$$I = n|E_x|^2 + n|E_y|^2, \quad (26-5)$$

where n is the refractive index of the medium in which the wave is propagating. Some elementary polarization states are listed in table 1.

Linear polarization: The phases of E_x and E_y are equal. Therefore, the phase-reduced Jones vector is real. The orientation ψ is a constant with the value

$$\psi = \text{atan} \frac{E_y}{E_x}. \quad (26-6)$$

Left-handed circular polarization: E_y advances E_x by $\pi/2$ in phase. Looking towards the direction of propagation, \vec{E} is traversing a circle anti-clockwise.

Right-handed circular polarization: E_y lags behind E_x by $\pi/2$ in phase. Looking towards the direction of propagation, \vec{E} is traversing a circle clockwise.

26.2.3

Ellipse of Polarization

The general state of a completely polarized wave is elliptical, i.e., for an arbitrary phase and amplitude difference the electric field rotates elliptically around the propagation direction.

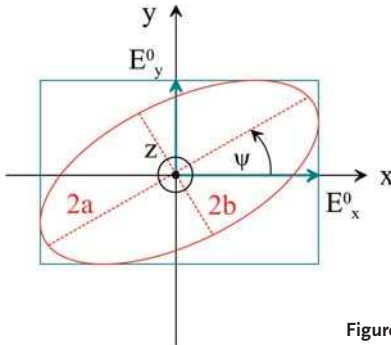


Figure 26-2: Elliptical polarization state.

The ellipse is characterized by two quantities: Its orientation ψ is [26-4]

$$\tan(2\psi) = \frac{2|E_x||E_y|}{|E_x|^2 - |E_y|^2} \cos \delta \quad (26-7)$$

and the ellipticity e as the quotient of the long and the short axis

$$e = \tan \chi = \pm \frac{b}{a} \quad (26-8)$$

with

$$\sin(2\chi) = \frac{2|E_x||E_y|}{|E_x|^2 + |E_y|^2} \sin \delta \quad (26-9)$$

and δ as the phase difference of the x - and y -component of the Jones vector

$$\delta = \arg(E_y) - \arg(E_x). \quad (26-10)$$

The sign of the ellipticity is positive for a left-handed polarization ($\delta = \pi/2$) and negative for a right-handed one ($\delta = -\pi/2$).

26.2.4

Orthogonal Jones Vectors

Two Jones vectors are orthogonal if their interference contrast is zero. Superposing two waves in orthogonal polarization states results in a straightforward addition of their intensities. Mathematically this corresponds to a vanishing inner product defined as

$$\langle \vec{E}_1, \vec{E}_2 \rangle = \vec{E}_1 \cdot \vec{E}_2^* = E_{x,1} E_{x,2}^* + E_{y,1} E_{y,2}^*. \quad (26-11)$$

Consequently, two orthogonal states \vec{E} and \vec{E}_{orth} obey

$$\vec{E} \cdot \vec{E}_{\text{orth}}^* = 0 \quad (26-12)$$

resulting in a simple expression for the components of the orthogonal state in terms of the components of the original one as

$$\begin{aligned} E_x^{\text{orth}} &= E_y^*, \\ E_y^{\text{orth}} &= -E_x^*. \end{aligned} \quad (26-13)$$

Linear x -polarization is orthogonal to linear y -polarization and left-circular polarization is orthogonal to right-circular polarization.

If the norm of two orthogonal Jones vectors is equal to one, they form an orthonormal set.

26.2.5

Jones Vectors in Different Bases

Two orthonormal Jones vectors form a basis for all completely polarized states, i.e., every Jones vector can be represented as the sum of two arbitrary orthonormal states. In explicit notation

$$\vec{E} = a\vec{E}_B + b\vec{E}_B^{\text{orth}} \quad (26-14)$$

and in vector notation

$$\vec{E} = \begin{pmatrix} a \\ b \end{pmatrix}_B. \quad (26-15)$$

The transformation from the Cartesian xy -basis into an arbitrary basis $(\vec{E}_B, \vec{E}_B^{\text{orth}})$ is performed by multiplication with a transfer matrix where the columns are formed by the basis vectors in xy -coordinates

$$\vec{E}(\text{basis } B) = \begin{bmatrix} \vec{E}_B & \vec{E}_B^{\text{orth}} \end{bmatrix} \vec{E}(\text{basis } xy). \quad (26-16)$$

Example: The transfer matrix from linear xy -basis states into left and right-circular basis states is

$$\begin{bmatrix} \vec{E}_B & \vec{E}_B^{\text{orth}} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & i \\ i & 1 \end{pmatrix}. \quad (26-17)$$

26.2.6

Unpolarized Light

If light is unpolarized, every possible polarizer (linear, circular, elliptical) transmits 50% of the incident intensity. This is equivalent to the statement that every polarization state contains half of the overall intensity.

Unpolarized light cannot be represented by a single Jones vector because Jones vectors describe completely polarized states, by definition. But it can be represented in the Jones vector calculus by the incoherent superposition of two orthogonal polarization states of equal intensity. In formal notation we may write

$$\text{Intensity of unpolarized light} \triangleq 0.5 \|\vec{E}\|^2 + 0.5 \|\vec{E}_{\text{orth}}\|^2. \quad (26-18)$$

26.2.7

Partial Polarization

Partially polarized light is a superposition of polarized and unpolarized light. The degree of polarization is defined as the intensity of the polarized part divided by the overall intensity

$$g = \frac{I_{\text{pol}}}{I_{\text{pol}} + I_{\text{unpol}}}, \quad (26-19)$$

i.e., an amount g of the intensity is in the polarized state and an amount $1 - g$ is in the unpolarized state.

Increasing the degree of polarization is known as “polarizing”, decreasing the degree of polarization is called “depolarizing”.

Partially polarized light can also be represented by the superposition of orthogonal Jones vectors.

$$\text{Intensity of partially polarized light} \triangleq \frac{1+g}{2} \|\vec{E}\|^2 + \frac{1-g}{2} \|\vec{E}_{\text{orth}}\|^2. \quad (26-20)$$

26.2.8

Polarization Matrix

A convenient way of describing partially polarized light mathematically is the coherence matrix \mathbf{P} , or polarization matrix (both names are used synonymously). It is the mean value in time of the dyadic product of the Jones vector with its complex conjugate.

$$\mathbf{P} = \langle \vec{E} \circ \vec{E}^* \rangle = \begin{pmatrix} \langle E_x E_x^* \rangle & \langle E_x E_y^* \rangle \\ \langle E_y E_x^* \rangle & \langle E_y E_y^* \rangle \end{pmatrix} = \begin{pmatrix} P_{xx} & P_{xy} \\ P_{xy}^* & P_{yy} \end{pmatrix}, \quad (26-21)$$

with the bracket defined as

$$\langle f(t) \rangle = \frac{1}{T} \int_0^T f(t) dt \quad (26-22)$$

where T denotes a time interval that is large compared with the inverse frequency.

The diagonal elements contain the intensity of the individual components while the non-diagonal elements contain their correlation. The latter are the complex conjugate of each other.

Properties of the Polarization Matrix

- a) \mathbf{P} is *hermitian*, i.e. $\mathbf{P} = \mathbf{P}^+$ (+ = complex conjugate of the transposed matrix), as follows from eq. (26-21) by inspection.
- b) The intensity of the wave equals the trace of \mathbf{P}

$$I_0 = \text{trace}(\mathbf{P}) = P_{xx} + P_{yy} = \langle E_x E_x^* \rangle + \langle E_y E_y^* \rangle \quad (26-23)$$

because the diagonal elements contain the intensity of the components.

- c) For *completely unpolarized light* the correlation between E_x and E_y is zero, i.e., the off-diagonal elements of the polarization matrix are zero. The intensities of the x - and y -components must be equal irrespective of the particular choice of coordinate system. Therefore the coherency matrix of unpolarized light is proportional to the identity matrix

$$\mathbf{P}_{\text{unpol}} = A \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \quad (26-24)$$

- d) For *completely polarized light* the polarization matrix contains only the components of the Jones vector

$$\mathbf{P}_{\text{pol}} = \begin{pmatrix} |E_x|^2 & E_x E_y^* \\ E_x^* E_y & |E_y|^2 \end{pmatrix}. \quad (26-25)$$

- e) The polarization matrix of the orthogonal polarization state follows from eq. (26-13) as

$$\mathbf{P}_{\text{pol}} = \begin{pmatrix} |E_y|^2 & -E_x E_y^* \\ -E_x^* E_y & |E_x|^2 \end{pmatrix}. \quad (26-26)$$

- f) *Partially polarized light* is composed of polarized light with an unpolarized background, i.e., the polarization matrix can be represented as the sum of the coherency matrices for polarized and unpolarized light

$$\mathbf{P} = \mathbf{P}_{\text{pol}} + A\mathbf{I} \quad \text{with} \quad A = \frac{\text{trace}(\mathbf{P}) - \text{trace}(\mathbf{P}_{\text{pol}})}{2} = I_0(1 - g). \quad (26-27)$$

- g) The *amount* A of *unpolarized light* can be derived from the solution of a quadratic equation. Computing the determinant of the polarization matrix yields

$$\begin{aligned} \det(\mathbf{P}) &= \langle E_x E_x^* \rangle \langle E_y E_y^* \rangle - \left| \langle E_x E_y^* \rangle \right|^2 \\ &= (A + |E_x|^2) (A + |E_y|^2) - |E_x|^2 |E_y|^2 \end{aligned} \quad (26-28)$$

(where we have made use of the fact that the non-diagonal elements of \mathbf{P} and \mathbf{P}_{pol} are equal, which yields $|\langle E_x E_y^* \rangle|^2 = |E_x|^2 |E_y|^2$) and from (26-27)

$$\text{trace}(\mathbf{P}_{\text{pol}}) = \text{trace}(\mathbf{P}) - 2A. \quad (26-29)$$

Inserting (26-29) into (26-28) yields a quadratic equation for A

$$A^2 - A(\langle E_x E_x^* \rangle + \langle E_y E_y^* \rangle) + \langle E_x E_x^* \rangle \langle E_y E_y^* \rangle - |\langle E_x E_y^* \rangle|^2 = 0 \quad (26-30)$$

where the terms in brackets correspond to the trace and the determinant of the polarization matrix \mathbf{P} , i.e.,

$$A^2 - A \text{trace}(\mathbf{P}) + \det(\mathbf{P}) = 0 \quad (26-31)$$

with the solution

$$A = \frac{\text{trace}(\mathbf{P})}{2} \left(1 \pm \sqrt{1 - \frac{4 \det(\mathbf{P})}{\text{trace}^2(\mathbf{P})}} \right). \quad (26-32)$$

The square-root varies between 0 for completely unpolarized light ($\text{trace}(\mathbf{P}) = 2A$, $\det(\mathbf{P}) = A^2$) and 1 for completely polarized light ($\det(\mathbf{P}) = 0$). For the latter case A must be equal to zero, i.e., only the negative sign in eq. (6.31) makes sense. With decreasing degree of polarization, A must increase monotonously. This requires the negative sign to be valid for all values of A .

h) The *degree of polarization* is

$$g = \frac{\text{trace}(\mathbf{P}_{\text{pol}})}{\text{trace}(\mathbf{P})} = \sqrt{1 - \frac{4 \det(\mathbf{P})}{\text{trace}^2(\mathbf{P})}}. \quad (26-33)$$

i) The polarization matrix defines the polarization of the field completely, but the information about the global phase is lost.

26.2.9

Stokes Vector

The Stokes vector is formed from the components of the coherency matrix. It has four real components with a distinct physical meaning:

$$\vec{S} = \begin{pmatrix} \langle E_x E_x^* \rangle + \langle E_y E_y^* \rangle \\ \langle E_x E_x^* \rangle - \langle E_y E_y^* \rangle \\ \langle E_x E_y^* \rangle + \langle E_y E_x^* \rangle \\ i(\langle E_y E_x^* \rangle - \langle E_x E_y^* \rangle) \end{pmatrix} = \begin{pmatrix} S_0 \\ S_1 \\ S_2 \\ S_3 \end{pmatrix}. \quad (26-34)$$

- The first element S_0 is the overall intensity
- The second element S_1 , is equal to the difference of the intensity transmitted by a linear x-polarizer and the intensity transmitted by a linear y-polarizer.
- The third element S_2 is equal to the difference of the intensity transmitted by a $+45^\circ$ and -45° oriented polarizer, respectively.
- The fourth element S_3 is equal to the difference of the intensity transmitted by a right-handed circular polarizer and a left-handed circular polarizer, respectively.
- There is a 1:1 relationship between the elements of the polarization matrix and the Stokes vector.

It is clear that there is no orthogonal polarization state for unpolarized light. Therefore, we call two polarization states orthogonal if their polarized parts are orthogonal. The orthogonal Stokes vector is obtained from the original one according to

$$\vec{S}_{\text{orth}} = \begin{pmatrix} \langle E_x E_x^* \rangle + \langle E_y E_y^* \rangle \\ \langle E_y E_y^* \rangle - \langle E_x E_x^* \rangle \\ -\langle E_x E_y^* \rangle - \langle E_y E_x^* \rangle \\ -i(\langle E_y E_x^* \rangle - \langle E_x E_y^* \rangle) \end{pmatrix} = \begin{pmatrix} S_0 \\ -S_1 \\ -S_2 \\ -S_3 \end{pmatrix}, \quad (26-35)$$

i.e., the sum of two Stokes vectors with orthogonal polarization states is non-zero only in the first component. Or stated another way: The incoherent superposition of two orthogonal polarization states of identical intensity provides unpolarized light.

An often used expression for the elements of the Stokes vector is obtained from the instantaneous amplitudes A_x and A_y of the fields (E_x and E_y) and their instantaneous phase difference δ

$$\vec{S} = \begin{pmatrix} \langle A_x^2 \rangle + \langle A_y^2 \rangle \\ \langle A_x^2 \rangle - \langle A_y^2 \rangle \\ 2\langle A_x A_y \cos \delta \rangle \\ 2\langle A_x A_y \sin \delta \rangle \end{pmatrix}. \quad (26-36)$$

Accordingly, in the case of a completely polarized state, the Stokes vector components can be obtained from the Jones vector from

$$\vec{S} = \begin{pmatrix} |E_x|^2 + |E_y|^2 \\ |E_x|^2 - |E_y|^2 \\ 2|E_x||E_y| \cos \delta \\ 2|E_x||E_y| \sin \delta \end{pmatrix}. \quad (26-37)$$

This relation can be reversed to obtain the Jones vector components from the Stokes vector

$$\begin{aligned}
 |E_x| &= \sqrt{\frac{S_0 + S_1}{2}}, \\
 |E_y| &= \sqrt{\frac{S_0 - S_1}{2}}, \\
 \delta &= \arctan\left(\frac{S_3}{S_2}\right).
 \end{aligned}
 \tag{26-38}$$

The relation of the parameters of the polarization ellipse to the Stokes vector components is

$$\begin{aligned}
 S_0 &= a^2(1 + \tan^2\chi), \\
 S_1 &= S_0 \cos 2\chi \cos 2\psi, \\
 S_2 &= S_0 \cos 2\chi \sin 2\psi, \\
 S_3 &= S_0 \sin 2\chi
 \end{aligned}
 \tag{26-39}$$

and the inverse relations are

$$\begin{aligned}
 \chi &= \frac{1}{2} \arcsin\left(\frac{S_3}{S_0}\right), \\
 \psi &= \frac{1}{2} \arctan\left(\frac{S_2}{S_1}\right), \\
 a &= \sqrt{\frac{S_0}{1 + \tan^2\chi}}.
 \end{aligned}
 \tag{26-40}$$

From the fact that the determinant of the polarization matrix is always positive one obtains

$$S_0^2 \geq S_1^2 + S_2^2 + S_3^2. \tag{26-41}$$

The degree of polarization of a Stokes vector is given by

$$g = \frac{\sqrt{S_1^2 + S_2^2 + S_3^2}}{S_0}. \tag{26-42}$$

The following table provides some examples of common Stokes vectors

Polarization	S_0	S_1	S_2	S_3
Unpolarized	1	0	0	0
Completely x -polarized	1	1	0	0
Completely y -polarized	1	-1	0	0
Right circularly polarized with degree 0.6	1	0	0	0.6

26.2.10

Poincaré Sphere

Consider Stokes vectors for which the degree of polarization g is equal. For these the squared sum of the components 1–3 divided by the element S_0 is constant

$$\frac{S_1^2}{S_0^2} + \frac{S_2^2}{S_0^2} + \frac{S_3^2}{S_0^2} = g. \quad (26-43)$$

The three components define the surface of a sphere of radius g , the Poincaré sphere.

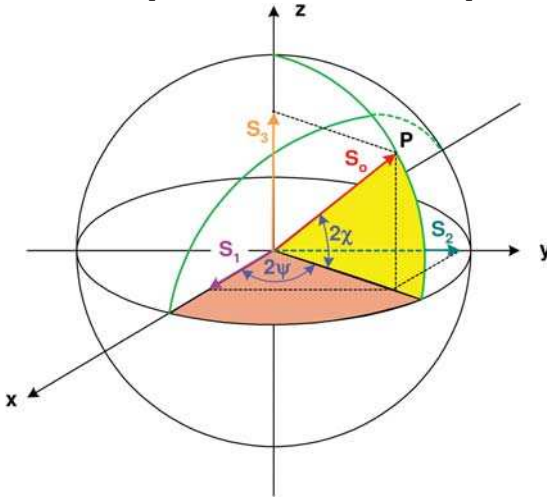


Figure 26-3: Poincaré sphere defined by the Stokes vector \vec{S} .

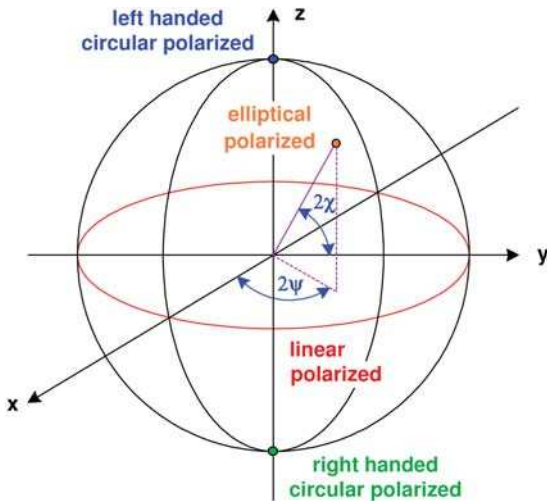


Figure 26-4: Polarization states on the Poincaré sphere.

Polarization states with an equal degree g of polarization are situated on the same sphere. The polarization state corresponds to a particular point on this sphere.

For orthogonal polarization states, S_1 – S_3 are of opposite sign. Therefore orthogonal states are located on opposite points of the sphere. The position of x/γ linear, $\pm 45^\circ$ linear and L/R circular are depicted in figure 26-4. Here, the left and right-circular state is represented by the poles of the sphere.

A distribution of polarization states as they appear within a Jones-pupil (cf. section 25.7) can be characterized by a distribution over the Poincaré sphere as illustrated in figure 26-5.

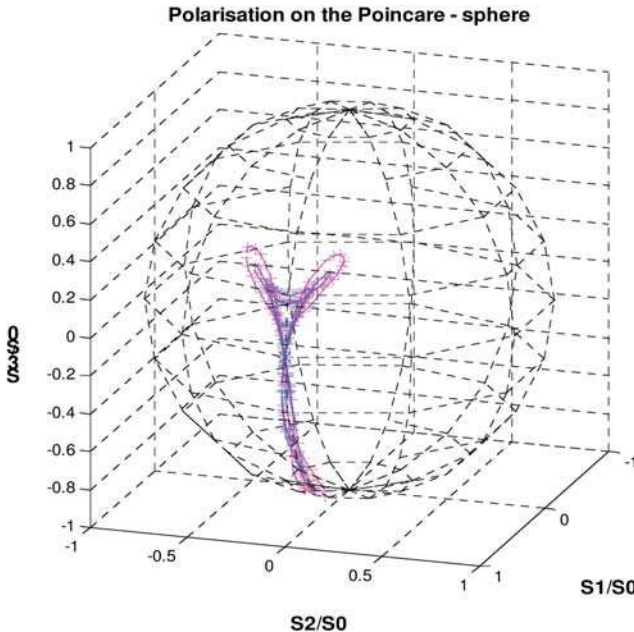


Figure 26-5: Example for a distribution of polarization states on the Poincaré sphere (Jones vectors after traversing a 111 CaF_2 plate with intrinsic birefringence).

26.3

Jones Matrix

26.3.1

Definition

Polarization optical elements change the polarization of the transmitted wave, i.e., they change the degree of polarization and/or the polarization state.

A non-depolarizing element cannot decrease the degree of polarization. It can either increase it or leave it constant. The action of non-depolarizing elements on polarized states is described by a complex 2×2 -matrix – the Jones matrix [26-3]:

$$\mathbf{J} = \begin{pmatrix} J_{xx} & J_{xy} \\ J_{yx} & J_{yy} \end{pmatrix}. \tag{26-44}$$

The Jones matrix transforms Jones vectors or polarization matrices.

A global phase factor has no effect on the polarization state. Frequently one is not interested in the complete Jones matrix but in one that is known up to a phase factor, i.e., a Jones matrix that is defined by

$$\mathbf{J} = \exp\{i\varphi_0\} \begin{pmatrix} J_{xx} & J_{xy} \\ J_{yx} & J_{yy} \end{pmatrix} \tag{26-45}$$

with an unknown φ_0 . It is called a *phase-reduced* Jones matrix.

26.3.2

Jones Matrix Acting on a Jones Vector

The action of a Jones matrix on a completely polarized state is described by a multiplication with the corresponding Jones vector (figure 26-6)

$$\vec{E}_{out} = \mathbf{J}\vec{E}_{in}. \tag{26-46}$$

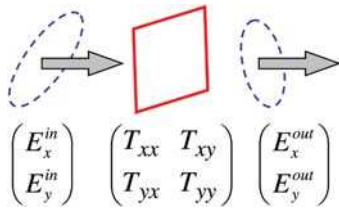


Figure 26-6: Change of polarization state due to a Jones matrix.

Therefore, the diagonal elements of the Jones matrix describe the change in the incident polarization state while the non-diagonal elements denote the cross-polarization, i.e., the excitation of the component that is orthogonal to the incident one.

26.3.3

Succession of Jones Matrices

The propagation of a Jones vector through a succession of N polarization optical elements (figure 26-7) can be described by a multiplication of the Jones vector with the product of the corresponding Jones matrices. It is important, however, to use the correct order (unlike the scalar product, the matrix product is sensitive to the order)

$$\vec{E}_{out} = \mathbf{J}_N \mathbf{J}_{N-1} \cdots \mathbf{J}_2 \mathbf{J}_1 \vec{E}_{in}. \tag{26-47}$$

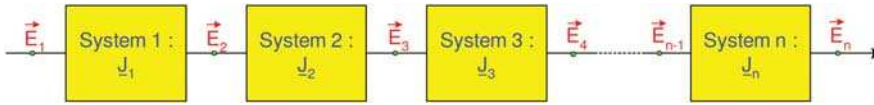


Figure 26-7: Jones vector propagation through a succession of systems described by their Jones matrices.

26.3.4

Jones Matrix Acting on a Polarization Matrix

The action of a Jones matrix on a partially polarized state is described by a two-sided multiplication according to

$$\mathbf{P}_{\text{out}} = \mathbf{J} \mathbf{P}_{\text{in}} \mathbf{J}^+ \quad (26-48)$$

Because a Jones vector transforms according to (26-46) the corresponding polarization matrices obey

$$\mathbf{P}_{\text{out}} = \langle \vec{\mathbf{E}}_{\text{out}} \circ \vec{\mathbf{E}}_{\text{out}}^+ \rangle = \langle \mathbf{J} \vec{\mathbf{E}}_{\text{in}} \circ (\mathbf{J} \vec{\mathbf{E}}_{\text{in}})^+ \rangle = \mathbf{J} \langle \vec{\mathbf{E}}_{\text{in}} \circ \vec{\mathbf{E}}_{\text{in}}^+ \rangle \mathbf{J}^+ = \mathbf{J} \mathbf{P}_{\text{in}} \mathbf{J}^+ \quad (26-49)$$

The “+” denotes the adjoint matrix, i.e. the complex-conjugate transpose. More explicitly:

$$\begin{pmatrix} P_{xx}^{\text{out}} & P_{xy}^{\text{out}} \\ P_{yx}^{\text{out}} & P_{yy}^{\text{out}} \end{pmatrix} = \begin{pmatrix} J_{xx} & J_{xy} \\ J_{yx} & J_{yy} \end{pmatrix} \begin{pmatrix} P_{xx}^{\text{in}} & P_{xy}^{\text{in}} \\ P_{yx}^{\text{in}} & P_{yy}^{\text{in}} \end{pmatrix} \begin{pmatrix} J_{xx}^* & J_{yx}^* \\ J_{xy}^* & J_{yy}^* \end{pmatrix} \quad (26-50)$$

Due to the multiplication with the adjoint matrix, any global phase factor in the Jones matrix cancels. \mathbf{J} can also be replaced by the phase-reduced Jones matrix. This is in accordance with the fact that the polarization matrix contains no information about the phase of the field.

26.3.5

Examples of Jones Matrices

Before we continue with the properties and representations of Jones matrices, some examples of “typical” and often applied polarization optical components will be given. All the following matrices are valid for collimated light only. Actual elements show a dependence of the Jones matrix on the direction of propagation that is not addressed in these matrices. To do this, more realistic models must be applied, as we will discover later.

Name	Matrix	Description
Neutral element	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	This is the identity matrix because the output polarization state is identical to the input state.
x-polarizer	$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$	y-component of incident field is deleted.
y-polarizer	$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$	x-component of incident field is deleted.
Partial x-polarizer	$\begin{pmatrix} 1 & 0 \\ 0 & g \end{pmatrix}$	The y-component is diminished by a factor $g < 1$. The resulting degree of polarization is $\frac{(1-g)^2}{(1+g)^2}$
Quarter-wave retarder	$\begin{pmatrix} 1 & -i \\ -i & 1 \end{pmatrix}$	Transforms the linear x-polarized state into the right circular polarized state
Half-wave retarder	$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$	Transforms the linear x-polarized state into the linear y-polarized state
General retarder	$\begin{pmatrix} \exp\{-i\frac{\Delta\phi}{2}\} & 0 \\ 0 & \exp\{i\frac{\Delta\phi}{2}\} \end{pmatrix}$	Introduces a phase-shift of $\Delta\phi$ between the x- and y-polarized components
Rotator	$\begin{pmatrix} \cos \alpha & \sin \alpha \\ -\sin \alpha & \cos \alpha \end{pmatrix}$	Rotates a linear polarization state by the angle α around the propagation direction.

26.3.6

Rotated and Mirrored Jones Matrix

Even the simple polarization optical elements of the section above can have different angular orientations provided by a rotation around the ray direction and they can be given in a mirrored coordinate system (figure 26-8).

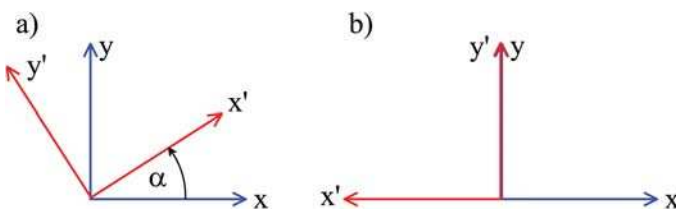


Figure 26-8: a) Local coordinate system rotated by an angle α with respect to the global coordinate system; b) local coordinate system mirrored with respect to the y-axis.

For a proper description of the action of a rotated component in the global coordinate system, the incident polarization state has to be transformed into the local coordinate system of the component and back again. This corresponds to the application

of a transformation matrix and its inverse. For a rotation angle α we obtain for a general Jones matrix \mathbf{J}

$$\mathbf{J}_\alpha = \mathbf{R}(-\alpha)\mathbf{J}\mathbf{R}(\alpha) = \begin{pmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{pmatrix} \begin{pmatrix} J_{xx} & J_{xy} \\ J_{yx} & J_{yy} \end{pmatrix} \begin{pmatrix} \cos \alpha & \sin \alpha \\ -\sin \alpha & \cos \alpha \end{pmatrix}. \quad (26-51)$$

The corresponding operation for the mirrored Jones matrix is

$$\mathbf{J}_M = \mathbf{Q}_y\mathbf{J}\mathbf{Q}_y = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} J_{xx} & J_{xy} \\ J_{yx} & J_{yy} \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} J_{xx} & -J_{xy} \\ -J_{yx} & J_{yy} \end{pmatrix}. \quad (26-52)$$

26.3.7

Jones Matrix for Different Basis Polarization States

The rotation and mirror operation are just special cases of a transformation into another basic polarization state. If a Jones matrix is given for a particular choice of basis polarization state (such as, for instance, left and right circular polarized) and its action is to be described for another set of basis polarization states (for instance xy -linear polarized) the same operation as in the section above is applied: The incident polarization state is transformed into the basis of the Jones matrix and back again. According to section 26.2.5 the corresponding transfer matrix is obtained from the new basis Jones vectors formulated in terms of the old basis Jones vectors. Denoting the resulting transfer matrix by \mathbf{T} we obtain

$$\mathbf{J}_T = \mathbf{T}^{-1}\mathbf{J}\mathbf{T}. \quad (26-53)$$

26.3.8

Eigenpolarizations of a Jones Matrix

The eigenpolarizations of a Jones matrix are those polarization states that are left invariant by the Jones matrix. They follow from a solution of the eigenvalue problem

$$\begin{pmatrix} J_{xx} & J_{xy} \\ J_{yx} & J_{yy} \end{pmatrix} \begin{pmatrix} E_{x1} & E_{x2} \\ E_{y1} & E_{y2} \end{pmatrix} = C \begin{pmatrix} 1 & 0 \\ 0 & D \exp\{i\Delta\varphi\} \end{pmatrix} \begin{pmatrix} E_{x1} & E_{x2} \\ E_{y1} & E_{y2} \end{pmatrix} \quad (26-54)$$

with the Jones vectors \vec{E}_1 and \vec{E}_2 of the two eigenpolarizations and a diagonal matrix of the eigenvalues. C is an arbitrary complex constant. D is the quotient of the amplitudes of the eigenpolarizations and is by this means a measure of the polarizing property of the element.

The eigenpolarizations of a plane wave reflected from a plane interface are those linear polarizations that are perpendicular and parallel to the plane of incidence. The eigenpolarizations of a uniaxial crystal retarder are parallel and perpendicular to the crystal axis. In the general case, the eigenpolarizations are elliptical.

26.3.9

Jones Matrix of a Retarder

A retarder has orthogonal eigenpolarizations on which it imposes a phase-shift $\Delta\phi$ (figure 26-9). $\Delta\phi$ is called the “retardance”.

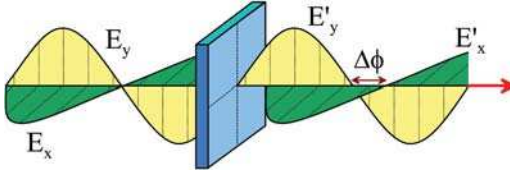


Figure 26-9: Illustration of the phase-shift between the x - and y -components of the Jones vector caused by a linear retarder.

If the eigenpolarizations are linear, the retarder is called a linear retarder (this case is illustrated in figure 26-9). A circular retarder has circular eigenpolarizations.

In the reference frame of the eigenpolarizations the Jones matrix is diagona

$$\mathbf{J}_R = e^{i\varphi_0} \begin{pmatrix} e^{-i\frac{\Delta\phi}{2}} & 0 \\ 0 & e^{i\frac{\Delta\phi}{2}} \end{pmatrix}, \quad (26-55)$$

where φ_0 is a mean phase-shift imposed on both components simultaneously.

Two cases are of particular importance and will be discussed briefly in the following: The *rotated linear retarder* and the *circular retarder*.

The Jones matrix of a linear retarder rotated by an angle α with respect to the global coordinate system is

$$\begin{aligned} \mathbf{J}_{R,\alpha} &= \mathbf{R}(-\alpha)\mathbf{J}_R\mathbf{R}(\alpha) \\ &= \begin{pmatrix} \cos \frac{\Delta\phi}{2} - i \sin \frac{\Delta\phi}{2} \cos 2\alpha & i \sin \frac{\Delta\phi}{2} \sin 2\alpha \\ i \sin \frac{\Delta\phi}{2} \sin 2\alpha & \cos \frac{\Delta\phi}{2} + i \sin \frac{\Delta\phi}{2} \cos 2\alpha \end{pmatrix}. \end{aligned} \quad (26-56)$$

For $\Delta\phi = \pi/2$ and $\alpha = 45^\circ$ the matrix represents a quarter-wave retarder under 45° to the incident linear polarization vector. The output polarization state is circular.

For $\Delta\phi = \pi$ and $\alpha = 45^\circ$ we obtain the Jones matrix of a half-wave retarder. This element flips an incident linear polarization state at its axis.

The Jones matrix of a stack of retarders is unitary (because of vanishing absorption), i.e., $\mathbf{J}^+ = \mathbf{J}^{-1}$. It can be shown that any unitary matrix has the shape

$$\mathbf{J}_{R\text{-Stack}} = \begin{pmatrix} a + ib & c + id \\ -c + id & a - ib \end{pmatrix}, \quad a^2 + b^2 + c^2 + d^2 = 1. \quad (26-57)$$

The eigenvalues of the Jones matrix of a stack of retarders are [26-7]

$$\lambda_{1,2} = a \pm \sqrt{a^2 - 1}, \quad (26-58)$$

i.e., they depend solely on the real part of the diagonal elements.

The eigenvectors have the form

$$\vec{E} = \begin{pmatrix} \cos \psi \\ e^{i\zeta} \sin \psi \end{pmatrix} \quad (26-59)$$

with

$$\tan \zeta = \frac{c}{d}, \quad \tan \psi = \frac{-b \pm \sqrt{1 - a^2}}{\sqrt{c^2 + d^2}} \quad (26-60)$$

The eigenvectors are linearly polarized if $c = 0$.

If retardance is present in an optical component and the incident polarization state is not equal to an eigenpolarization, energy is transferred from the incident polarization state into the orthogonal state. A linear polarized incident state of intensity I_0 is transformed by the retarder into an elliptical polarized state

$$E_{\text{out}} = \mathbf{J}_{R,\alpha} \begin{pmatrix} \sqrt{I_0} \\ 0 \end{pmatrix} = \sqrt{I_0} \begin{pmatrix} \cos \frac{\Delta\phi}{2} - i \sin \frac{\Delta\phi}{2} \cos 2\alpha \\ i \sin \frac{\Delta\phi}{2} \sin 2\alpha \end{pmatrix}. \quad (26-61)$$

The intensity of light in the intended polarization state is diminished, i.e., a polarization loss occurs. A quantitative measure of the polarization loss is the intensity in the orthogonal polarization state divided by the overall transmitted intensity.

$$\Delta P = \frac{I_{\perp}}{I_{\parallel} + I_{\perp}}. \quad (26-62)$$

The intensity of the component, which is polarized orthogonal to the incident one, is

$$\frac{I_{\perp}}{I_0} = \sin^2 \frac{\Delta\phi}{2} \sin^2 2\alpha. \quad (26-63)$$

The dependence on both $\Delta\phi$ and α is sketched in figure 25.8.

The intensity of the orthogonal polarized component becomes a maximum for orientations that are odd multiples of 45° .

A left-circular polarized incident field of intensity I_0 provides an output Jones vector of

$$\vec{E}_{\text{out}} = \mathbf{J}_{R,\alpha} \sqrt{\frac{I_0}{2}} \begin{pmatrix} 1 \\ i \end{pmatrix} = \sqrt{\frac{I_0}{2}} \left[\cos \frac{\Delta\phi}{2} \begin{pmatrix} 1 \\ i \end{pmatrix} - e^{-i2\alpha} \sin \frac{\Delta\phi}{2} \begin{pmatrix} i \\ 1 \end{pmatrix} \right], \quad (26-64)$$

i.e., a right-circular polarized component exits, which has a relative intensity of

$$\frac{I_{\perp}}{I_0} = \sin^2 \frac{\Delta\phi}{2} \quad (26-65)$$

and a phase-shift of -2α .

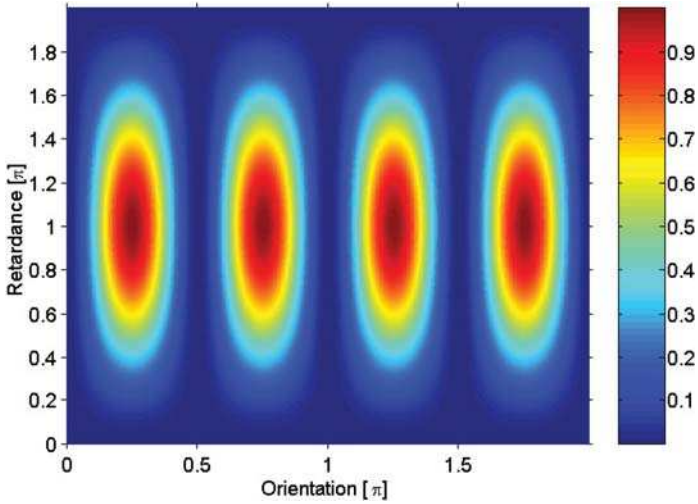


Figure 26-10: Relative intensity of the orthogonal component after traversing a rotated linear retarder.

In contrast to the linearly polarized incidence, the intensity of the orthogonally polarized component is now independent of the orientation. The orientation determines instead the phase of the orthogonal component.

A revolving half-wave plate with a retardance of π transforms left-circular into right-circular polarized light with phase 2α , i.e., it can be used as a variable phase-shifter.

To describe the effect of a circular retarder on an incident linear polarized state, its Jones matrix is conveniently transformed into xy -linear polarized states by the application of (26-17) and (26-53)

$$\begin{aligned}
 \mathbf{J}_{R,c} &= \frac{1}{2} \begin{pmatrix} 1 & i \\ i & 1 \end{pmatrix} \begin{pmatrix} e^{-i\frac{\Delta\phi}{2}} & 0 \\ 0 & e^{i\frac{\Delta\phi}{2}} \end{pmatrix} \begin{pmatrix} 1 & -i \\ -i & 1 \end{pmatrix} \\
 &= \begin{pmatrix} \cos \frac{\Delta\phi}{2} & -\sin \frac{\Delta\phi}{2} \\ \sin \frac{\Delta\phi}{2} & \cos \frac{\Delta\phi}{2} \end{pmatrix} = \mathbf{R} \left(-\frac{\Delta\phi}{2} \right)
 \end{aligned} \tag{26-66}$$

which results in a rotation matrix, i.e., a circular retarder of retardance $\Delta\phi$ rotates a traversing linear polarization state by an amount $-\Delta\phi/2$.

26.3.10

Jones Matrix of a Partial Polarizer

If an unpolarized wave is incident on a partial polarizer it is afterwards partially polarized. A partial polarizer has two eigenpolarizations belonging to eigenvalues of different amplitude (figure 26-11). This phenomenon is also called diattenuation.

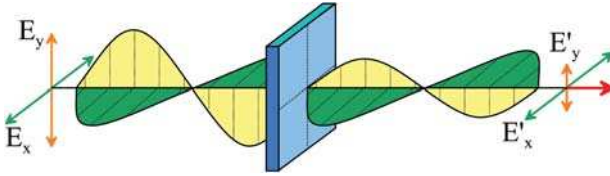


Figure 26-11: Illustration of the diattenuation between the x - and y -component of the Jones vector caused by a partial polarizer with linear eigenpolarizations.

If the eigenpolarizations are linear, the partial polarizer is called a linear partial polarizer (or linear diattenuator – this case is illustrated in the figure). For circular eigenpolarizations it is called a circular partial polarizer.

In the reference frame of the eigenpolarizations its Jones matrix is diagonal

$$J_P = \begin{pmatrix} 1 & 0 \\ 0 & A \end{pmatrix}, \quad (26-67)$$

with a real quantity $A < 1$ describing the attenuation of the component orthogonal to the transmitted one. $1 - A$ is the diattenuation of the element.

Incident unpolarized light attains a degree of polarization of (figure 26-12)

$$g = \frac{1 - A^2}{1 + A^2}. \quad (26-68)$$

Likewise, in the opposite way, we obtain the diattenuation from the degree of polarization by the use of

$$A = \sqrt{\frac{1 - g}{1 + g}}. \quad (26-69)$$

The Jones matrix of a linear partial polarizer rotated by an angle α with respect to the global coordinate system is

$$\begin{aligned} J_{P,\alpha} &= \mathbf{R}(-\alpha) J_P \mathbf{R}(\alpha) \\ &= \begin{pmatrix} (1 + A) + (1 - A) \cos 2\alpha & (1 - A) \sin 2\alpha \\ (1 - A) \sin 2\alpha & (1 + A) - (1 - A) \cos 2\alpha \end{pmatrix}. \end{aligned} \quad (26-70)$$

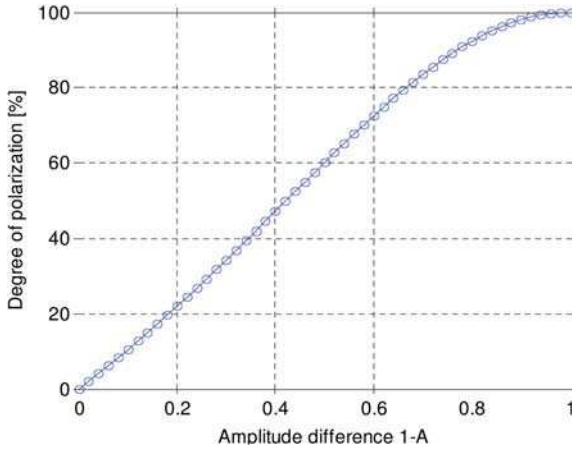


Figure 26-12: Degree of polarization depending on the amplitude difference of both eigenpolarizations of a partial polarizer.

Because the Jones matrix is purely real, a linear polarized incident state of intensity I_0 is transformed into an attenuated rotated linear polarized state

$$\vec{E}_{\text{out}} = J_{P,\alpha} \begin{pmatrix} \sqrt{I_0} \\ 0 \end{pmatrix} = \sqrt{I_0} \begin{pmatrix} (1 + A) + (1 - A) \cos 2\alpha \\ (1 - A) \sin 2\alpha \end{pmatrix}.$$

The intensity of the component polarized orthogonal to the incident one is

$$\frac{I_{\perp}}{I_0} = (1 - A)^2 \sin^2 2\alpha. \tag{26-71}$$

The dependence on both A and α is sketched in figure 26-13.

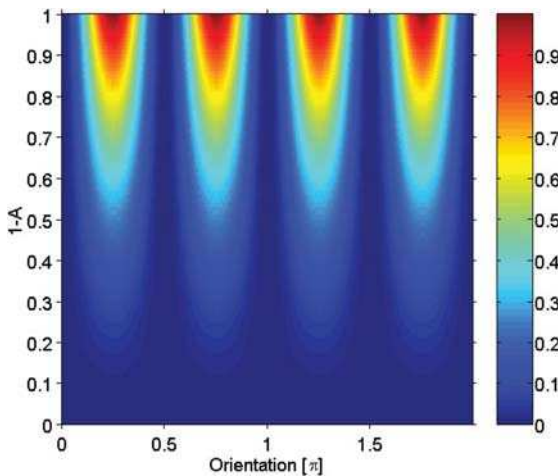


Figure 26-13: Relative intensity of the orthogonal component after traversing a rotated linear partial polarizer.

26.3.11

Pauli's Spin Matrices

Pauli's spin matrices form a basis for the complex 2×2 matrices. They are defined as

$$\mathbf{s}_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{s}_1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \mathbf{s}_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{s}_3 = \begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix}. \quad (26-72)$$

Interpreted as single Jones matrices they have a distinct physical meaning:

- \mathbf{s}_0 is the identity matrix, i.e., it represents the neutral Jones matrix.
- \mathbf{s}_1 applies a phase-shift of π to the y -component, i.e., it is a half-wave plate where the y -axis is the fast axis.
- \mathbf{s}_2 exchanges the orientation of the x - and y -component. It replaces a state by its mirror state
- \mathbf{s}_3 does the same with an additional phase-shift of $+90^\circ$ for the x and -90° for the y -component.

26.3.12

Jones Matrix Decomposition

For analysis and synthesis of Jones matrices their decomposition into elementary Jones matrices is very convenient (figure 26-14). The main idea is to write the Jones matrix as a product of elementary matrices that correspond either to simple matrices from a mathematical point of view or to simple polarization-optical components from an experimental point of view.

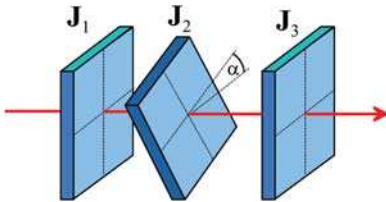


Figure 26-14: Series of rotated Jones matrices.

This task is performed by the Jones matrix decomposition. Several decompositions exist. In the following we proceed from simple decomposition for special cases to general ones.

- The Jones matrix of a non-absorbing component preserves the norm of an incident Jones vector. Therefore it is unitary, i.e., the adjoint Jones matrix is equal to the inverse Jones matrix

$$\mathbf{J}^{-1} = \mathbf{J}^+. \quad (26-73)$$

Any unitary matrix can be written as the product of a rotator and a rotated retarder.

$$\mathbf{J} = \text{Rotator} \cdot \text{Retarder} = \mathbf{R}(\beta)\mathbf{J} = \mathbf{R}(\beta)\mathbf{R}(-\alpha)\mathbf{J}_R\mathbf{R}(\alpha) \quad (26-74)$$

- b) Furthermore, the Jones matrix of a non-absorbing (unitary) component can be described as the product of two rotated quarter-wave retarders and one half-wave retarder

$$\mathbf{J} = \text{QWP}(\alpha_1) \cdot \text{HWP}(\alpha_2) \cdot \text{QWP}(\alpha_3). \quad (26-75)$$

The task can be provided with any succession of the elements. A good review of the elements and their derivation can be found in [26-8].

This combination is called Simon-Mukunda Polarization Gadget after its two inventors. A useful special case is the combination of two half-wave retarders, rotated against each other about an angle α :

$$\begin{aligned} & \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \cos \alpha & \sin \alpha \\ -\sin \alpha & \cos \alpha \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{pmatrix} \\ & = \begin{pmatrix} \cos^2 \alpha - \sin^2 \alpha & -2 \cos \alpha \sin \alpha \\ 2 \cos \alpha \sin \alpha & \cos^2 \alpha - \sin^2 \alpha \end{pmatrix} = \begin{pmatrix} \cos(2\alpha) & -\sin(2\alpha) \\ \sin(2\alpha) & \cos(2\alpha) \end{pmatrix}. \end{aligned} \quad (26-76)$$

These are equivalent to a rotator with a rotation angle of 2α .

- c) A general Jones matrix \mathbf{J} containing eight independent real quantities can be decomposed into two rotators, two retarders and a partial polarizer [26-9].

$$\mathbf{J} = \mathbf{R}(\alpha_1)\mathbf{J}_R(\alpha_2, \beta_2)\mathbf{J}_P(\sigma_1, \sigma_2)\mathbf{J}_R(-\alpha_3, \beta_2)\mathbf{R}(-\alpha_4) \quad (26-77)$$

where \mathbf{J}_R denotes a rotated retarder according to eq. (26-51) and \mathbf{J}_P denotes a partial polarizer.

- d) The singular value decomposition (SVD) of a square complex matrix is based on the following theorem: Any complex $N \times N$ matrix \mathbf{J} can be written as the product of a unitary $N \times N$ matrix \mathbf{U} , a diagonal matrix \mathbf{D} with real, positive elements (the singular values) and the adjoint of a second unitary matrix \mathbf{V} [26-12]

$$\mathbf{J} = \mathbf{U}\mathbf{D}\mathbf{V}^+. \quad (26-78)$$

The SVD can be found as a standard routine in any scientific computer library.

- e) The polar decomposition theorem for Jones matrices states that any Jones matrix can be decomposed into a partial polarizer and a retarder [26-10]

$$\mathbf{J} = \mathbf{J}_R\mathbf{J}_P = \mathbf{J}'_P\mathbf{J}_R. \quad (26-79)$$

It can be derived from the singular value decomposition and – more importantly – it can be easily computed. Inserting the identity matrix in the form $\mathbf{U}^+\mathbf{U}$ into the right-hand side of the SVD (26-78) yields

$$\mathbf{J} = \mathbf{U}\mathbf{D}\mathbf{U}^+\mathbf{U}\mathbf{V}^+, \quad (26-80)$$

i.e., the partial polarizer and retarder are

$$\mathbf{J}_P = \mathbf{U}\mathbf{D}\mathbf{U}^+, \quad \mathbf{J}_R = \mathbf{U}\mathbf{V}^+. \quad (26-81)$$

If we instead insert $\mathbf{V}^+\mathbf{V}$ in the left side of the SVD (26-78) we obtain the second form

$$\mathbf{J} = \mathbf{U}\mathbf{V}^+\mathbf{V}\mathbf{D}\mathbf{V}^+, \quad (26-82)$$

i.e., the retarder and partial polarizer are

$$\mathbf{J}_R = \mathbf{U}\mathbf{V}^+, \quad \mathbf{J}_P = \mathbf{V}\mathbf{D}\mathbf{V}^+ \quad (26-83)$$

\mathbf{J}_P and \mathbf{J}'_P are hermitian because $(\mathbf{U}\mathbf{D}\mathbf{U}^+)^+ = \mathbf{U}(\mathbf{D})^+ = \mathbf{U}\mathbf{D}\mathbf{U}^+$. \mathbf{J}_R is unitary because $\mathbf{U}\mathbf{V}^+(\mathbf{U}\mathbf{V}^+)^+ = \mathbf{U}\mathbf{V}^+\mathbf{V}\mathbf{U}^+ = \mathbf{1}$.

26.4 Müller Matrix

The Jones matrix calculus does not allow the description of a true depolarization, i.e., a reduction of the degree of polarization. The reason for this fact is that the Jones matrix is acting on single polarization states (Jones vectors), but to decrease the degree of polarization we need to reduce the correlation between two orthogonal states, and to achieve that, both states must be transformed together, i.e., instead of a 2×2 matrix we would expect a 4×4 matrix.

26.4.1

Definition

The Müller matrix \mathbf{M} transforms the Stokes vectors.

$$\vec{\mathcal{S}}_{\text{out}} = \mathbf{M}\vec{\mathcal{S}}_{\text{in}}. \quad (26-84)$$

Accordingly it is a real 4×4 matrix

$$\mathbf{M} = \begin{pmatrix} M_{00} & M_{01} & M_{02} & M_{03} \\ M_{10} & M_{11} & M_{12} & M_{13} \\ M_{20} & M_{21} & M_{22} & M_{23} \\ M_{30} & M_{31} & M_{32} & M_{33} \end{pmatrix}. \quad (26-85)$$

It also describes – in contrast to a Jones matrix – depolarizing components (a Jones matrix can only increase the degree of polarization or leave it constant).

Comparing the matrix with the physical meaning of the Stokes vector components, we can immediately deduce the physical meaning of the single components.

- The first column contains the effect on the overall intensity.
- The second column contains the effect on the $x\gamma$ -linearly polarized part.
- The third column contains the effect on the 45° linearly polarized part.
- The fourth column contains the effect on the circularly polarized part.

26.4.2

Examples

Name	Matrix	Description
Neutral element	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$	Identity matrix. The output polarization state is identical to the input state
x -polarizer	$\frac{1}{2} \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$	Because of $I_x = \frac{1}{2}(S_0 + S_1)$, $I_y = 0$
y -polarizer	$\frac{1}{2} \begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$	Because of $I_y = \frac{1}{2}(S_0 - S_1)$, $I_x = 0$
Quarter-wave retarder	$\frac{1}{2} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{pmatrix}$	Transforms the linear x -polarized state into the right-circular polarized state
Half-wave retarder	$\frac{1}{2} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$	Transforms the linear x -polarized state into the linear y -polarized state
General retarder	$\frac{1}{2} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos \Delta\varphi & \sin \Delta\varphi \\ 0 & 0 & -\sin \Delta\varphi & \cos \Delta\varphi \end{pmatrix}$	Introduces a $\Delta\varphi$ phase-shift between the x - and y -polarized component.
Depolarizer	$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$	Reduces the degree of polarization to zero.

26.5 Müller–Jones Matrix

A Müller–Jones matrix is a Müller matrix that is equivalent to a Jones matrix. For such a non-depolarizing Müller matrix, additional relations must hold because a Müller matrix contains 16 independent real elements while a phase-reduced Jones matrix contains only seven independent real elements. It can be shown [26-11] that a modified inner product of two columns, in each case, must be zero:

$$M_{0k}M_{0j} - \sum_{i=1}^3 M_{ik}M_{ij} = 0, \quad j, k = 1, \dots, 3, \quad j \neq k, \quad (26-86)$$

$$M_{0k}^2 - \sum_{i=1}^3 M_{ik}^2 + M_{00}^2 - \sum_{i=1}^3 M_{i0}^2 = 0, \quad k = 1, \dots, 3.$$

These relations are used in [26-11] to determine whether a measured Müller matrix is actually a Jones matrix. To this end, the above relations must be valid within the measurement accuracy.

A non-depolarizing Müller matrix must leave the degree of polarization invariant. Therefore it must satisfy the condition [26-18]

$$\text{trace}(\mathbf{M}^T \mathbf{M}) = 4M_{00}^2. \quad (26-87)$$

The Müller–Jones matrix \mathbf{M} can be reduced to its generating Jones matrix \mathbf{J} by means of the following relations.

Based on the relation between a Jones matrix \mathbf{J} and the corresponding Müller–Jones matrix \mathbf{M} with elements M_{ij} and Pauli’s spin-matrices \mathbf{s}_i [26-5]

$$M_{ij} = 0.5 \text{ trace}(\mathbf{J} \mathbf{s}_j \mathbf{J} \mathbf{s}_i), \quad i, j = 0, 1, 2, 3. \quad (26-88)$$

We obtain 16 equations of the type

$$M_{00} = \frac{1}{2} (|J_{xx}|^2 + |J_{xy}|^2 + |J_{yx}|^2 + |J_{yy}|^2), \quad (26-89)$$

$$M_{10} = \dots,$$

from which the Jones matrix elements can be calculated.

The amplitude of the Jones matrix elements follows from

$$|J_{xx}| = \frac{1}{\sqrt{2}} \sqrt{M_{00} + M_{01} + M_{10} + M_{11}}$$

$$|J_{xy}| = \frac{1}{\sqrt{2}} \sqrt{M_{00} - M_{01} + M_{10} - M_{11}}$$

$$|J_{yx}| = \frac{1}{\sqrt{2}} \sqrt{M_{00} + M_{01} - M_{10} - M_{11}}$$

$$|J_{yy}| = \frac{1}{\sqrt{2}} \sqrt{M_{00} - M_{01} - M_{10} + M_{11}} \quad (26-90)$$

and the phase

$$\begin{aligned}
 \arg(J_{xx}) &= 0 \\
 \arg(J_{xy}) &= \arctan\left(\frac{-M_{03} - M_{13}}{M_{02} + M_{12}}\right) \\
 \arg(J_{yx}) &= \arctan\left(\frac{M_{30} + M_{31}}{M_{20} + M_{21}}\right) \\
 \arg(J_{yy}) &= \arctan\left(\frac{M_{32} - M_{23}}{M_{22} + M_{33}}\right).
 \end{aligned} \tag{26-91}$$

The previous two real relations can be combined to a complex one

$$\begin{aligned}
 J_{xx} &= \frac{1}{\sqrt{2}} \sqrt{M_{00} + M_{01} + M_{10} + M_{11}} \\
 J_{xy} &= \frac{1}{\sqrt{2}} \sqrt{\frac{M_{00} - M_{01} + M_{10} - M_{11}}{(M_{02} + M_{12})^2 + (M_{03} + M_{13})^2}} (M_{02} + M_{12} - i[M_{03} + M_{13}]) \\
 J_{yx} &= \frac{1}{\sqrt{2}} \sqrt{\frac{M_{00} + M_{01} - M_{10} - M_{11}}{(M_{20} + M_{21})^2 + (M_{30} + M_{31})^2}} (M_{20} + M_{21} + i[M_{30} + M_{31}]) \\
 J_{yy} &= \frac{1}{\sqrt{2}} \sqrt{\frac{M_{00} - M_{01} - M_{10} + M_{11}}{(M_{22} + M_{33})^2 + (M_{32} - M_{23})^2}} (M_{22} + M_{33} + i[M_{32} - M_{23}]) .
 \end{aligned} \tag{26-92}$$

26.6

Light in Anisotropic Media

26.6.1

Anisotropic Media

A medium is called anisotropic if the refractive index depends on the direction of propagation. Assuming the validity of Hook's law, the relation between the induced polarization and the electric field, and therefore between the displacement current and the electric field, is linear.

It follows that in an anisotropic medium the relation between the displacement current and the electric field is given by the linear tensor relation

$$D_j = \sum_l \epsilon_{jl} E_l + \sum_{l,m} \gamma_{jlm} \nabla_m E_l + \sum_{l,m,q} \alpha_{jimq} \nabla_m \nabla_q E_l + \dots \tag{26-93}$$

The first tensor product describes the anisotropy that depends on the electric field at the considered spatial point itself. It is called birefringence and is governed by the dielectric tensor of second order

$$\boldsymbol{\epsilon} = (\epsilon_{ik}). \tag{26-94}$$

The following tensor products stem from a Taylor expansion of the electric field around the considered point, i.e., they describe the fact that the displacement current depends also on the field in an infinitesimal area surrounding the considered point. This phenomenon is called “spatial dispersion” in analogy to (time) dispersion where the displacement current at a given instant depends, not only on the electric field at that instant, but on the field strength infinitesimally shortly before it.

The second tensor product relates a tensor of third order with the spatial derivative of the electric field. It yields the circular anisotropy or “optical activity”.

The third term is the product of a tensor of fourth order with the second spatial derivative of the electric field. It provides the anisotropy caused by “spatial dispersion” also called “intrinsic birefringence”.

Considering monochromatic plane waves with propagation direction \vec{k}

$$\vec{E}(\vec{r}) = \vec{E}_0 \exp\{i\vec{k} \cdot \vec{r}\} \quad (26-95)$$

the spatial derivatives can be carried out and we obtain

$$D_j = \sum_l \varepsilon_{jl} E_l + i \sum_{l,m} \gamma_{jlm} k_m E_l - \sum_{l,m,q} \alpha_{jlmq} k_m k_q E_l + \dots \quad (26-96)$$

For a given propagation direction \vec{k} the higher-order tensors can be reduced to a dielectric tensor of second order by the relation

$$\varepsilon'_{jl} = \varepsilon_{jl} + i \sum_{m=1}^3 \gamma_{jlm} k_m - \sum_{m=1}^3 \sum_{q=1}^3 \alpha_{jlmq} k_m k_q + \dots \quad (26-97)$$

It follows from the conservation of energy that for non-magnetic, lossless media, the dielectric tensor is hermitian [26-14],

$$\varepsilon'_{ik} = \varepsilon'^{*}_{ki}. \quad (26-98)$$

26.6.2

Principal Refractive Indices of an Anisotropic Medium Without Spatial Dispersion and Optical Activity

If the coordinate system is aligned with the principal axes of the medium, the dielectric tensor becomes diagonal (a hermitian tensor can always be diagonalized)

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_{xx} & 0 & 0 \\ 0 & \varepsilon_{yy} & 0 \\ 0 & 0 & \varepsilon_{zz} \end{pmatrix} = \begin{pmatrix} n_x^2 & 0 & 0 \\ 0 & n_y^2 & 0 \\ 0 & 0 & n_z^2 \end{pmatrix}, \quad (26-99)$$

where ε_{ii} and n_i are the principal dielectric constants and the principal refractive indices, respectively. From the diagonal elements follows the well-known distinction

of three types of media: isotropic ($n^2 = \epsilon_{xx} = \epsilon_{yy} = \epsilon_{zz}$), uniaxial ($n_o^2 = \epsilon_{xx} = \epsilon_{yy}$, $n_e^2 = \epsilon_{zz}$) and biaxial ($n_x^2 = \epsilon_{xx}$, $n_y^2 = \epsilon_{yy}$, $n_z^2 = \epsilon_{zz}$).

In general, the electric field and the displacement are not parallel. It is only along the principal axes that they are, in fact, parallel, i.e.,

$$D_x = \epsilon_{xx} E_x, \quad D_y = \epsilon_{yy} E_y, \quad D_z = \epsilon_{zz} E_z. \quad (26-100)$$

26.6.3

Fresnel Ellipsoid

From the equation for the energy density of the electric field

$$\rho_E = \frac{1}{2} \vec{E} \cdot \vec{D} \quad (26-101)$$

we obtain by insertion of the dielectric tensor

$$\sum_{i=1}^3 \sum_{j=1}^3 \epsilon_{ij} E_i E_j = 2\rho_E = \text{const.} \quad (26-102)$$

This is the equation of an ellipsoid called the “Fresnel ellipsoid”. The fact that it is actually an ellipsoid becomes more obvious by the transformation into the coordinate system of the principal axes

$$\epsilon_{xx} E_x^2 + \epsilon_{yy} E_y^2 + \epsilon_{zz} E_z^2 = 2\rho_E = \text{const.} \quad (26-103)$$

Comparing this with the equation of an ellipse of the main axes a , b , and c

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1. \quad (26-104)$$

The main axes of the Fresnel ellipsoid are formed by the reciprocals of the principal refractive indices, i.e., by n_x^{-1} , n_y^{-1} , and n_z^{-1} (figure 26-15) which are proportional to the phase velocity of light.

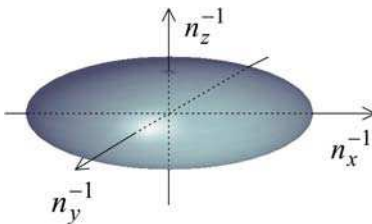


Figure 26-15: Fresnel ellipsoid.

26.6.4

Index Ellipsoid

Instead of obtaining the displacement current from the electric field by multiplication with the dielectric tensor we can instead solve the inverse problem, i.e., obtain the electric field by multiplication of the displacement current with the inverse of the dielectric tensor. This tensor can also be diagonalized and it, too, yields an ellipsoid, the so-called “index ellipsoid”.

In the coordinate system of the principal axes the equation follows from an inversion of the diagonalized tensor, which is, of course, just the inverse of the diagonal elements, and we obtain

$$\frac{D_x^2}{n_x^2} + \frac{D_y^2}{n_y^2} + \frac{D_z^2}{n_z^2} = 2\rho_E = \text{const.} \quad (26-105)$$

Here, the main axes are given directly by the principal refractive indices.

26.6.5

Types of Birefringent Media

The epsilon tensor must obey the symmetries of the interaction of light with matter.

Isotropic Media

If all three principal refractive indices are equal, the medium is optically isotropic:

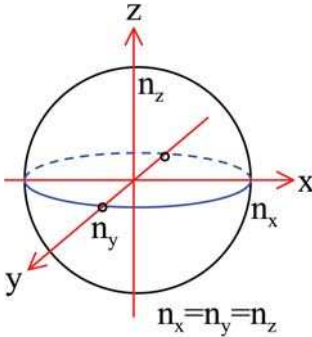


Figure 26-16: Index ellipsoid of isotropic media.

$$\epsilon = \begin{pmatrix} n^2 & 0 & 0 \\ 0 & n^2 & 0 \\ 0 & 0 & n^2 \end{pmatrix} = n^2 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (26-106)$$

The refractive index is independent of the polarization and therefore it is independent of the direction of propagation.

Isotropic media are media without internal order: Usually gases, fluids and amorphous solids like glass and fused silica. Any breaking of the symmetry such as anisotropic pressure applied to a solid body or the flow of a fluid may induce optical anisotropy.

Cubic crystals like CaF_2 and BaF_2 are isotropic in terms of the second-order dielectric tensor ϵ_{ij} , but the fourth-order dielectric tensor is non-zero, i.e., they show the phenomenon of spatial dispersion.

Uniaxial Media

Two different principal refractive indices, i.e., two axes of Fresnel's ellipsoid are identical:

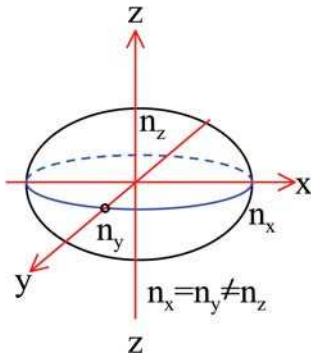


Figure 26-17: Index ellipsoid of uniaxial media.

$$\boldsymbol{\epsilon} = \begin{pmatrix} n_o^2 & 0 & 0 \\ 0 & n_o^2 & 0 \\ 0 & 0 & n_e^2 \end{pmatrix}. \quad (26-107)$$

The refractive index that appears twice is called the ordinary refractive index n_o and the refractive index that appears only once is called the extraordinary refractive index n_e . The index ellipsoid becomes an ellipsoid of revolution. These media are called uniaxial, because they possess one optical axis, i.e., one propagation direction for which the refractive indices of both eigenpolarizations are equal (see section 26.7.2). If $n_e > n_o$, the medium is called *positive uniaxial*, if $n_e < n_o$ it is called *negative uniaxial*.

Tetragonal, hexagonal and trigonal crystals are uniaxial.

Examples of negative uniaxial media:

Material	Crystal class	λ [nm]	Refractive indices	$\Delta n = n_e - n_o$
Calcite (Ca_2CO_3)	trigonal	589	$n_o = 1.6584$, $n_e = 1.4864$	-0.172
Tourmaline	tetragonal	589	$n_o = 1.669$, $n_e = 1.638$	-0.031
Beryll	hexagonal	589	$n_o = 1.570$, $n_e = 1.567$	-0.007

Examples of positive uniaxial media:

Material	Crystal class	λ [nm]	Refractive indices	Δn
Cinnabar	trigonal	589	$n_o = 2.854, n_e = 3.201$	0.347
Rutil	tetragonal	589	$n_o = 2.6158, n_e = 2.9029$	0.287
Quartz (SiO ₂)	trigonal	589	$n_o = 1.5442, n_e = 1.5533$	0.0091
Ice (H ₂ O)	hexagonal	589	$n_o = 1.309, n_e = 1.313$	0.004
MgF ₂	tetragonal	157	$n_o = 1.468, n_e = 1.482$	0.014

Biaxial Media

Three different principal refractive indices, i.e. all three axes of Fresnel's ellipsoid are different:

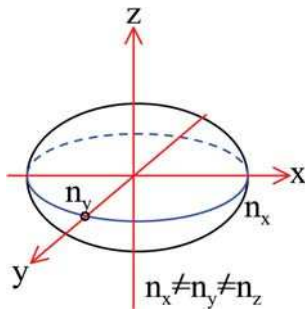


Figure 26-18: Index ellipsoid of biaxial media.

$$\epsilon = \begin{pmatrix} n_x^2 & 0 & 0 \\ 0 & n_y^2 & 0 \\ 0 & 0 & n_z^2 \end{pmatrix}. \quad (26-108)$$

These media are called biaxial, because they possess two optical axes (see section 26.7.2).

Triclinic, monoclinic and rhombic crystals are biaxial.

Examples:

Material	Crystal class	λ [nm]	n_x	n_y	n_z
Calcium sulfate (gypsum)	monoclinic	589	1.5208	1.5228	1.5298
Mica	monoclinic	589	1.5612	1.5944	1.5993
Sugar		589	1.5382	1.5658	1.5710

Optical Activity

Optical activity stems from the tensor of third order on the relation between the displacement and the electric field. For a plane wave of propagation constant $\vec{k}(k_1, k_2, k_3)$ the resulting dielectric tensor becomes

$$\varepsilon_{jl}^{\text{oa}} = i \sum_{m=1}^3 \gamma_{jlm} k_m = i\gamma_{jl}. \quad (26-109)$$

The optical activity is zero for crystals with a center of symmetry. It results in a rotation of the plane of polarization for an incident linearly polarized plane wave.

Because of the anti-symmetry of γ_{jlm} and in accordance with the hermiticity of ε , γ is antisymmetrical [26-13]

$$\gamma_{jl} = -\gamma_{ij}. \quad (26-110)$$

Because an antisymmetrical second-order tensor can be replaced by an axial vector according to

$$\vec{\gamma} = \begin{pmatrix} \gamma_{23} \\ \gamma_{31} \\ \gamma_{12} \end{pmatrix} = - \begin{pmatrix} \gamma_{32} \\ \gamma_{13} \\ \gamma_{21} \end{pmatrix}. \quad (26-111)$$

The relation between the displacement and the electric field becomes

$$\vec{D} = \varepsilon \vec{E} - i\vec{\gamma} \times \vec{E} \quad (26-112)$$

where γ is the “gyration vector” describing a rotation of \vec{E} .

Spatial Dispersion

Spatial dispersion of second order is described by a tensor of fourth order. For a plane wave of propagation constant $\vec{k}(k_1, k_2, k_3)$ the resulting dielectric tensor becomes

$$\varepsilon_{ij}^{\text{IBR}} = \sum_{l=1}^3 \sum_{m=1}^3 \alpha_{ijlm} k_l k_m. \quad (26-113)$$

The abbreviation IBR stands for “intrinsic birefringence”. The term stems from its appearance in optical lithography. Below a wavelength of 193 nm CaF₂ is almost the sole applicable material for refractive components. Being a cubic crystal it was firstly considered as isotropic. Polarimetric measurements at the HeNe laser line of 633 nm showed no effect apart from the stress-induced birefringence, so the industrial production of CaF₂ crystals began. First measurements of the birefringence at the target wavelength of 157 nm, however, proved that CaF₂ also shows birefringence stemming from spatial dispersion with a magnitude of 11 nm/cm [26-16], which is far above the specification for stress-induced birefringence. Because this birefringence is a property of the material and could not be eliminated by tighter specification it was called “intrinsic”.

Not all elements of the tensor α are independent. In fact, most values are equal because of symmetry considerations. For a cubic crystal, as in CaF₂, the symmetry provides the relations [26-17]

- a) $\alpha_{xxxx}=\alpha_{yyyy}=\alpha_{zzzz}$,
- b) $\alpha_{xxzz}=\alpha_{xxyy}=\alpha_{yyxx}=\alpha_{zzxx}=\alpha_{yyzz}=\alpha_{zzyy}$, (26-114)
- c) $\alpha_{xyxy}=\alpha_{yxyx}=\alpha_{xyyx}=\alpha_{yxyx}=\alpha_{xzxz}=\alpha_{zxzx}=\alpha_{xzzx}=\alpha_{zzxz}=\alpha_{zyyz}=\alpha_{yzyz}=\alpha_{zyyz}=\alpha_{yzzz}$.

The maximum birefringence appears along the 110-direction. It becomes

$$2n_0\Delta n = \frac{k_0^2}{2}(\alpha_{1111} - \alpha_{1122} - 2\alpha_{1212}) \quad (26-115)$$

where $k_0 = 2\pi/\lambda_0$ (λ_0 =free space wavelength) and n_0 denotes the mean refractive index.

26.7 Eigenwaves in Anisotropic Media

In this section the propagation of plane waves in anisotropic media is addressed. We show that a plane wave decomposes into two orthogonally polarized eigenwaves that usually propagate with different phase velocities, and we discuss their properties.

26.7.1 Plane Waves in Anisotropic Media

Consider a plane wave with propagation vector \vec{k} and arbitrary polarization in a source-free and non-magnetic, but otherwise arbitrary, anisotropic medium.

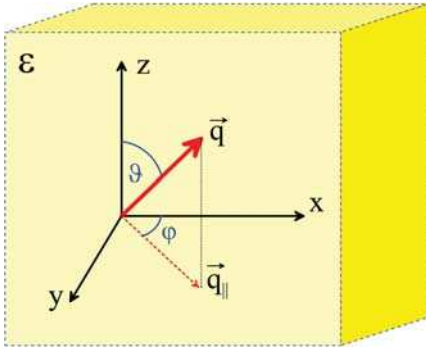


Figure 26-19: Propagation direction \mathbf{q} in coordinate system.

The direction of propagation is denoted by the vector \vec{q} , which is the normalized propagation vector (figure 26-19). It is defined by

$$\vec{k} = \frac{2\pi}{\lambda_0} n\vec{q}. \quad (26-116)$$

The Maxwell equations reduce for plane waves to a set of algebraic equations:

$$\begin{aligned}
 n\vec{q} \times \vec{E} &= Z_0 \vec{H}, \\
 n\vec{q} \times \vec{H} &= -Z_0^{-1} \vec{D}, \\
 \vec{q} \cdot \vec{D} &= 0, \\
 \vec{q} \cdot \vec{H} &= 0, \\
 \vec{D} &= \varepsilon \vec{E},
 \end{aligned}
 \tag{26-117}$$

where Z_0 denotes the free space impedance. The relation between the electric field and the displacement becomes

$$\vec{D} = -n^2 \vec{q} \times (\vec{q} \times \vec{E}) = -n^2 [(\vec{q} \cdot \vec{E}) \vec{q} - \vec{E}].
 \tag{26-118}$$

Inspection yields the following properties of a plane wave in an anisotropic medium:

- 1) \vec{D} and \vec{E} are not usually parallel.
- 2) \vec{D} is transverse, i.e., \vec{D} and \vec{q} are orthogonal.
- 3) \vec{E} is usually not transverse.
- 4) \vec{H} is transverse.
- 5) \vec{E} , \vec{D} and \vec{q} are coplanar.

The relative orientation of the fields is sketched in figure 26-20.

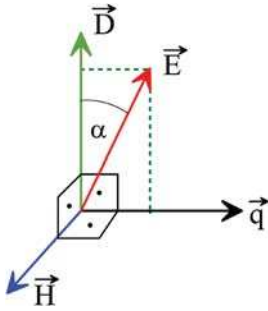


Figure 26-20: Relative orientation of \vec{E} , \vec{D} , \vec{H} , and \vec{q} .

From the non-transversality of the electric field follows the non-parallelism of the Poynting vector to the propagation direction of the plane wave. This fact and the resulting beam-splitting will be discussed in detail in section 26.10.

26.7.2

Eigenwaves and their Polarization

Inserting the tensor relation for \vec{D} into (26-118) yields an equation for \vec{E} alone

$$n^2 \vec{q} \times \vec{q} \times \vec{E} = -\varepsilon \vec{E}.
 \tag{26-119}$$

The cross-product with \vec{q} can be written in matrix form

$$\vec{q} \times \vec{A} = \begin{pmatrix} 0 & -q_z & q_y \\ q_z & 0 & -q_x \\ -q_y & q_x & 0 \end{pmatrix} \begin{pmatrix} A_x \\ A_y \\ A_z \end{pmatrix} = \mathbf{q} \vec{A}, \quad (26-120)$$

and we obtain a matrix equation for (26-119)

$$n^2 \mathbf{q}^2 \vec{E} = -\epsilon \vec{E}. \quad (26-121)$$

The solutions of this equation lead to the eigenpolarizations and the corresponding refractive indices. It can also be formulated in terms of the displacement vector \vec{D}

$$n^2 \mathbf{q}^2 \epsilon^{-1} \vec{D} = -\vec{D}. \quad (26-122)$$

For evaluation, two forms are useful:

- 1) An eigenvalue equation

$$\epsilon^{-1} \mathbf{q}^2 \vec{E} = -n^{-2} \vec{E}. \quad (26-123)$$

- 2) A homogeneous linear system of equations

$$(n^2 \mathbf{q}^2 + \epsilon) \vec{E} = 0. \quad (26-124)$$

In the coordinate system of the principal axes of the medium, the dielectric tensor ϵ is diagonal and the explicit form becomes

$$\begin{pmatrix} \epsilon_x - (q_y^2 + q_z^2) & n^2 q_x q_y & n^2 q_x q_z \\ n^2 q_x q_y & \epsilon_y - (q_x^2 + q_z^2) & n^2 q_y q_z \\ n^2 q_x q_z & n^2 q_y q_z & \epsilon_z - (q_x^2 + q_y^2) \end{pmatrix} \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \quad (26-125)$$

For a non-trivial solution the matrix must be singular, i.e.,

$$\det(n^2 \mathbf{q}^2 - \epsilon) = 0, \quad (26-126)$$

which yields, after some algebra, Fresnel's quartic equation

$$\begin{aligned} & \left(q_x^2 \epsilon_x + q_y^2 \epsilon_y + q_z^2 \epsilon_z \right) n^4 - \left[q_x^2 \epsilon_x (\epsilon_y + \epsilon_z) + q_y^2 \epsilon_y (\epsilon_x + \epsilon_z) + q_z^2 \epsilon_z (\epsilon_x + \epsilon_y) \right] n^2 \\ & + \epsilon_x \epsilon_y \epsilon_z = 0 \end{aligned} \quad (26-127)$$

It is of fourth degree in n and yields two solutions for n^2 . Another form of Fresnel's quartic equation is in terms of the phase velocity $v = c_0/n$

$$\sum_{i=1}^3 \frac{q_i^2}{v^2 - v_i^2} = 0, \quad (26-128)$$

where v_i are the phase velocities along the principal axes and q_i are the components of the propagation vectors in the coordinate system of the principal axes.

In the coordinate planes, i.e., for q_x , q_y or $q_z = 0$, Fresnel's quartic becomes particularly simple. In the yz -plane it is

$$q_x = 0 : \left[\varepsilon_x - n^2 (q_y^2 + q_z^2) \right] \left\{ (\varepsilon_y - q_z^2 n^2) (\varepsilon_z - q_y^2 n^2) - n^4 q_y^2 q_z^2 \right\} = 0 \quad (26-129)$$

which becomes with $q_y^2 + q_z^2 = 1$

$$q_x = 0 : (\varepsilon_x - n^2) \left\{ (\varepsilon_y - q_z^2 n^2) (\varepsilon_z - q_y^2 n^2) - n^4 q_y^2 q_z^2 \right\} = 0. \quad (26-130)$$

The analogous equations for the xy and xz -planes are obtained from a cyclic permutation of x , y , and z .

To satisfy the equation one of both brackets must be zero. The first one yields

$$\varepsilon_x = n^2 \quad (26-131)$$

i.e., the refractive index $n = \sqrt{\varepsilon_x}$ occurs for all propagation directions in the yz -plane. Because of the directional independency it is called the *ordinary index of refraction*.

The second bracket is equivalent to the equation of an ellipse

$$\frac{q_y^2 n^2}{\varepsilon_z} + \frac{q_z^2 n^2}{\varepsilon_y} = 1 \quad (26-132)$$

with the refractive index

$$n = \sqrt{\frac{\varepsilon_y \varepsilon_z}{\varepsilon_y q_y^2 + \varepsilon_z q_z^2}} \quad (26-133)$$

which depends on the direction of propagation and is therefore called the *extraordinary index of refraction*.

The existence of two refractive indices means that the Ewald sphere of a homogeneous medium becomes a two-leaf Ewald surface. Slices through the surface at the coordinate planes are plotted in figures 26-21 and 26-22.

While, for a uniaxial crystal, the complete 3D-surfaces are just an ellipsoid and a sphere that touch each other at two opposing points, they are somewhat more complex for a biaxial crystal. The inner surface bulges outwards at four points and the outer surface bulges inwards, so that both surfaces touch at four points [26-13].

The optical axes are the directions along which the refractive index is equal for both eigenpolarizations. Uniaxial media have one axis through the opposite touching points of the sphere for the ordinary, and the ellipsoid for the extraordinary index (cf. figure 26-21). For biaxial media two optical axes exist, one for each pair of opposite touching points of the inner and the outer sphere (figure 26-23).

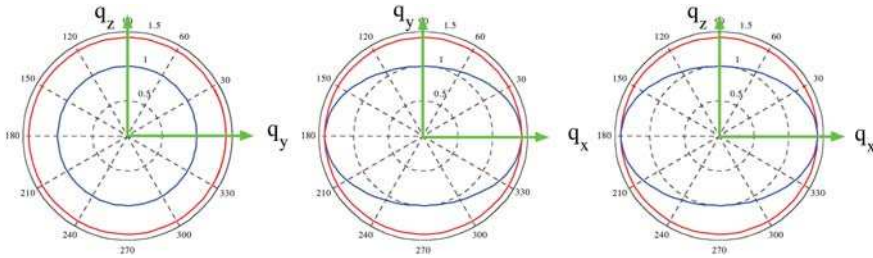


Figure 26-21: Refractive indices in coordinate planes for a uniaxial crystal with $\epsilon_x = 1$, $\epsilon_y = 2$, $\epsilon_z = 2$, red: ordinary, blue: extraordinary refractive index.

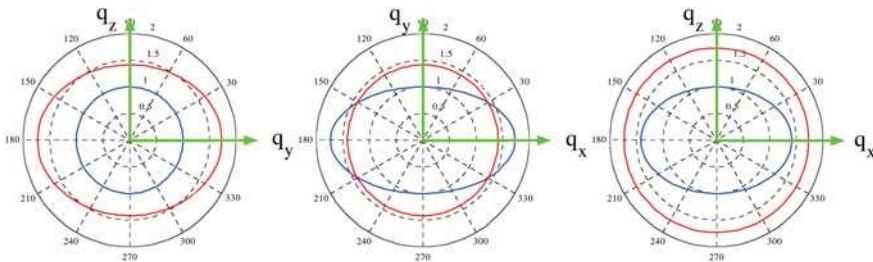


Figure 26-22: Refractive indices in coordinate planes for a biaxial crystal with $\epsilon_x = 1$, $\epsilon_y = 2$, $\epsilon_z = 3$, red: ordinary, blue: extraordinary refractive index.

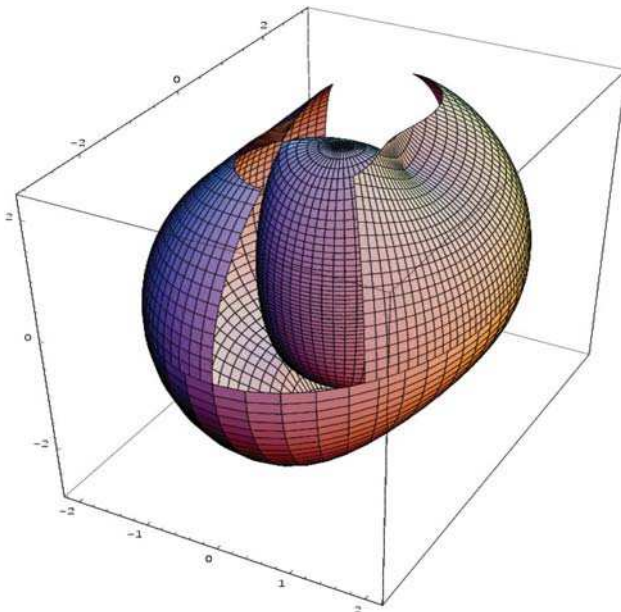


Figure 26-23: 3D-surface of refractive indices for a biaxial crystal.

26.7.3

Properties of the Eigenpolarizations

For a given propagation direction in an anisotropic medium two orthogonal polarized waves that propagate with different phase-velocities exist. The corresponding eigenpolarizations are linear and orthogonal when the matrix $\epsilon^{-1}\mathbf{q}^2$ is real and symmetrical. The polarizations are “eigenpolarizations”, because they are the states of polarization that do not change while the wave is propagating.

$$\vec{D}_j(\vec{q}, \vec{r}) = \vec{D}_j e^{in_j \frac{2\pi}{\lambda_0} \vec{q} \cdot \vec{r}} \quad (26-134)$$

The fields and the wave vectors obey the following relations

$$\begin{aligned} \text{a) } \vec{D}_{1,2} \text{ are transverse and orthogonal: } & \vec{D}_1 \cdot \vec{q} = \vec{D}_2 \cdot \vec{q} = 0, \vec{D}_1 \cdot \vec{D}_2 = 0, \\ \text{b) } \vec{E}_{1,2} \text{ are orthogonal: } & \vec{E}_1 \cdot \vec{E}_2 = 0, \\ \text{c) } \vec{H}_{1,2} \text{ are transverse and orthogonal: } & \vec{H}_1 \cdot \vec{q} = \vec{H}_2 \cdot \vec{q} = \vec{H}_1 \cdot \vec{H}_2 = 0. \end{aligned} \quad (26-135)$$

26.7.4

The Intersection Ellipse

The index ellipsoid provides a visual means of obtaining the refractive indices of a plane wave in a given propagation direction.

The phase velocities are proportional to the reciprocals of the corresponding refractive indices, which can be obtained by constructing the intersection ellipse. Consider a propagation direction \vec{q} . A plane that contains the center of the ellipsoid and that is orthogonal to the direction \vec{q} intersects the index ellipsoid with an elliptical curve (figure 26-24). The direction of the main axes of this ellipse yields the linear polarization vectors and their length the corresponding refractive indices. The intersection ellipse is parallel to the wavefront of the plane wave given by the propagation vector.

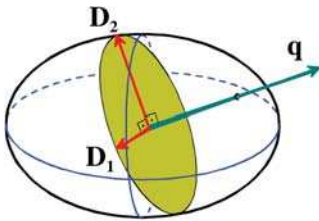


Figure 26-24: Index ellipsoid and construction of a pair of orthogonal \vec{D} -vectors for a given propagation direction.

For an isotropic medium, the intersection curve with a plane through the origin is always a circle. The long and the short axis of the intersection ellipse are identical and for every direction a single wave propagates with the same propagation velocity independent of the direction.

For a uniaxial medium, the long and short axis become different with one exception: If the propagation direction is parallel to the optical axis the ellipse reduces to a circle and the phase velocity is independent of the polarization. Therefore these me-

dia are called uniaxial: There is one axis along which the phase velocity is independent of the polarization. The maximum difference in the refractive indices occurs for a propagation direction which is perpendicular to the optical axis.

In general, the propagation direction encloses an angle ψ with the optical axis. One axis of the intersection ellipse is always equal to the ordinary refractive index while the other is given by

$$\frac{1}{n(\psi)^2} = \frac{\cos^2\psi}{n_o^2} + \frac{\sin^2\psi}{n_e^2} . \tag{26-136}$$

The eigenpolarization \vec{D}_e corresponding to the extraordinary refractive index is given by the projection of the optical axis onto the wavefront, i.e., \vec{D}_e, \vec{q} and the optical axis are in the same plane. The orthogonal eigenpolarization, corresponding to the ordinary refractive index, is orthogonal to this plane.

A point source inside a uniaxial crystal provides a spherical wave which is polarized perpendicular to the crystal axis, and an elliptical wave for the orthogonal polarization.

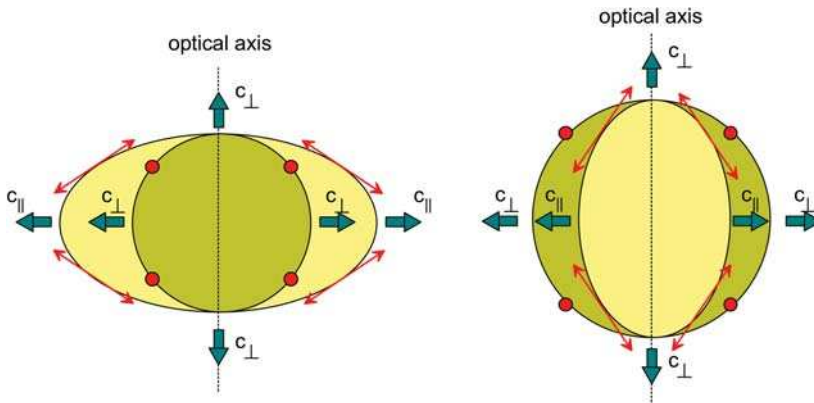


Figure 26-25: Point source in a uniaxial crystal. Left: negative uniaxial crystal, right: positive uniaxial crystal.

26.8 Jones Matrix of Propagation

The birefringence of anisotropic media is described by two eigenpolarizations together with the corresponding refractive indices. For non-absorbing media ϵ is hermitian and therefore the eigenpolarizations \vec{D}_1, \vec{D}_2 are orthogonal. Because of this, every polarized wave in the medium can be decomposed uniquely into a linear combination of both normalized eigenpolarizations. The propagation of the field \vec{D}_{in} over a distance d becomes

$$\vec{D}_{out} = (\vec{D}_{in} \cdot \hat{D}_1^*) \hat{D}_1 e^{ik_0 n_1 d} + (\vec{D}_{in} \cdot \hat{D}_2^*) \hat{D}_2 e^{ik_0 n_2 d} \tag{26-137}$$

which can be written in compact form in Jones matrix notation

$$\vec{D}_{\text{out}} = \begin{pmatrix} |D_{1x}|^2\phi_1 + |D_{2x}|^2\phi_2 & D_{1x}D_{1y}^*\phi_1 + D_{2x}D_{2y}^*\phi_2 \\ D_{1x}^*D_{1y}\phi_1 + D_{2x}^*D_{2y}\phi_2 & |D_{1y}|^2\phi_1 + |D_{2y}|^2\phi_2 \end{pmatrix} \begin{pmatrix} D_x^{\text{in}} \\ D_y^{\text{in}} \end{pmatrix} \quad (26-138)$$

with

$$\phi_i = e^{ik_0 n_i d}. \quad (26-139)$$

A concise form of the Jones matrix is obtained by using the dyadic product °

$$\mathbf{J} = \vec{D}_1 \circ \vec{D}_1^* \phi_1 + \vec{D}_2 \circ \vec{D}_2^* \phi_2. \quad (26-140)$$

By this means the Jones matrix of a series of different anisotropic media can be computed (however, only the effect due to the medium is considered whereas interfaces yielding double refraction and Fresnel effects are not included).

26.9

Jones Matrices of Propagation for Common Media

26.9.1

Eigenpolarizations and -values

The angular dependence of the eigenpolarizations and -values is computed from the Jones matrix of propagation via a solution of the eigenvalue equation

$$\mathbf{J}\vec{D}_{1,2} = \Lambda\vec{D}_{1,2} \quad (26-141)$$

with the matrix Λ of the eigenvalues

$$\Lambda = \begin{pmatrix} a_1 e^{i\varphi_1} & 0 \\ 0 & a_2 e^{i\varphi_2} \end{pmatrix}. \quad (26-142)$$

The retardation is the phase difference of the eigenvalues

$$\Delta\varphi = \varphi_1 - \varphi_2 \quad (26-143)$$

and the orientation is given by the orientation of the slow axis, i.e., the orientation of the eigenvector with the higher phase value. If the eigenvector corresponds to an elliptical polarization, the orientation of the main axis is taken.

26.9.2

Coordinate Systems

For correct computation of the change of the state of polarization of a ray traversing an anisotropic medium, three coordinate systems have to be considered: The ray coordinate system, the crystal coordinate system and the global coordinate system.

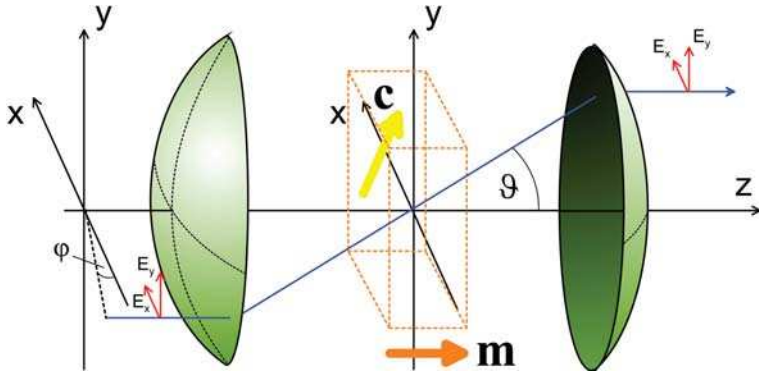


Figure 26-26: Coordinate system for the Jones matrix of propagation of a crystal. *c* = crystal axis.

The angular distributions of the Jones matrices of the following examples are shown for $\vartheta = 0 - 90^\circ$ and $\varphi = 0 - 2\pi$ (cf. figure 26-26). The direction of the ray along the optical axis in the crystal is denoted by its Miller indices (m_x, m_y, m_z) for cubic crystals. For uniaxial and biaxial media (m_x, m_y, m_z) denote the directions in the coordinate system of the principal indices.

26.9.3

Uniaxial Crystal

Consider a uniaxial crystal with a mean refractive index of 1.55840 and a maximum birefringence of 10^{-6} . The principal refractive indices become

$$\begin{aligned} n_1 &= 1.558401, \\ n_2 &= 1.558400, \\ n_3 &= 1.558400. \end{aligned}$$

The angular distributions of the Jones matrices are shown in figure 26-27.

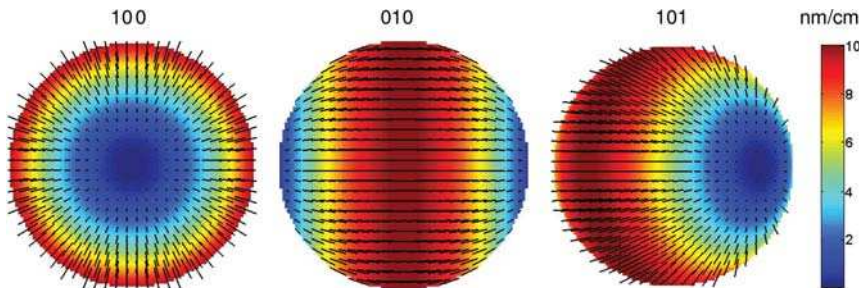


Figure 26-27: Birefringence distribution of a uniaxial crystal.

Looking along the optical axis the eigenpolarizations are radial with azimuthal symmetry. This case is shown on the left. Looking in a direction orthogonal to the crystal axis (plot in the center), the maximum retardance is observed. Please note that the eigenpolarizations are not parallel to each other. They are inclined for increasing height.

26.9.4

Biaxial Crystal

Consider a biaxial crystal with a mean refractive index of 1.55840 and a maximum birefringence of 10^{-6} and 2×10^{-6} . The principal refractive indices become

$$\begin{aligned} n_1 &= 1.558402, \\ n_2 &= 1.558401, \\ n_3 &= 1.558400. \end{aligned}$$

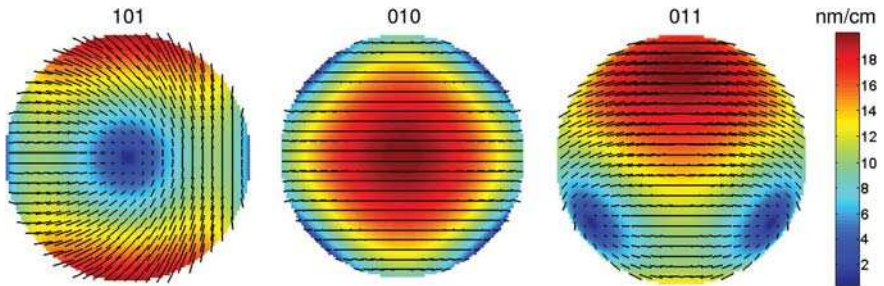


Figure 26-28: Birefringence distribution of a biaxial crystal.

The two optical axes are orthogonal to each other.

Principal refractive indices of

$$\begin{aligned} n_1 &= 1.5584015, \\ n_2 &= 1.5584010, \\ n_3 &= 1.5584000. \end{aligned}$$

yield a similar pattern, but now the optical axes are not orthogonal.

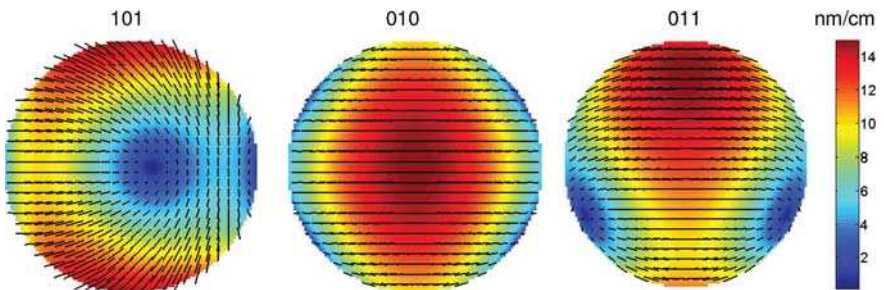


Figure 26-29: Birefringence distribution of a biaxial crystal.

26.9.5

CaF₂ with Spatial Dispersion at $\lambda = 193$ nm

The spatial dispersion of calcium fluoride yields a distribution of eigenpolarizations that is considerably more complicated. The images below show the retardances and eigenpolarizations of CaF₂ at a wavelength of 193 nm. The birefringence is $-0.34 \cdot 10^{-6}$. Along seven directions the retardance becomes zero, i.e., we may speak of a hepta-axial crystal.

For the use of the material in optical lithography the 100 and the 111 directions are of particular importance, because there the retardance for the ray parallel to the optical axes vanishes. That means usually that in the center of the pupil the resulting retardance becomes zero. A correct combination of material of different crystal cuts (different directions of axial rays) may also reduce the retardance for oblique rays down to a level that is acceptable for high precision optics as is required for lithography.

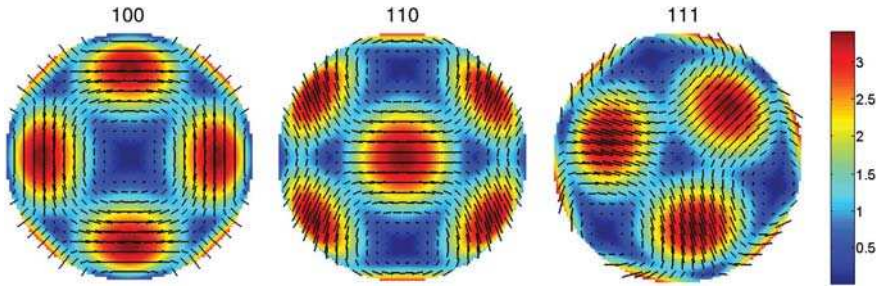


Figure 26-30: Birefringence distribution of CaF₂ (a crystal with spatial dispersion).

26.10
Beam-splitting in an Anisotropic Medium

The direction of the energy flow for a plane wave can be represented by the direction of the Poynting vector:

$$\vec{S} = \vec{E} \times \vec{H}. \quad (26-144)$$

Because \vec{E} , \vec{D} and \vec{q} are coplanar and $\vec{H} \times \vec{q} = 0$, \vec{S} is located in the same plane, i.e., all four vectors \vec{E} , \vec{D} , \vec{S} , and \vec{q} are coplanar. Their orientation is sketched in figure 26-31.

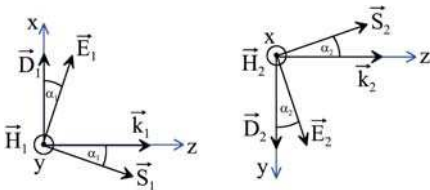


Figure 26-31: Relative orientation of the electromagnetic field vectors of the eigenwaves. Note the coplanarity of \vec{E} , \vec{D} , \vec{S} and \vec{k} .

Consequences:

- a) The direction of the energy flow is not parallel to the wave direction. It encloses the same angle α that exists between the electric vector and the displacement vector, i.e.

$$\cos \alpha = \frac{\vec{E} \cdot \vec{D}}{|\vec{E}| |\vec{D}|}. \quad (26-145)$$

In the coordinate system of the principal axes we get

$$\cos \alpha = \frac{n_x^2 E_x^2 + n_y^2 E_y^2 + n_z^2 E_z^2}{\sqrt{(E_x^2 + E_y^2 + E_z^2)(n_x^4 E_x^2 + n_y^4 E_y^2 + n_z^4 E_z^2)}}. \quad (26-146)$$

- b) Although, for the two eigenwaves, for a given direction of propagation the wave vectors are parallel (because the direction is identical), the corresponding Poynting vectors $\vec{S}_{1,2}$, are not, because the E -vectors $\vec{E}_{1,2}$ usually point in different directions. Therefore, there is an inherent beam split during the propagation. For the example of a uniaxial crystal above, the energy propagation of the ordinary wave is parallel to the plane-wave propagation. Therefore the beam split is already obtained from eq. (26-146).

Example

In order to quantify the maximum beam split for a given birefringence we consider a uniaxial crystal with the principal refractive indices $n_o = n_x = n_z$ and $n_e = n_y$ and an xz -polarized wave propagating under 45° to the optical axis (z -direction). For a weak birefringence the electric field is approximately $\vec{E} = (1, 0, 1)$ and we obtain from eq. (26-146)

$$\cos \alpha = \frac{n_e^2 + n_o^2}{\sqrt{2(n_e^4 + n_o^4)}} \quad (26-147)$$

which results for CaF_2 with $n_o = 1.56$ and $n_e = n_o + 10^{-6}$ at 157 nm to $\alpha = 3.7 \cdot 10^{-5}^\circ$. For CaCO_3 in the visible range of the spectrum ($\lambda = 546$ nm) we get, with $n_o = 1.662$ and $n_e = 1.488$, a beam split of $\alpha = 6.3^\circ$ resulting in a ray displacement of 1.1 mm for a traversed distance of 1 cm.

Rays as Plane-wave Bundles

Besides the fact that the beam split results merely in a displacement, it is not necessary to treat the beam split of rays independently: The beam split follows from the representation of a beam as a bundle of plane waves.

To demonstrate this fact, consider a beam that is represented as a bundle of plane waves into different directions \vec{q}_Ω ,

$$U = \int_{\Delta\Omega} U(\Omega) e^{i n(\Omega) \vec{q} \cdot \vec{r}} d\Omega \quad (26-148)$$

where Ω denotes the spatial angle. Now we argue as follows:

- a) The propagation of the beam is equivalent to the propagation of the point of maximum interference.
- b) The spatial propagation of the point of maximum interference follows the direction where the phase velocity of all contributing plane waves are equal,

$$n(\Omega)\vec{q} \cdot \vec{S} = \text{const. in } \Omega, \quad (26-149)$$

i.e., a direction of stationary phase

$$\frac{d\vec{q}_\Omega}{d\Omega} \cdot \vec{S} = 0 \quad (26-150)$$

with

$$\vec{q}_\Omega = n(\Omega)\vec{q}. \quad (26-151)$$

- c) Therefore, the direction of energy propagation \vec{S} is perpendicular to the angular derivative of the phase-propagation vector.
- d) For an isotropic medium, $d\vec{q}_\Omega$ is perpendicular to the plane-wave direction with the consequence that the directions of energy and plane-wave propagation are parallel.
- e) In an anisotropic medium, the phase velocity depends on the direction of propagation, i.e., $d\vec{q}_\Omega$ is not perpendicular to \vec{q} and therefore \vec{S} is not parallel to \vec{q} .

To illustrate this continuous beam displacement, we simulated the propagation of a beam consisting of plane waves with an angular spectrum within $\pm 5^\circ$ where the refractive index changes by $\Delta n = -0.1$ around a mean index of 1.56, i.e.,

$$U = \sum_{j=-M/2}^{M/2} \exp \left\{ i2\pi \left(n_0 + \Delta n \left(\sin \left(\frac{2j\phi_{\max}}{M} \right) x + \cos \left(\frac{2j\phi_{\max}}{M} \right) z \right) \right) \right\}. \quad (26-152)$$

The result in form of the real part of the scalar field U is shown in figure 25.23. The angle between the wave propagation and the continuous maximum displacement is 20° .

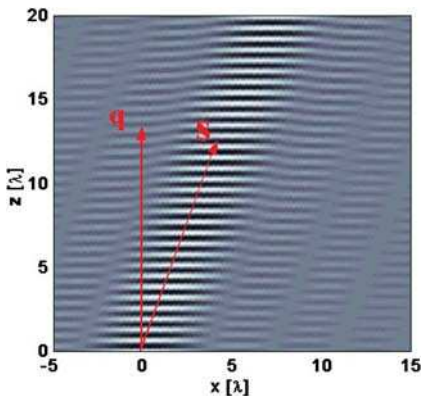


Figure 26-32: Simulation of the beam displacement in an anisotropic medium. The plotted quantity is the real part of the field.

Double Refraction

The split of the beams, i.e., the slit in the direction of energy propagation, should be distinguished carefully from the double refraction at interfaces. While the beam split already occurs for the bulk medium, the double refraction is caused by the different angles of refraction at an interface.

The phenomenon of double refraction for an arbitrary anisotropic medium is depicted in figure 26-33: A plane wave incident under the angle ϑ_i onto the interface is refracted into two plane waves with propagation vectors $\vec{k}_1 = n_1(\vartheta_1)\vec{q}_1$ and $\vec{k}_2 = n_2(\vartheta_2)\vec{q}_2$.

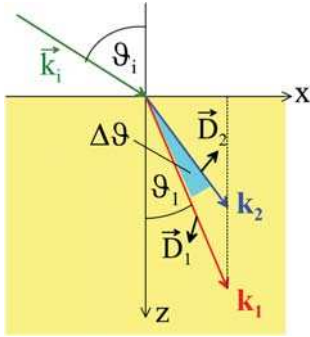


Figure 26-33: Double refraction at an interface to an anisotropic medium.

In order to obtain the angular split and phase difference of both waves after a length z of propagation, the angles and the refractive indices are referred to one wave, i.e., we write

$$\begin{aligned}\vartheta_2 &= \vartheta_1 + \Delta\vartheta = \vartheta + \Delta\vartheta, \\ n_2(\vartheta_2) &= n_1(\vartheta_1) + \Delta n = n + \Delta n.\end{aligned}\quad (26-153)$$

The angular split is calculated from

$$\sin \vartheta_i = n \sin \vartheta = (n + \Delta n) \sin (\vartheta + \Delta\vartheta) \quad (26-154)$$

yielding

$$\begin{aligned}\Delta\vartheta &= \arcsin\left(\frac{n \sin \vartheta}{n + \Delta n}\right) - \vartheta \\ &= \arcsin\left(\frac{\sin \vartheta_i}{n + \Delta n}\right) - \arcsin\left(\frac{\sin \vartheta_i}{n}\right).\end{aligned}\quad (26-155)$$

Neglecting terms of quadratic order (Δn^2 , $\Delta\vartheta^2$, $\Delta n \Delta\vartheta$) we obtain for the angular split

$$\Delta\vartheta \approx -\frac{\Delta n}{n} \tan \vartheta. \quad (26-156)$$

In particular, for a plane parallel plate, the angular split of the waves occurs for a plane only within the material. Outside the waves are again parallel, because, for any refractive index, a plane plate introduces only a beam displacement (figure 26-34)

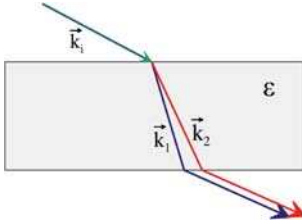


Figure 26-34: Parallel beams behind a birefringent parallel plate.

Concluding this chapter we compute the retardance that occurs at a given distance behind the refracting interface of an anisotropic medium.

The retardance should be referred to a certain point $r = (x, z)$ in space

$$\Delta\varphi = (\vec{k}_2 - \vec{k}_1) \cdot \vec{r} = (k_{2z} - k_{1z})z + (k_{2x} - k_{1x})x. \quad (26-157)$$

Because of the continuity of the tangential fields, the tangential component of the propagation vector is continuous while traversing the interface, i.e.,

$$(k_{2x}, k_{2y}) = (k_{1x}, k_{1y}) \quad \text{and} \quad (k_{1x}, k_{1y}) = (k_{ix}, k_{iy}). \quad (26-158)$$

This implies

$$(k_{2x}, k_{2y}) = (k_{1x}, k_{1y}) \quad (26-159)$$

and we obtain for the retardance

$$\begin{aligned} \Delta\varphi &= (k_{2z} - k_{1z})z \\ &= [(n + \Delta n) \cos(\vartheta + \Delta\vartheta) - n \cos(\vartheta)]z. \end{aligned} \quad (26-160)$$

Again neglecting terms of quadratic order we get

$$\Delta\varphi \approx [\Delta n \cos(\vartheta) - n\Delta\vartheta \sin(\vartheta)]z. \quad (26-161)$$

Inserting the approximate formula for the angular split $\Delta\vartheta$ yields finally

$$\Delta\varphi \approx \frac{\Delta n}{\cos(\vartheta)} z = \Delta n L \quad (26-162)$$

where $L = z/\cos\vartheta$ denotes the geometrical path length. This is an equation we expect if both polarizations traverse the same path but are subject to different refractive indices. It implies that we may neglect the angular split for small Δn and $\Delta\vartheta$ but still obtain valid retardances.

26.11

Examples of Polarization-optical Elements

26.11.1

Quarter-wave and Half-wave Retarder

A quarter-wave retarder (quarter-wave plate) with a retardance of $\pi/2$ and a half-wave retarder with a retardance of π , usually consists of a plate of a birefringent crystal (calcite, quartz, mica).

The thickness of an m^{th} -order quarter-wave retarder (where n_o and n_{ao} are the ordinary and the extraordinary refractive indexes, respectively) is

$$d = (4m - 3) \cdot \frac{\lambda}{4 \cdot (n_o - n_{ao})}.$$

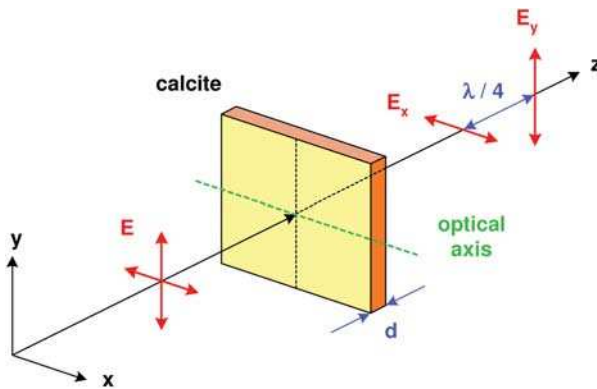


Figure 26-35: Quarter-wave retarder.

Higher order plates are fabricated from two plates of which the thickness is one order apart. Compared to a zero-order plate ($m = 1$) they have a couple of disadvantages:

1. higher sensitivity for thickness variations,
2. higher sensitivity for temperature variations,
3. higher sensitivity for angular variations.

26.11.2

Babinet–Soleil Compensator

The Babinet–Soleil compensator is a combination of two prisms made from a uniaxial crystal with orthogonal optical axes. The wedge angle is small so that the double refraction at the interface is negligible.

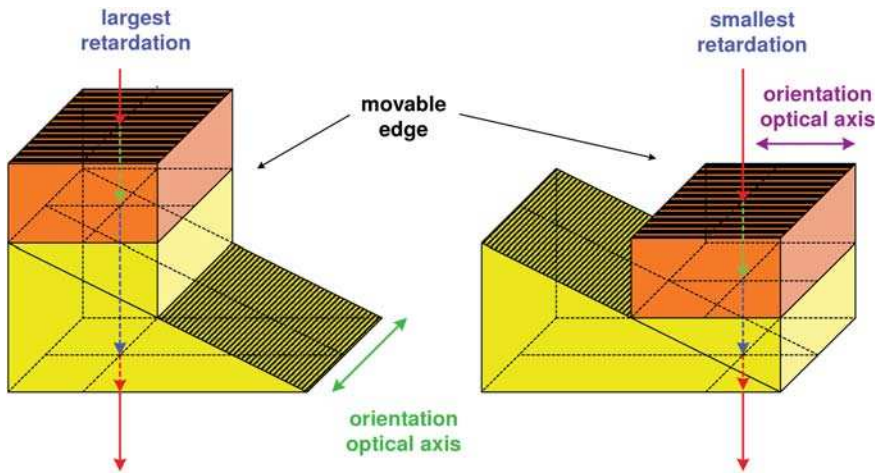


Figure 26-36: Principle of the Babinet-Soleil compensator.

A lateral movement changes the difference in the optical path length for the ordinary and the extraordinary refractive index:

$$\Delta\varphi = k_0(n_e - n_o)(d_1 - d_2).$$

By this means the retardance can be adjusted, i.e., the element is a variable retarder.

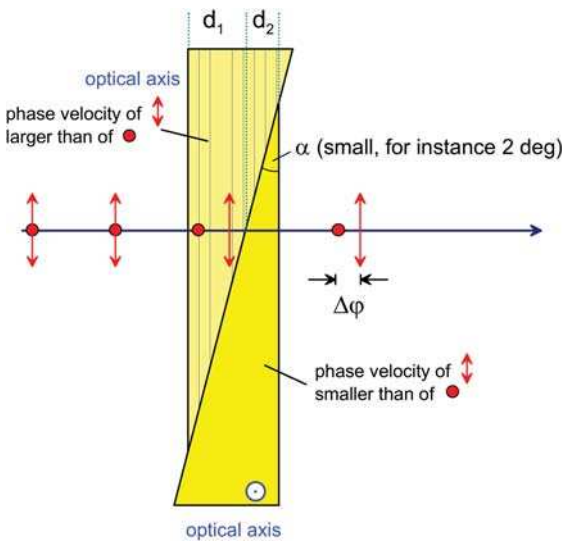


Figure 26-37: Retardance of the Babinet-Soleil compensator.

Faraday Rotator

An optical rotator can be realized using the Faraday effect: An external magnetic field rotates the plane of vibration of a traversing electromagnetic field proportional to the external field B and the traversed length L . The constant of proportionality is the Verdet constant of the material (figure 26-38)

$$\beta = |\vec{B}| \cdot L \cdot V.$$

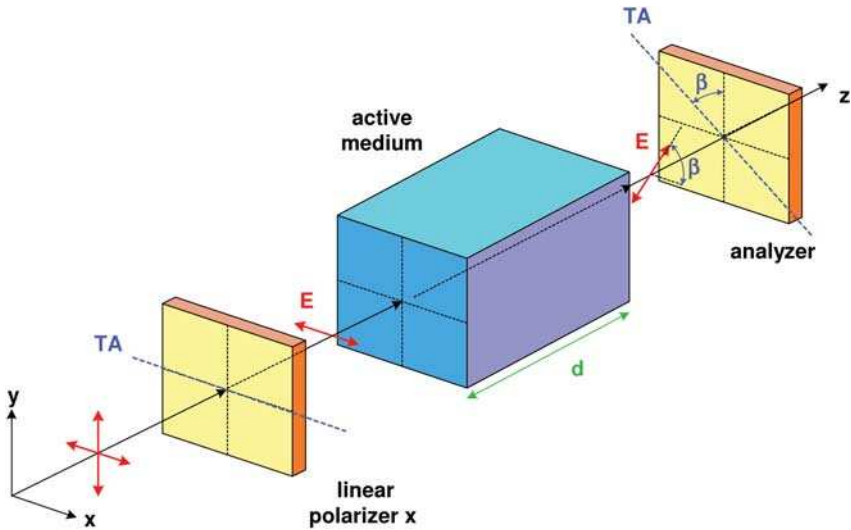


Figure 26-38: Parallel beams behind a birefringent parallel plate.

The direction of rotation induced by the Faraday effect is independent of the direction of propagation of the traversing light. Because of this the angle of rotation increases if a light ray traverses the element after reflection a second time. This property is applied in an optical isolator: It consists of a Faraday cell that rotates a linear input-polarization about 45 degrees. The cell is situated between two polarizers rotated about 45 degrees with respect to each other (figure 26-39).

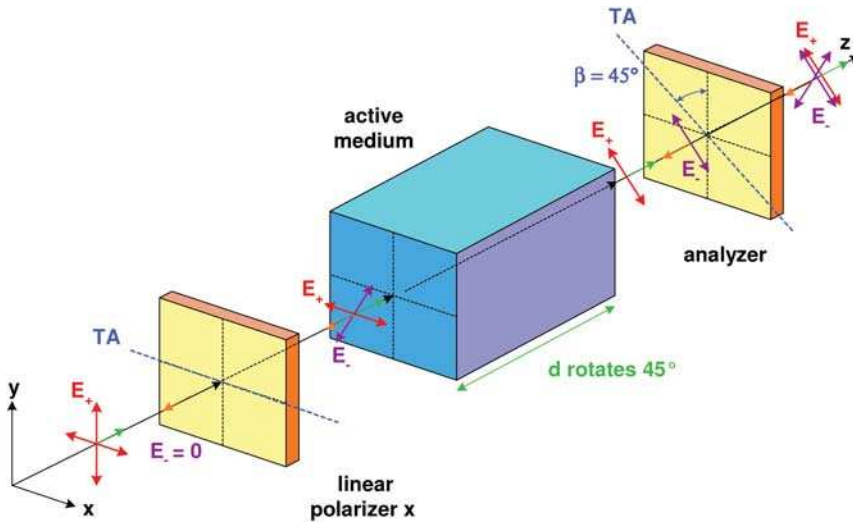


Figure 26-39: Optical isolator.

As an optical activity, the Faraday effect can be interpreted as circular birefringence. However, unlike the optical activity of passive chiral materials the Faraday effect remains its helicity even if the direction of propagation is reversed. Because of this an optical isolator can be realized only with the Faraday effect.

26.11.4

Brewster Plate

A Brewster plate of angle $\tan(\theta) = n$ has a Jones matrix

$$J_{BP}(n) = \begin{pmatrix} \left(\frac{2n}{n^2 + 1}\right)^2 & 0 \\ 0 & 1 \end{pmatrix}.$$

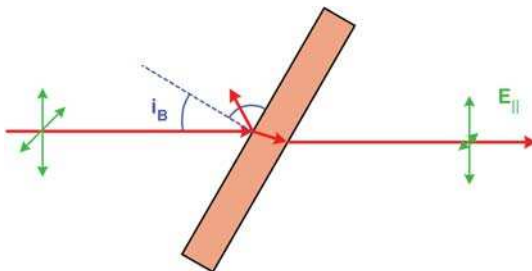


Figure 26-40: Brewster plate.

A stack of m plates in Brewster configuration has the Jones matrix

$$J_{SBP}(n) = \begin{pmatrix} \left(\frac{2n}{n^2+1}\right)^{2m} & 0 \\ 0 & 1 \end{pmatrix}.$$

The degree of polarization that can be obtained using such a stack is

$$p = \frac{(n^2+1)^{4m} - (2n)^{4m}}{(n^2+1)^{4m} + (2n)^{4m}}.$$

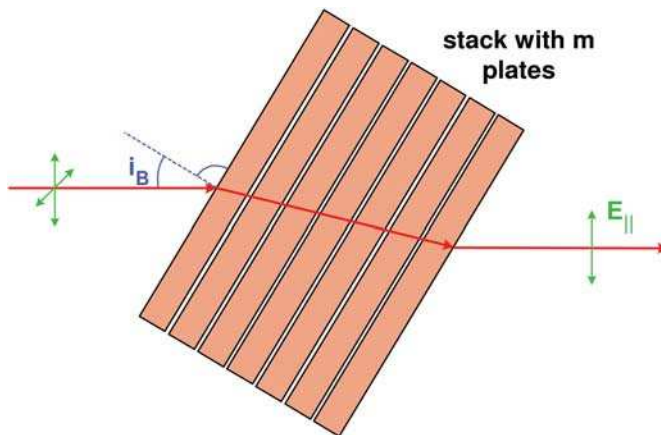


Figure 26-41: Stack of Brewster plates.

26.12

Literature

- 26-1 E. L. Malus, *Mém. Soc. d'Arcueil* 2, 143, 149 (1808–1809). *Mém. Savants étrangers* 2, 303 (1810–1811). *Théorie de la double refraction* (Paris, 1810).
- 26-2 C. Huygens, *Traité de la Lumière*, Leiden, 1690.
- 26-3 R.C. Jones, "A new calculus for the treatment of optical systems: I. Description and discussion of the calculus", *J. Opt. Soc. Am.* **31** (1941) 488–493.
- 26-4 M. Born, E. Wolf, *Principles of Optics*, 6th ed. (Pergamon Press, 1980).
- 26-5 L. Mandel, E. Wolf, *Optical Coherence and Quantum Optics* (Cambridge University Press, 1995) p. 352, p. 362.
- 26-6 R. M. A. Azzam, N. M. Bashara, *Ellipsometry and Polarized Light* (North-Holland, Amsterdam, 1977) p. 60f.
- 26-7 Sudha, A. V. Gopala Ruo, "Polarization elements: a group theoretical study", *J. Opt. Soc. Am. A* **18** (2001) 3130–3134.
- 26-8 V. Bagini, R. Borghi, F. Gori, M. Santriero, F. Frezzat, G. Schettini, G.S. Spagnolo, "The Simon-Mukunda polarization gadget", *Eur. J. Phys.* **17** (1996) 279–283.
- 26-9 R. Barakat, "Jones matrix equivalence theorems for polarization theory", *Eur. J. Phys.* **19** (1998) 209–216.

- 26-10** S-Y.L. Lu, R.A. Chipman, "Homogeneous and inhomogeneous Jones matrices", *J. Opt. Soc. Am. A* **11** (1994) 766–773.
- 26-11** R. Barakat, "Bilinear constraints between elements of the 4×4 Mueller-Jones transfer matrix of polarization theory", *Opt. Commun.* **38** (1981) 159–161.
- 26-12** W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, "Numerical Recipes in C", 2nd Ed. (Cambridge University Press, 1994), p. 59 ff.
- 26-13** A. Sommerfeld, *Vorlesungen über Theoretische Physik*, Vol. 4: Optik, Reprint of the 2nd edition from 1958 (Verlag Harri Deutsch, Frankfurt, 1978) p. 118.
- 26-14** S. Huard, *Polarization of Light*, (Wiley, New York, 1997).
- 26-15** C. Brosseau, *Fundamentals of Polarized Light* (Wiley, New York, 1998).
- 26-16** J. H. Burnet, Z. H. Levine, E. L. Shirley, J. H. Bruning, "Symmetry of spatial-dispersion induced birefringence and its implications for CaF₂ ultraviolet optics", *J. Microolithgr. Microfabr. Microsyst.* **1** (2002) 213–224.
- 26-17** V. M. Agranovich, V. L. Ginzburg, *Crystal Optics with Spatial Dispersion and Excitons* Springer Series in Solid-State Sciences Vol. 42, (Springer, Berlin, 1984), S. 129ff.
- 26-18** J. J. Gil, E. Bernabeu, "Obtainment of the polarization and retardation parameters of a non-depolarizing optical system from the polar decomposition of its Mueller matrix", *Optik* **76** (1987) 67–71.

27

Vector Diffraction

- 27.1 Introduction 524
- 27.2 Focus Computation for Polarized Fields 525
 - 27.2.1 Geometry for Focus Computation 525
 - 27.2.2 Richards–Wolf integral 526
 - 27.2.3 Plane Wave Expansion 531
 - 27.2.4 Focus Fields for Various Input Polarizations 533
- 27.3 Vector Kirchhoff Diffraction Integral 538
- 27.4 Analytical Solutions 538
 - 27.4.1 Plane Interface: Fresnel's Equations 540
 - 27.4.2 Diffraction at a Circular Cylinder 542
 - 27.4.3 Mie Scattering 547
- 27.5 Numerical Methods for Arbitrary Structures 553
- 27.6 Coupled Dipole Method 553
- 27.7 Integral Equation Approach and Moment Method 555
 - 27.7.1 The Moment Method 555
 - 27.7.2 Form of Scattering Operator 556
 - 27.7.3 Scattering in Three-layer Medium 557
- 27.8 Fourier Modal Method 563
 - 27.8.1 Theory 563
 - 27.8.2 Diffraction Efficiency 568
- 27.9 Finite-difference Method 568
 - 27.9.1 Boundary Conditions 570
 - 27.9.2 Implicit Paraxial Wave Equation in Two Dimensions 572
 - 27.9.3 Paraxial Wave Equation in Cylindrical Coordinates 572
 - 27.9.4 ADI-formulation of the Paraxial Wave Equation in Three Dimensions 575
 - 27.9.5 Split-step-beam Propagation Method 576
- 27.10 Rigorous Diffraction in Optical Imaging 579
 - 27.10.1 Dielectrics and Metals 579
- 27.11 Simulation of Polarized Imaging by use of Rigorous Diffraction 583
- 27.12 Literature 587

27.1

Introduction

In the field of classical optics the general diffraction problem is an electromagnetic one. Given an incident electromagnetic field $(\vec{E}_i(r), \vec{H}_i(r))$ and an inhomogeneous medium described by a space-variant permittivity $\epsilon(r)$, permeability $\mu(r)$ or conductivity $\sigma(r)$ the task is to find the resulting field distribution in the near and far-field of the medium (Figure 27-1).

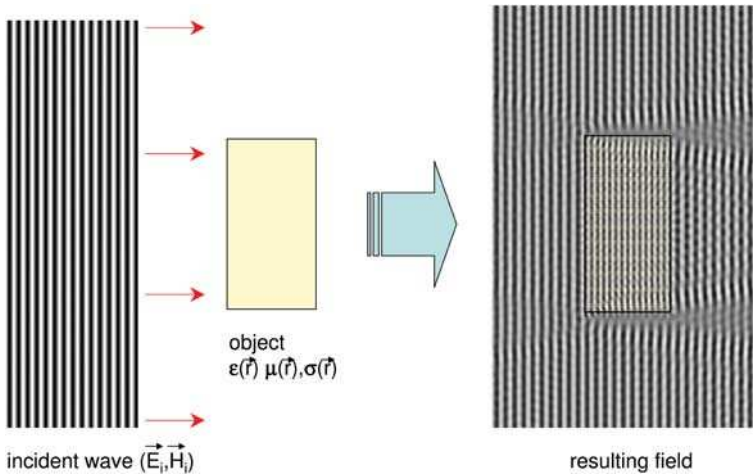


Figure 27-1: General diffraction problem.

This requires the solution of Maxwell's equations for an inhomogeneous medium. Three options are available:

1. Approximative solutions.
2. Analytical solutions.
3. Numerical solutions.

Approximative methods were discussed in detail in chapter 18. The classical scalar diffraction integrals do not solve a boundary condition problem but assume boundary values from which the diffracted field is computed. This method can be extended to the electromagnetic field in a straightforward manner. We will show here the computation of the polarized focus in some detail and discuss the vector Kirchhoff diffraction integral.

Analytical solutions are possible only for a very limited number of cases where the surface of the diffracting object coincides with a coordinate surface. Then the field tangential to the surface can be separated from the field normal to it, allowing a direct solution of the field inside and outside the diffracting object. The homogeneous half-space (solved by the well known Fresnel equations), the circular cylinder, and the sphere (Mie scattering) will be presented in some detail. Analytic solutions

are also possible for perfect conductors like Sommerfeld's diffraction at the perfectly conducting half-plane.

The importance of numerical solutions has shown a remarkable increase recently due to the advances in computation capabilities. Methods like the rigorous coupled wave analysis (RCWA; and also FMM or Fourier Modal Method) and the Finite Difference Time Domain Method (FDTD) allow a rigorous solution of Maxwell's equations by transforming the diffraction problem into a discretized version suitable for a computer. These methods are now the workhorse of electromagnetic diffraction calculations. Their description will close this chapter.

27.2

Focus Computation for Polarized Fields

The focus computation is an actual boundary value problem. The field in the exit pupil of an optical system is given and we search the resulting field distribution in the focus. The system aspect including aberrations of this computation will be discussed in detail in chapter 28. Here we compute the focus for a perfectly focussed field.

27.2.1

Geometry for Focus Computation

To compute the focus of electromagnetic fields we consider an optical system with a plane entrance pupil and a spherical exit pupil (Figure 27-2). The field distribution in the entrance pupil is described by a distribution of Jones vectors in pupil coordinates. Two-dimensional Jones vectors are sufficient because the electromagnetic field is transverse.

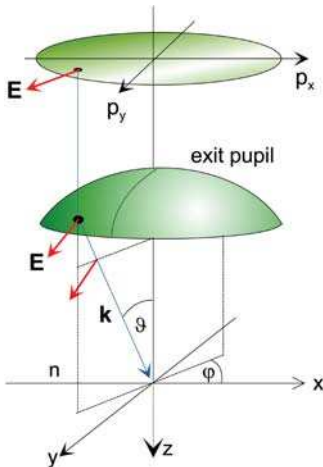


Figure 27-2: Geometry for high-NA focussing.

It should be noted that there is a certain contradiction in this argument: If we allow a phase distribution of the electric field in the entrance pupil, the direction of propagation cannot be purely axial. For vector fields this results in a non-vanishing z -component, but for macroscopic pupils (with diameter of at least some mm) and phase errors around a few wavelengths, this component is vanishingly small.

Two-dimensional, dimensionless pupil coordinates are denoted by (p_x, p_y) which are related to the propagation angles of the corresponding plane wave in image space by

$$\begin{aligned} p_x &= n \sin \vartheta \cos \varphi, \\ p_y &= n \sin \vartheta \sin \varphi. \end{aligned} \quad (27-1)$$

with the refractive index n in image space and

$$\sqrt{p_x^2 + p_y^2} \leq n \sin \vartheta = NA. \quad (27-2)$$

The inverse relations are

$$\begin{aligned} \sin \vartheta &= \frac{1}{n} \sqrt{p_x^2 + p_y^2}, \\ \sin \varphi &= \frac{p_y}{\sqrt{p_x^2 + p_y^2}}, \\ \cos \varphi &= \frac{p_x}{\sqrt{p_x^2 + p_y^2}}. \end{aligned} \quad (27-3)$$

27.2.2

Richards–Wolf integral

The computation of the focus fields can be done along the lines of the classical paper by Richards and Wolf [27-1] and [27-2]. The field components of the point-spread function follows from the Debye integral for an incoming field, linearly polarized in the x -direction, in the form

$$\vec{E}(r, z) = \vec{E}_0 \cdot e^{-\frac{iy}{4 \sin^2 \theta_0 / 2}} \cdot \begin{pmatrix} -i \cdot (I_0 + I_2 \cdot \cos 2\varphi) \\ -i \cdot I_2 \cdot \sin 2\varphi \\ -2 \cdot I_1 \cdot \cos \varphi \end{pmatrix} \quad (27-4)$$

where φ is the angle of the azimuth and ϑ_0 is the angle of the full numerical aperture. The following three auxiliary function are used in the above expression

$$I_0(r, z) = \int_0^{\vartheta_0} \sqrt{\cos \vartheta} \cdot \sin \vartheta \cdot (1 + \cos \theta) \cdot J_0(kr \sin \vartheta) \cdot e^{ikz \cos \vartheta} d\vartheta, \quad (27-5)$$

$$I_1(r, z) = \int_0^{\vartheta_0} \sqrt{\cos \vartheta} \cdot \sin^2 \vartheta \cdot J_1(kr \sin \vartheta) \cdot e^{ikz \cos \vartheta} d\vartheta, \quad (27-6)$$

$$I_2(r, z) = \int_0^{\vartheta_0} \sqrt{\cos \vartheta} \cdot \sin \vartheta \cdot (1 - \cos \vartheta) \cdot J_2(kr \sin \vartheta) \cdot e^{ikz \cos \vartheta} d\vartheta. \quad (27-7)$$

If the normalized coordinates u and v with

$$v = kr \cdot \sin \vartheta_0, \quad (27-8)$$

$$u = 4k \cdot z \cdot \sin^2 \frac{\vartheta_0}{2} = 2k \cdot z \cdot (1 - \cos \vartheta_0) \quad (27-9)$$

are used, the integrals I_0 , I_1 and I_2 can be written in the scaled form

$$I_0(u, v) = \int_0^{\vartheta_0} \sqrt{\cos \vartheta} \cdot \sin \vartheta \cdot (1 + \cos \vartheta) \cdot J_0\left(\frac{v \cdot \sin \vartheta}{\sin \vartheta_0}\right) \cdot e^{\frac{i u \cdot \sin^2 \vartheta / 2}{\sin^2 \vartheta_0 / 2}} d\vartheta, \quad (27-10)$$

$$I_1(u, v) = \int_0^{\vartheta_0} \sqrt{\cos \vartheta} \cdot \sin^2 \vartheta \cdot J_1\left(\frac{v \cdot \sin \vartheta}{\sin \vartheta_0}\right) \cdot e^{\frac{i u \cdot \sin^2 \vartheta / 2}{\sin^2 \vartheta_0 / 2}} d\vartheta, \quad (27-11)$$

$$I_2(u, v) = \int_0^{\vartheta_0} \sqrt{\cos \vartheta} \cdot \sin \vartheta \cdot (1 - \cos \vartheta) \cdot J_2\left(\frac{v \cdot \sin \vartheta}{\sin \vartheta_0}\right) \cdot e^{\frac{i u \cdot \sin^2 \vartheta / 2}{\sin^2 \vartheta_0 / 2}} d\vartheta. \quad (27-12)$$

Figure 27-3 shows the field components of a point-spread function of a typical focussing system with high numerical aperture as calculated with the formulas above. The dependence on the azimuth angle φ generated the remarkable intensity pattern for the orthogonal polarization and the axial field component. The color coding in this illustration is renormalized for all components, in reality, the x -component has a dominant size.

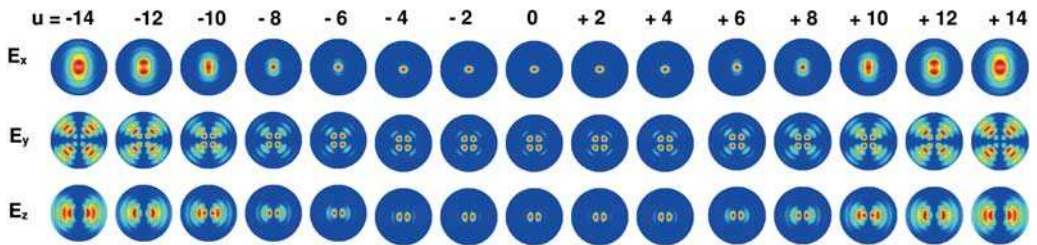


Figure 27-3: Field components of the point-spread function for a focussing setup with high numerical aperture as calculated with the diffraction integral according to Richards and Wolf.

The practicality of the diffraction integral in the above formulation is limited, since it cannot be calculated with a fast algorithm and the assumed rotational symmetry restricts the general application. But this framework allows a good estimation of the size of the vectorial effects.

The incoming field is linearly polarized in the x -direction. The vectorial effect causes a coupling of the field components, so the field in the focal region has a non-vanishing field strength in the y and the z direction. I_0 has a much larger value than the other two integrals. Therefore the ratio of the Integrals I_1 and I_2 to I_0 gives a relative measure for the influence of the vector effects. In Figure 27-4, these two

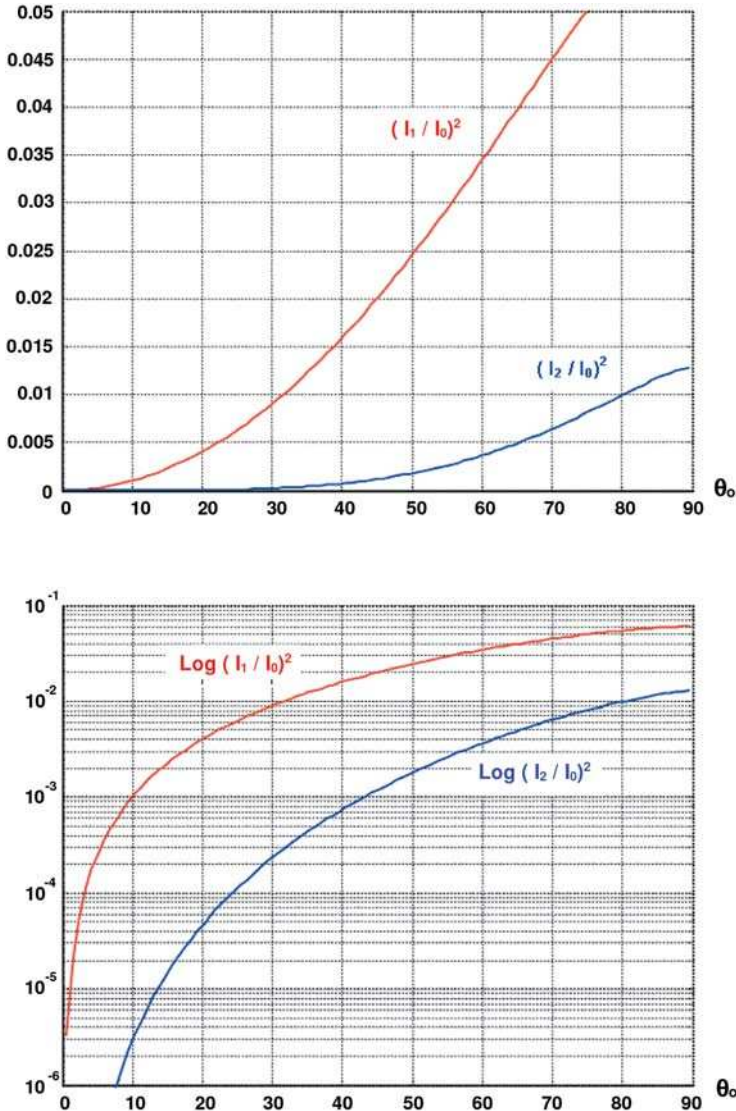


Figure 27-4: Relative size of the orthogonal axial (red) and the transverse (blue) component of the vectorial point-spread function in the image plane as a function of the numerical aperture in linear and logarithmic scaling.

ratios are shown in the image plane $u = 0$ as a function of the angle θ_0 of the numerical aperture. It can be seen that the axial component grows more rapidly than the transverse component. The axial part obtains a relative size of 1% for an angle of 30° . A corresponding size of the transverse component is reached for an angle of 80° . Since, in the practical situation, the numerical aperture is the more usual term to consider, a corresponding picture is shown in Figure 27-5. Table 1 contains the relative error of the scalar model for some selected values of the numerical aperture.

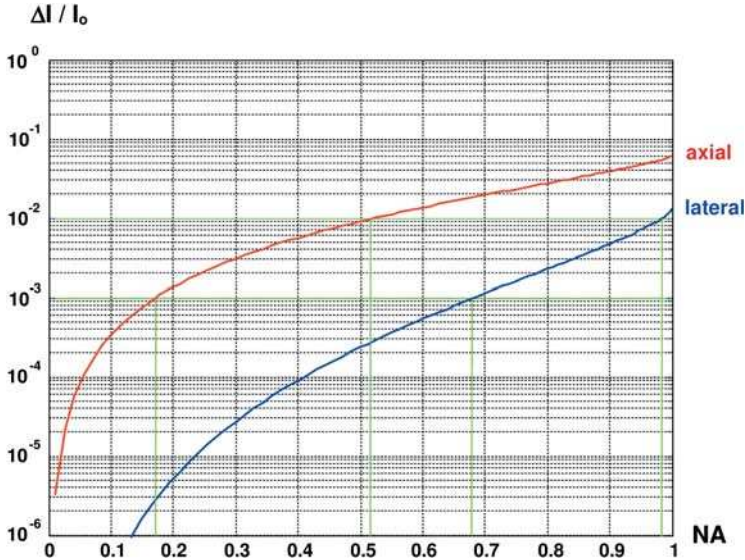


Figure 27-5: Relative size of the vectorial effects as a function of the numerical aperture separated into the transverse and the axial part.

Table 1: Relative error of the scalar point-spread function for some selected numerical apertures in axial and transverse field components.

NA	Axial error	Transverse error
0.2	0.0013	0
0.3	0.0031	0
0.4	0.0057	0.0001
0.5	0.0090	0.0002
0.6	0.0137	0.0005
0.7	0.0193	0.0011
0.8	0.0274	0.0023
0.9	0.0392	0.0048

Figure 27-6 shows the distribution of the two field strengths as a function of the defocussing parameter u and the transverse coordinate v . The vectorial effect dominates in the off-axis regime in the neighborhood of the image plane.

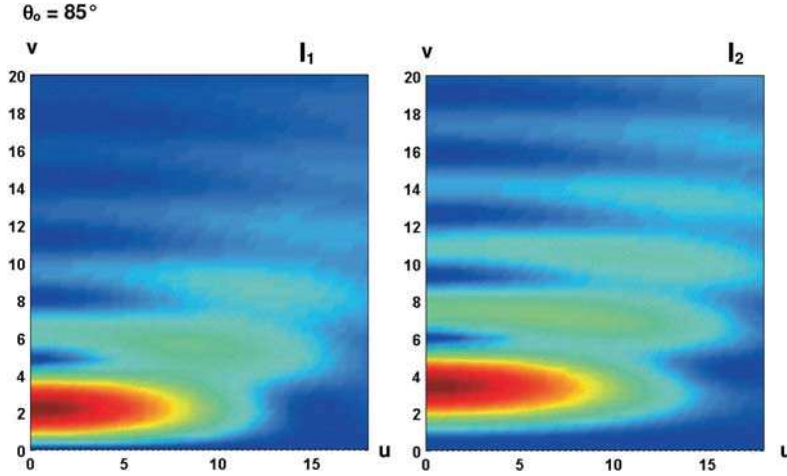


Figure 27-6: Relative size of the orthogonal transverse and the axial component of the vectorial point-spread function in the image plane as a function of the defocussing and the transverse coordinate v for an angle of 85° which corresponds to an NA of 0.996.

If the sine condition is not valid, the diffraction integrals in the Richards–Wolf formulation can be generalized with an apodization function $P(\vartheta)$ similar to the scalar case. The auxiliary functions are then expressed as

$$I_0(u, v) = \int_0^{\vartheta_0} P(\vartheta) \cdot \sqrt{\cos \vartheta} \cdot \sin \vartheta \cdot (1 + \cos \vartheta) \cdot J_0\left(\frac{v \cdot \sin \vartheta}{\sin \vartheta_0}\right) \cdot e^{\frac{iu \cdot \sin^2 \vartheta / 2}{2 \sin^2 \vartheta_0 / 2}} d\vartheta, \quad (27-13)$$

$$I_1(u, v) = \int_0^{\vartheta_0} P(\vartheta) \cdot \sqrt{\cos \vartheta} \cdot \sin^2 \vartheta \cdot J_1\left(\frac{v \cdot \sin \vartheta}{\sin \vartheta_0}\right) \cdot e^{\frac{iu \cdot \sin^2 \vartheta / 2}{2 \sin^2 \vartheta_0 / 2}} d\vartheta, \quad (27-14)$$

$$I_2(u, v) = \int_0^{\vartheta_0} P(\vartheta) \cdot \sqrt{\cos \vartheta} \cdot \sin \vartheta \cdot (1 - \cos \vartheta) \cdot J_2\left(\frac{v \cdot \sin \vartheta}{\sin \vartheta_0}\right) \cdot e^{\frac{iu \cdot \sin^2 \vartheta / 2}{2 \sin^2 \vartheta_0 / 2}} d\vartheta. \quad (27-15)$$

In the special case of a very narrow annular pupil, the vectorial form of the diffraction integral can be solved in an analytical form. According to [27-3] one gets, for the image plane position $u = 0$, the intensity

$$\begin{aligned} I(v, \varphi) = & J_0^2(v) + 2 \tan^2 \frac{\vartheta_0}{2} \cdot J_1^2(v) + 2 \tan^2 \frac{\vartheta_0}{2} \cdot J_2^2(v) \\ & + 2 \cos(2\varphi) \cdot \tan^2 \frac{\vartheta_0}{2} \cdot [J_1^2(v) + J_0(v) \cdot J_2(v)]. \end{aligned} \quad (27-16)$$

For a circular input polarization, the corresponding intensity can be written as

$$I_{\text{circ}}(\nu) = J_0^2(\nu) + 2 \tan^2 \frac{\vartheta_o}{2} \cdot J_1^2(\nu) + \tan \frac{\vartheta_o}{2} \cdot J_2^2(\nu). \quad (27-17)$$

A radial polarization delivers the expression

$$I_{\text{rad}}(\nu) = J_0^2(\nu) + \cos \vartheta_o \cdot J_1^2(\nu). \quad (27-18)$$

Figure 27-7 shows the transverse intensity profiles for different starting polarizations, together with the scalar annular pupil and the Airy distribution of a circular aperture in the image plane for an aperture angle of 72° , shown for comparison.

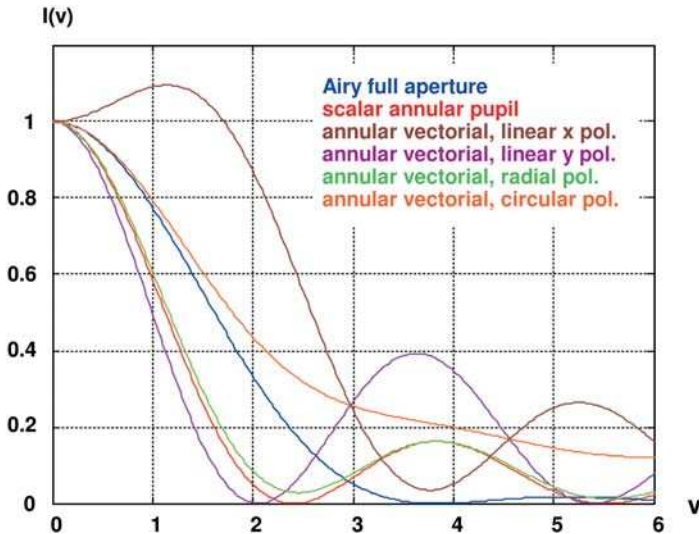


Figure 27-7: Point-spread function profiles for an annular pupil with different polarizations, the scalar case and the Airy distribution.

27.2.3

Plane Wave Expansion

In the diffraction computation according to the Debye approximation, each point of the exit pupil is taken as the source of a plane wave with amplitude, phase and polarization prescribed by this point. This provides a versatile method for focus computation that is easily generalized to non-circular pupils. The propagation vector corresponding to the pupil coordinate (p_x, p_y) with $k_0 = 2\pi/\lambda$ in a medium of refractive index n is

$$\vec{k} = k_0 \begin{pmatrix} p_x \\ p_y \\ p_z \end{pmatrix} \quad (27-19)$$

with

$$p_z = \sqrt{n^2 - p_x^2 - p_y^2} . \quad (27-20)$$

While the electric field \vec{E}_{pupil} in the pupil coordinates is a two-component phasor, the electric field \vec{E}_{3D} in global coordinates in image space usually has three components and can be derived by application of a 3×2 transformation matrix $\boldsymbol{\psi}$

$$\vec{E}_{3D} = \boldsymbol{\psi}(\vec{p}) \vec{E}_{\text{pupil}} . \quad (27-21)$$

The transformation matrix follows from the change in the direction of the propagation vector. The succession of three rotational matrices that provides this change of direction

$$\vec{E}_{3D} = \begin{pmatrix} \cos \varphi & -\sin \varphi & 0 \\ \sin \varphi & \cos \varphi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos \vartheta & 0 & \sin \vartheta \\ 0 & 1 & 0 \\ -\sin \vartheta & 0 & \cos \vartheta \end{pmatrix} \begin{pmatrix} \cos \varphi & \sin \varphi & 0 \\ -\sin \varphi & \cos \varphi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} E_x \\ E_y \\ 0 \end{pmatrix} \quad (27-22)$$

eventually yields the transformation matrix

$$\mathbf{T} = \begin{pmatrix} \cos^2 \varphi \cos \vartheta + \sin^2 \varphi & -\sin \varphi \cos \varphi (1 - \cos \vartheta) \\ -\sin \varphi \cos \varphi (1 - \cos \vartheta) & \sin^2 \varphi \cos \vartheta + \cos^2 \varphi \\ -\cos \varphi \sin \vartheta & -\sin \varphi \sin \vartheta \end{pmatrix} \quad (27-23)$$

and we finally have the diffraction integral

$$\vec{E}(\vec{r}) = \int \int_{EP} \boldsymbol{\psi}(p_x, p_y) \vec{E}(p_x, p_y) \exp\{i\vec{k} \cdot \vec{r}\} (\cos \vartheta)^{-1/2} dp_x dp_y \quad (27-24)$$

which can be implemented as an inverse Fourier transform

$$\begin{aligned} \vec{E}(x, y, z) &= \int \int_{\text{pupil}} \vec{E}_{\text{glob}}(p_x, p_y) e^{ik_0(p_x x + p_y y + p_z z)} dp_x dp_y \\ &= \text{IFT} \left\{ \vec{E}_{\text{glob}}(p_x, p_y) e^{ik_0 p_z z} \right\} . \end{aligned} \quad (27-25)$$

The intensity distribution in the focus is the sum of the intensities of the three field components

$$\begin{aligned} I(\vec{r}) &= \left\| \vec{E}(\vec{r}) \right\|^2 = |E_x(\vec{r})|^2 + |E_y(\vec{r})|^2 + |E_z(\vec{r})|^2 \\ &= I_x + I_y + I_z . \end{aligned} \quad (27-26)$$

27.2.4

Focus Fields for Various Input Polarizations

In this section, examples of high NA foci are shown, which are computed with the plane wave expansion. The polarization states in the entrance pupil are linear, circular, radial, and tangential.

A) Focus of a Linearly Polarized Pupil

Consider an entrance pupil in which the field is linearly polarized in the x direction (Figure 27-8). The amplitude is constant. This case has already been considered in section 27.2.2.

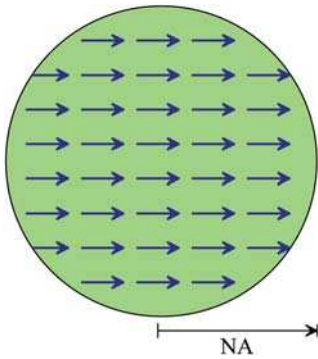


Figure 27-8: Linearly polarized pupil.

While the focus shows circular symmetry for small NA , it becomes asymmetrical with increasing NA as is shown in Figure 27-9. The focus width is larger along the direction of polarization than it is perpendicular to it. This is a direct consequence of the vector effect: Perpendicular to the polarization direction, the interfering electric field components are parallel to each other. Along the polarization direction they enclose the same angle as the propagation vectors. Because of this, the contrast of interference is reduced which yields a broader focus. Finally, this results in the dependence of the image of a line structure, on the orientation of it. The image contrast is asymmetrical with higher contrast perpendicular to the polarization direction.

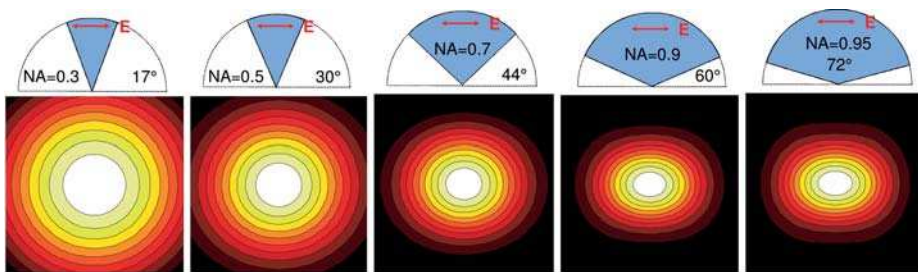


Figure 27-9: Focus of linearly polarized field for increasing numerical aperture.

The vector effect is significant for high- NA projection optics where a demagnified image of a structure is formed. In optical microscopy the image is magnified, i.e., the image is formed with low NA . (For a microscopy lens of $NA = 0.9$ and a magnification of $100\times$ the NA in image space is 0.009). Therefore, the vector effect is entirely negligible there.

Figure 27-10 shows the summed intensity and the x , y , and z components separately. The field components show a characteristic pattern. While the x -polarized focus is quite symmetric, the summed focus is asymmetric, mainly due to the super-

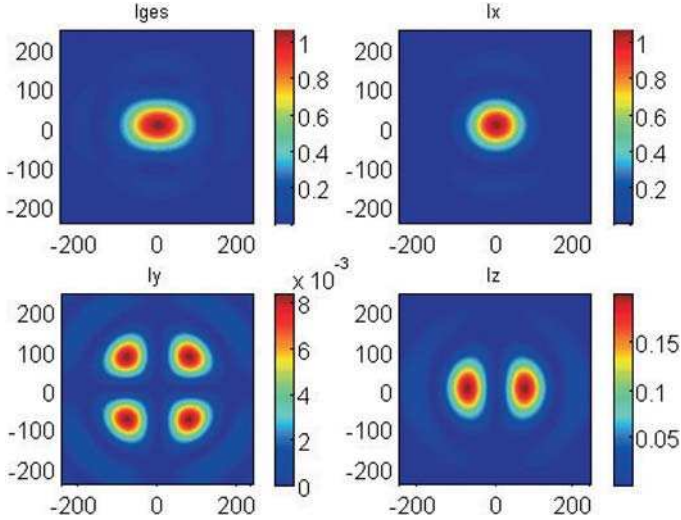


Figure 27-10: x , y , z component and summed intensity (I_{ges}) for an x -linearly polarized incident field of $NA = 0.9$ in air.

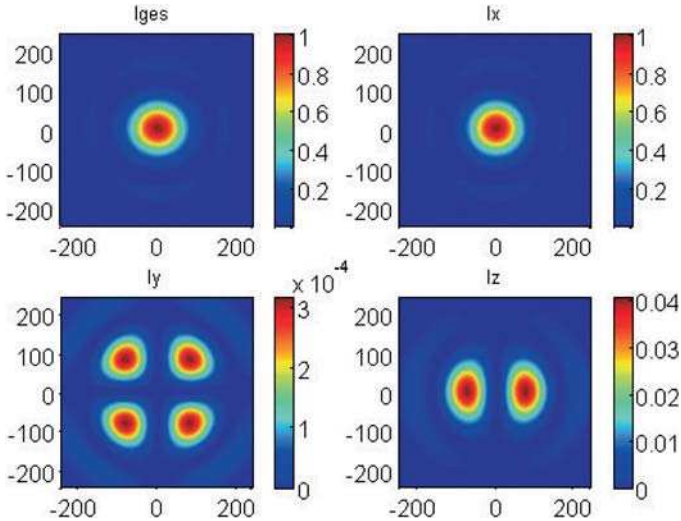


Figure 27-11: x , y , z component and summed intensity (I_{ges}) for an x -linearly polarized incident field in a photoemulsion of refractive index $n = 1.8$.

position of the z -polarized focus. Due to symmetry it is zero on the axis and shows two separated maxima along the x -direction. The y -polarized focus is considerably weaker with two zero lines along the x and y coordinate axes.

If the image is formed in a dielectric medium with a high index of refraction, the interfering plane waves become more parallel because of refraction into the photo-emulsion. The z -component is strongly diminished and the summed intensity distribution shows almost circular symmetry as for a low-NA focus (Figure 27-11).

B) Focus of a Circularly Polarized Pupil

Consider an entrance pupil in which the field is circularly polarized (27-12). The amplitude is constant. The focus is shown in Figure 27-13 for a wavelength of 193 nm and NA of 0.95. It has circular symmetry for any NA. The focus of the z component has circular symmetry in the shape of a doughnut and vanishes on the optical axis. The x and y component are both asymmetrical but with orthogonal axes so that their sum is again symmetrical.

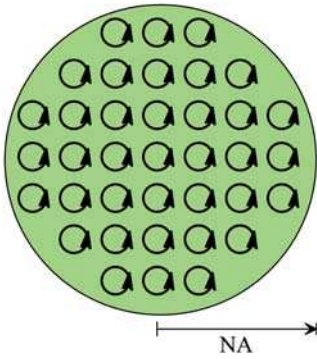


Figure 27-12: Circularly polarized pupil.

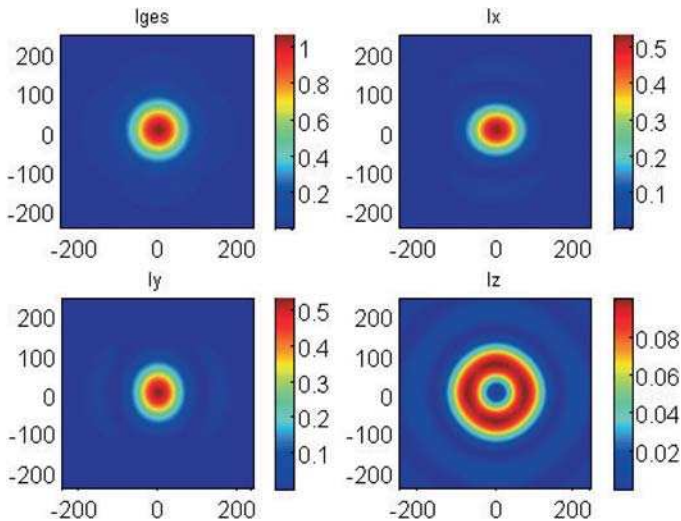


Figure 27-13: x , y , z component and summed intensity (I_{ges}) for a circularly polarized incident field.

C) Focus of a Radially Polarized Pupil

For a radially polarized pupil, the amplitude cannot be constant because of the singularity in the center. Here we choose a linearly increasing amplitude and obtain the pupil field shown in Figure 27-14.

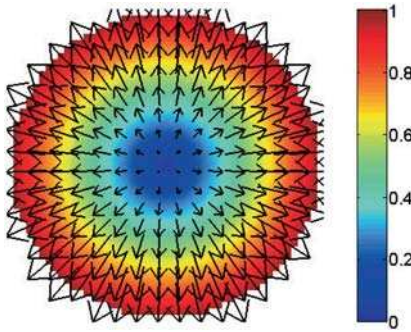


Figure 27-14: Radially polarized pupil.

The radially polarized focus has a sharp intensity peak on the optical axis caused by the now constructive interference of the z -component. The z -component alone provides an even better focus than the overall intensity [27-4]. Because of this, a ring-shaped aperture is advantageous here: it reduces the x and y -polarized intensity. These have zero intensity perpendicular to the polarization direction, due to symmetry.

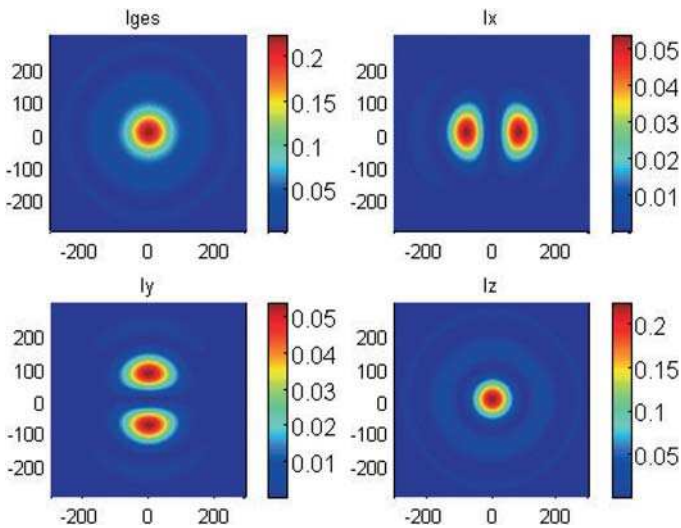


Figure 27-15: x , y , z component and summed intensity (I_{ges}) for a radially polarized incident field.

D) Focus of a Tangentially Polarized Pupil

As for the radially polarized pupil, the amplitude of the tangentially polarized pupil cannot be constant because of the singularity in the center. Again we choose a linearly increasing amplitude and obtain the pupil field shown in Figure 27-16.

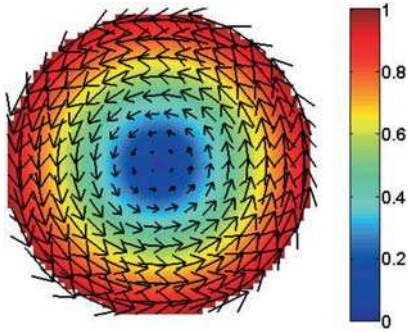


Figure 27-16: Tangentially polarized pupil.

A tangentially polarized pupil yields a doughnut-focus with a vanishing intensity on axis (27-17). The reason is that opposing parts of the pupil are π -phase shifted with respect to each other. For this configuration the z -component is zero, which makes the tangential polarization an excellent alternative for high-NA image generation.

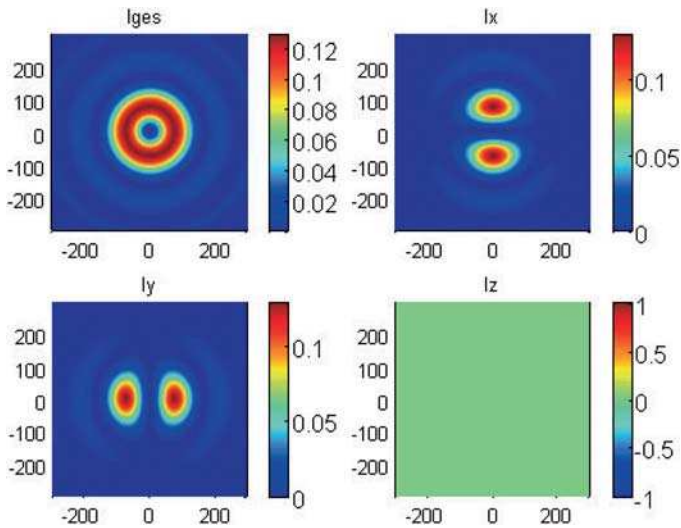


Figure 27-17: x , y , z component and summed intensity (I_{ges}) for a tangentially polarized incident field.

27.3

Vector Kirchhoff Diffraction Integral

The vector Kirchhoff diffraction integral, also called the Stratton-Chu integral ([27-5], [27-6]) after the inventors, results from an application of the vector form of Green's relation. For a closed volume V with a surface S the following relation holds

$$\iint_S \left(\vec{P} \times (\nabla \times \vec{Q}) - \vec{Q} (\nabla \times \vec{P}) \right) \cdot \vec{s} \, d\sigma = \iiint_V \left(\vec{Q} \cdot \nabla \times (\nabla \times \vec{P}) - \vec{P} \cdot \nabla \times (\nabla \times \vec{Q}) \right) dV \quad (27-27)$$

where \vec{P} and \vec{Q} are defined in the volume V and have a continuous second derivative there.

Assuming that V and S are free of charges and currents, we obtain the following integral for the electric field [27-5]

$$\vec{E}(\vec{r}) = \frac{1}{4\pi} \iint_S \left[i\omega\mu_0\mu_r (\vec{s} \times \vec{H}) G + (\vec{s} \times \vec{E}) \times \nabla G + (\vec{s} \cdot \vec{E}) \nabla G \right] d\sigma \quad (27-28)$$

with

$$G = \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|} . \quad (27-29)$$

Applying this formula to a planar aperture A in a black screen in the xy plane, yields the radiation field of the aperture as

$$\begin{aligned} \vec{E}(\vec{r}) = & \frac{ik_0}{4\pi} \iint_A \left\{ a(R) [\vec{e}_z \times \vec{E}(\vec{r}_t)] \times \vec{e}_r - Z_0 b(R) [\vec{e}_z \times \vec{H}(\vec{r}_t)] \right. \\ & \left. + Z_0 c(R) [(\vec{e}_z \times \vec{H}(\vec{r}_t)) \cdot \vec{e}_r] \vec{e}_r \right\} dx_t dy_t \end{aligned} \quad (27-30)$$

with $\vec{e}_z = (0,0,1)$, $\vec{e}_r = (\vec{r}-\vec{r}_t)/|\vec{r}-\vec{r}_t|$ and $(R = |\vec{r}-\vec{r}_t|)$

$$\begin{aligned} a(R) &= 1 - \frac{1}{ik_0 R} , \\ b(R) &= 1 - \frac{1}{ik_0 R} + \frac{1}{(ik_0 R)^3} , \\ c(R) &= 1 - \frac{3}{ik_0 R} + \frac{3}{(ik_0 R)^3} . \end{aligned} \quad (27-31)$$

27.4

Analytical Solutions

Analytical solutions are invaluable for judging the validity of numerical methods, but they exist for a very limited number of cases only. These include the following:

- a) Perfect conductors of simple geometry like Sommerfeld's solution for diffraction at a perfectly conducting half-plane [27-7] or Bethe's solution for diffraction at sub-wavelength apertures [27-8]. Aperture arrays in perfectly conducting screens see [27-9].
- b) Simple geometries where the physical surface of the diffraction structure coincides with a coordinate surface. This allows a separation approach in the solution of the Maxwell equations.

The most common examples of the latter are Fresnel equations [27-10] in Cartesian coordinates, scattering at a circular cylinder [27-11][27-12] in cylinder coordinates and Mie scattering [27-13] in spherical coordinates. Cylinder and Mie scattering have been extended to elliptical cylinders [27-14] and ellipsoids. All methods have been extended to multilayer systems.

The mathematical procedure for the solution of the diffraction problem is very similar for all geometries:

1. Express the electromagnetic field in the appropriate coordinate system by use of Maxwell's equations in that coordinate system.
2. Expand the reflected (scattered) and transmitted (internal) field as a superposition of orthogonal modes with the reflection and transmission coefficients as variables.

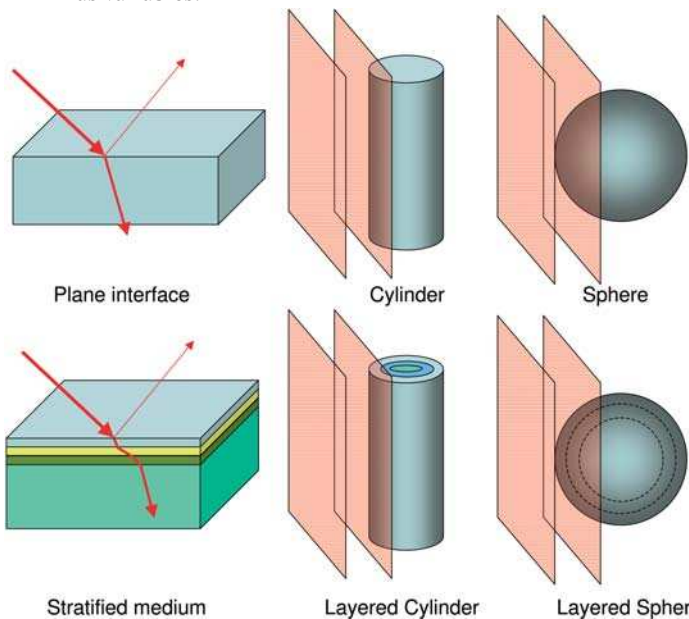


Figure 27-18: Schematic view of the most common examples for analytical solutions of electromagnetic diffraction.

3. Apply the continuity of the tangential electric and magnetic field component at the interface for each mode.
4. Solve for the reflection and transmission coefficient.

27.4.1

Plane Interface: Fresnel's Equations

The Fresnel equations are discussed in detail in section 3.2 of Vol. 1. Here we rederive them from the continuity of the tangential fields. The appropriate coordinate system for plane interfaces is Cartesian (Figure 27-19). The x and the y components of the fields are tangential to the interface. The z component is normal to it. A plane wave is incident from medium 1 onto the interface to medium 2 under an angle ϑ_i .

The electromagnetic wave is called s-polarized if the electric field \vec{E} is tangential to the interface, i.e., it has a y component only. Then the y component of the magnetic field \vec{H} is vanishing. The orthogonal polarization is called p-polarized. Here, the magnetic field is tangential to the interface (has a y component only) and the y component of the electric field is vanishing.

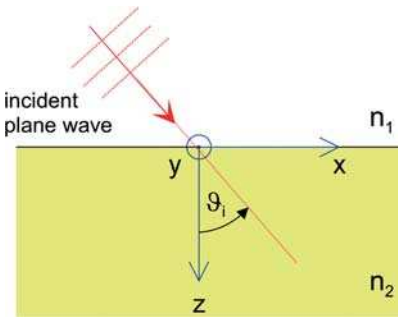


Figure 27-19: Geometry for Fresnel's equations.

The relation between \vec{E} and \vec{H} is provided by Maxwell's equations in Cartesian coordinates

$$\text{s-polarization : } \vec{E} = \begin{pmatrix} 0 \\ E_y \\ 0 \end{pmatrix} \rightarrow \vec{H} = \frac{-i}{\omega\mu} \nabla \times \vec{E} = \frac{i}{\omega\mu} \begin{pmatrix} \partial E_y / \partial z \\ 0 \\ -\partial E_y / \partial x \end{pmatrix} \quad (27-32)$$

$$\text{p-polarization : } \vec{H} = \begin{pmatrix} 0 \\ H_y \\ 0 \end{pmatrix} \rightarrow \vec{E} = \frac{i}{\omega\varepsilon} \nabla \times \vec{H} = \frac{i}{\omega\varepsilon} \begin{pmatrix} -\partial H_y / \partial z \\ 0 \\ \partial H_y / \partial x \end{pmatrix}$$

yielding the fields E_{tang} and H_{tang} tangential to the interface

	E_{tang}	H_{tang}
s-pol.	E_y	$\frac{i}{\omega\mu} \frac{\partial E_y}{\partial z}$
p-pol.	$-\frac{i}{\omega\varepsilon} \frac{\partial H_y}{\partial z}$	H_y

(27-33)

The symmetry of these equations allows the computation to be carried out for s-polarization and the results to be derived for p-polarization from symmetry considerations. Therefore, we consider an s-polarized plane wave that is propagated with an angle ϑ_i to the z axis. The tangential field components are, for s-polarization,

$$\begin{aligned}
 E_y^i &= e^{in_1 k_0(x \sin \vartheta_1 + z \cos \vartheta_1)}, \\
 H_x^i &= -\frac{n_1 k_0}{\omega\mu} \cos \vartheta_1 E_y^i.
 \end{aligned}
 \tag{27-34}$$

The reflected field becomes, according to the law of reflection,

$$\begin{aligned}
 E_y^r &= r_s e^{in_1 k_0(x \sin \vartheta_1 - z \cos \vartheta_1)}, \\
 H_x^r &= \frac{n_1 k_0}{\omega\mu} \cos \vartheta_1 E_y^r
 \end{aligned}
 \tag{27-35}$$

and the transmitted field

$$\begin{aligned}
 E_y^t &= t_s e^{in_2 k_0(x \sin \vartheta_2 + z \cos \vartheta_2)}, \\
 H_x^t &= -\frac{n_2 k_0}{\omega\mu} \cos \vartheta_2 E_y^t.
 \end{aligned}
 \tag{27-36}$$

The continuity at the interface requires

$$\begin{aligned}
 E_y^i + E_y^r &= E_y^t \text{ @ } z = 0 \rightarrow 1 + r_s = t_s, \\
 H_x^i + H_x^r &= H_x^t \text{ @ } z = 0 \rightarrow n_1 \cos \vartheta_1 - r_s n_1 \cos \vartheta_1 = t_s n_2 \cos \vartheta_2.
 \end{aligned}
 \tag{27-37}$$

where the law of reflection was applied and a non-magnetic medium ($\mu_r = 1$) was assumed. The solution is

$$\begin{aligned}
 r_s &= \frac{n_1 \cos \vartheta_1 - n_2 \cos \vartheta_2}{n_1 \cos \vartheta_1 + n_2 \cos \vartheta_2}, \\
 t_s &= \frac{2n_1 \cos \vartheta_1}{n_1 \cos \vartheta_1 + n_2 \cos \vartheta_2}.
 \end{aligned}
 \tag{27-38}$$

The computation for p-polarization is analogous, yielding the reflection and transmission coefficients

$$r_p = \frac{n_2 \cos \vartheta_1 - n_1 \cos \vartheta_2}{n_2 \cos \vartheta_1 + n_1 \cos \vartheta_2}, \quad (27-39)$$

$$t_p = \frac{2n_1 \cos \vartheta_1}{n_2 \cos \vartheta_1 + n_1 \cos \vartheta_2}.$$

Reflection and transmission curves for various cases are presented in section 3.2 of Vol. 1.

27.4.2

Diffraction at a Circular Cylinder

The solution for the diffraction of an incident plane wave by a cylinder of circular cross-section has been known since 1918 [27-11]. More general cases are the obliquely incident plane wave (onto a circular cylinder) [27-12] and the cylinder of elliptical cross-section [27-14].

The basic geometry is sketched in Figure 27-20. A circular cylinder of refractive index n_2 and radius R_0 is immersed in a homogeneous medium of refractive index n_1 . Although a general magnetic cylinder presents no problems, we consider here only the non-magnetic case, i.e., $\mu_r = 1$. The cylinder is situated in the origin of a Cartesian coordinate system with its axis parallel to the y axis.

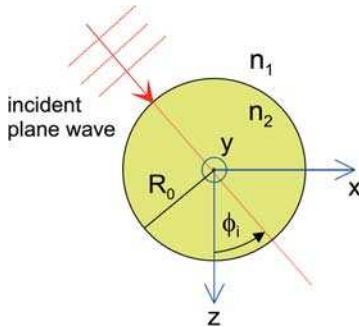


Figure 27-20: Geometry for the diffraction of a plane wave by a circular cylinder.

The electromagnetic wave is called TE-polarized if the electric field E has a y component only. Then the y component of the magnetic field H is vanishing. The orthogonal polarization is called TM-polarized. Here, the magnetic field has a y component only and the y component of the electric field is vanishing. For an incident wave in the xz plane, the electric field component is tangential to the cylinder surface in TE-polarization. In TM polarization the magnetic field is tangential.

\vec{E} and \vec{H} are represented in polar coordinates

$$\vec{E} = E_r \vec{e}_r + E_\phi \vec{e}_\phi + E_y \vec{e}_y, \quad (27-40)$$

$$\vec{H} = H_r \vec{e}_r + H_\phi \vec{e}_\phi + H_y \vec{e}_y.$$

The relation between \vec{E} and \vec{H} is provided by Maxwell's equations in polar coordinates

$$\text{TE-polarization : } \vec{E} = E_y \vec{e}_y \rightarrow \vec{H} = -\frac{i}{\omega\mu} \left(\frac{1}{r} \frac{\partial E_y}{\partial \phi} \vec{e}_r - \frac{\partial E_y}{\partial r} \vec{e}_\phi \right) \quad (27-41)$$

$$\text{TM-polarization : } \vec{H} = H_y \vec{e}_y \rightarrow \vec{E} = \frac{i}{\omega\varepsilon} \left(\frac{1}{r} \frac{\partial H_y}{\partial \phi} \vec{e}_r - \frac{\partial H_y}{\partial r} \vec{e}_\phi \right)$$

yielding the fields E_{tang} and H_{tang} tangential to the interface

	E_{tang}	H_{tang}
TE-pol.	E_y	$\frac{i}{\omega\mu} \frac{\partial E_y}{\partial r}$
TM-pol.	$-\frac{i}{\omega\varepsilon} \frac{\partial H_y}{\partial r}$	H_y

(27-42)

The symmetry of these equations again allows one to do the computation for s-polarization and to derive the results for p-polarization from symmetry considerations. Therefore, we consider a TE-polarized plane wave that is propagated at an angle ϕ_i to the z axis. The fields are expanded into a Bessel series. The tangential field components are, for TE-polarization,

$$E_y^i = e^{in_1 k_0 (x \sin \phi_i + z \cos \phi_i)} = \sum_{m=-\infty}^{\infty} (-i)^m J_m(k_1 r) e^{im(\phi - \phi_i)}, \quad (27-43)$$

$$H_\phi^i = \frac{ik_1}{\omega\mu} \sum_{m=-\infty}^{\infty} (-i)^m J_m'(k_1 r) e^{im(\phi - \phi_i)}.$$

The reflected (scattered) field is

$$E_y^r = \sum_{m=-\infty}^{\infty} a_m H_m(k_1 r) e^{im\phi}, \quad (27-44)$$

$$H_\phi^r = \frac{ik_1}{\omega\mu} \sum_{m=-\infty}^{\infty} a_m H_m'(k_1 r) e^{im\phi},$$

and the transmitted field inside the cylinder

$$E_y^t = \sum_{m=-\infty}^{\infty} b_m J_m(k_2 r) e^{im\phi}, \quad (27-45)$$

$$H_\phi^t = \frac{ik_2}{\omega\mu} \sum_{m=-\infty}^{\infty} b_m J_m'(k_2 r) e^{im\phi},$$

where $k_1 = n_1 k_0$ is the propagation constant in the surrounding medium and $k_2 = n_2 k_0$ inside the cylinder. J_m and H_m are the Bessel function and the Hankel function of the first kind and m th order. The dash denotes the derivative. The continuity at the interface requires

$$\begin{aligned}
E_y^i + E_y^r &= E_y^t @ r = R_0 \rightarrow (-i)^m J_m(k_1 R_0) + a_m H_m(k_1 R_0) = b_m J_m(k_2 R_0) \\
H_x^i + H_x^r &= H_x^t @ r = R_0 \rightarrow (-i)^m J_m'(k_1 R_0) + a_m H_m'(k_1 R_0) = n b_m J_m'(k_2 R_0)
\end{aligned}
\tag{27-46}$$

with $n = \frac{n_2}{n_1}$. Computing a_m and b_m and resolving the complex exponential we obtain the electric field E_y inside the cylinder.

$$E_y = \sum_{m=0}^{\infty} B_m \varepsilon_m (-i)^m \cos [m(\phi - \phi_i)] J_m(k_1 r) . \tag{27-47}$$

And outside

$$E_y = \sum_{m=0}^{\infty} (J_m(k_2 r) + A_m H_m(k_2 r)) \varepsilon_m (-i)^m \cos [m(\phi - \phi_i)] \tag{27-48}$$

with the factor $\varepsilon_m = 1$ for $m=0$ and 2 otherwise and with the coefficients

$$\begin{aligned}
A_m &= \frac{n J_m(k_2 R_0) J_m'(k_1 R_0) - J_m'(k_2 R_0) J_m(k_1 R_0)}{J_m(k_1 R_0) H_m'(k_2 R_0) - n J_m'(k_1 R_0) H_m(k_2 R_0)} , \\
B_m &= \frac{J_m(k_2 R_0) H_m'(k_2 R_0) - J_m'(k_2 R_0) H_m(k_2 R_0)}{J_m(k_1 R_0) H_m'(k_2 R_0) - n J_m'(k_1 R_0) H_m(k_2 R_0)} .
\end{aligned}
\tag{27-49}$$

The corresponding solution for a TE-polarized incident wave differs from the TM-polarized one in only two respects. E_y is replaced by H_y and n by its inverse, i.e., for TE-polarization we have the magnetic field E_y inside the cylinder.

$$H_y = \sum_{m=0}^{\infty} B_m' \varepsilon_m (-i)^m \cos [m(\phi - \phi_0)] J_m(k_1 r) . \tag{27-50}$$

And outside

$$H_y = Z_0^{-1} \sum_{m=0}^{\infty} (J_m(k_2 r) + A_m' H_m(k_2 r)) \varepsilon_m (-i)^m \cos [m(\phi - \phi_0)] \tag{27-51}$$

with the coefficients

$$\begin{aligned}
A_m' &= \frac{n^{-1} J_m(k_2 R_0) J_m'(k_1 R_0) - J_m'(k_2 R_0) J_m(k_1 R_0)}{J_m(k_1 R_0) H_m'(k_2 R_0) - n^{-1} J_m'(k_1 R_0) H_m(k_2 R_0)} , \\
B_m &= \frac{J_m(k_2 R_0) H_m'(k_2 R_0) - J_m'(k_2 R_0) H_m(k_2 R_0)}{J_m(k_1 R_0) H_m'(k_2 R_0) - n^{-1} J_m'(k_1 R_0) H_m(k_2 R_0)} .
\end{aligned}
\tag{27-52}$$

Figure 27-21 shows as an example the diffracted field for a cylinder with a refractive index of 1.5 and an incident plane wave with a wavelength of 1 μm . The real part of the electric field is shown on the right-hand side. It shows the “waves” at a certain

instant, providing by these means a simultaneous representation of the amplitude and phase of the field. On the right-hand side the amplitude distribution is shown. For the $4\ \mu\text{m}$ cylinder, for example, you see in the real-part image how the wave is focussed inside the cylinder to its rear and you also see how a spherical wave is emanating from the “focus”. The amplitude distribution on the right of it shows the resulting amplitude distribution with a maximum at the focus position. A standing wave has formed in front of the cylinder while its rear shows the interference pattern of a spherical wave and a plane wave propagation, mainly in the same direction. Inside the cylinder a standing wave pattern has formed. With increasing cylinder diameter the transition to a more ray-optical propagation can be observed (figure 27.22).

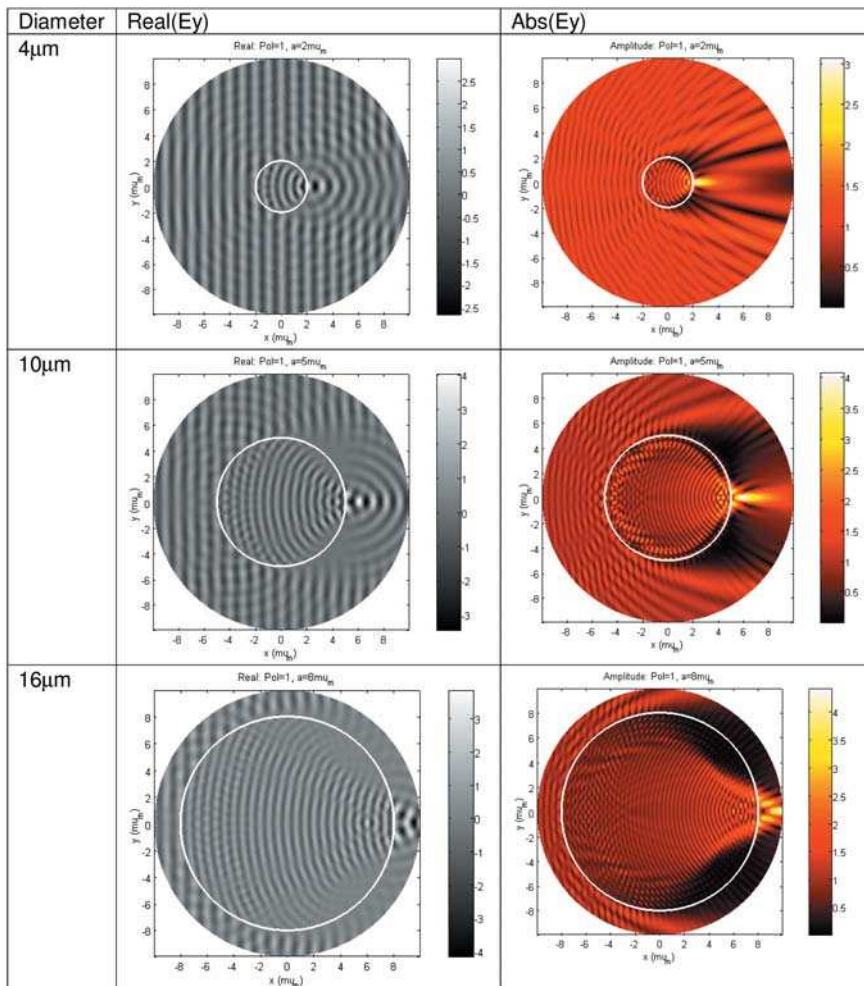


Figure 27-21: Diffraction at a dielectric cylinder in TE-polarization. $\lambda = 1\ \mu\text{m}$, $n = 1.5$, increasing diameter. Left: real part of the field, right: amplitude.

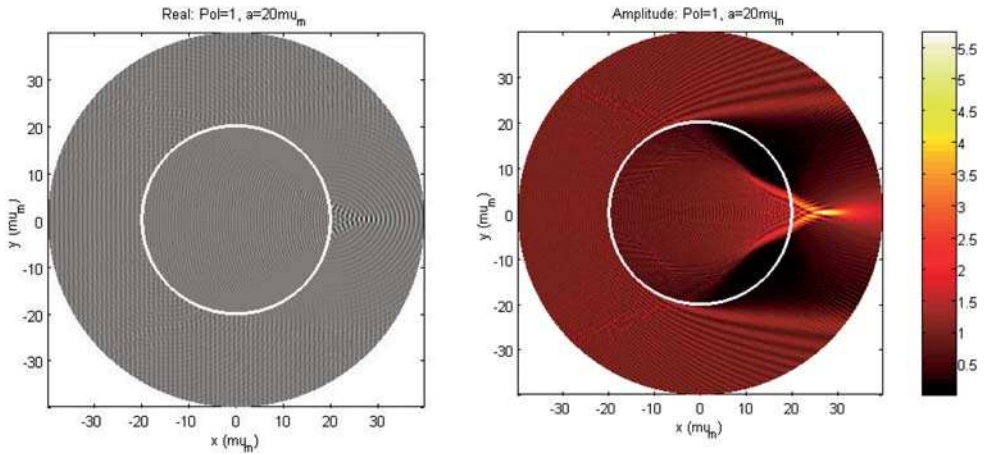


Figure 27-22: Diffraction at a dielectric cylinder in TE-polarization. $\lambda = 1 \mu\text{m}$, $n = 1.5$, $40 \mu\text{m}$ diameter. Left: real part of the field, right: amplitude.

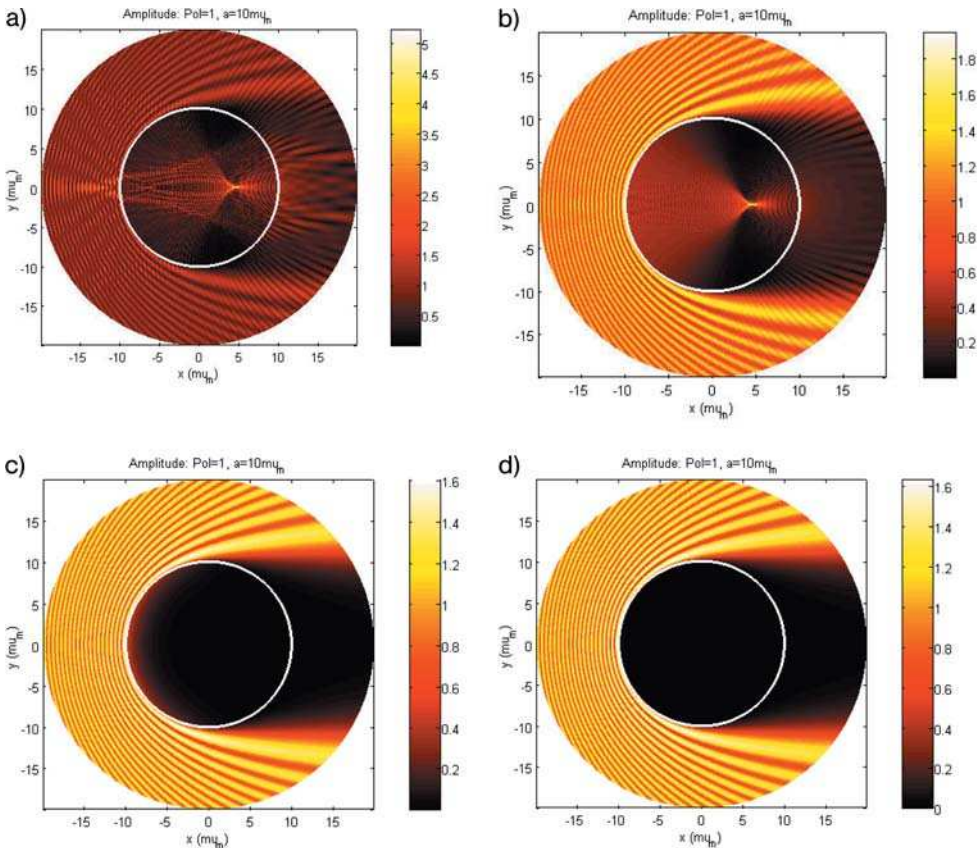


Figure 27-23: Diffraction amplitude at a dielectric cylinder in TE-polarization. $\lambda = 1 \mu\text{m}$, $20 \mu\text{m}$ diameter. The imaginary part of n is increasing: a) $n_2 = 3$, b) $n_2 = 3 + 0.01i$, c) $n_2 = 3 + 0.1i$, d) $n_2 = 3 + i$.

The refractive index must not be real. However, it can also be complex allowing, by these means, the diffraction at metallic cylinders to be computed. Figure 27-23 shows the diffraction at cylinders of equal real part of the refractive index ($n_r=3$) but increasing imaginary part ($n_i=0, 0.01, 0.1, 1$). Note how the field diminishes within the cylinder and a clear shadow develops behind it (the changing color map outside the cylinder is due to the fact that the amplitude range for the color map changes because the focus within the cylinder becomes weaker).

The diffraction at “inverse cylinders”, i.e., cylinders that have a smaller refractive index than their surroundings, can be computed as well (Figure 27-24). The inverse cylinder acts as a diverging lens producing no focus.

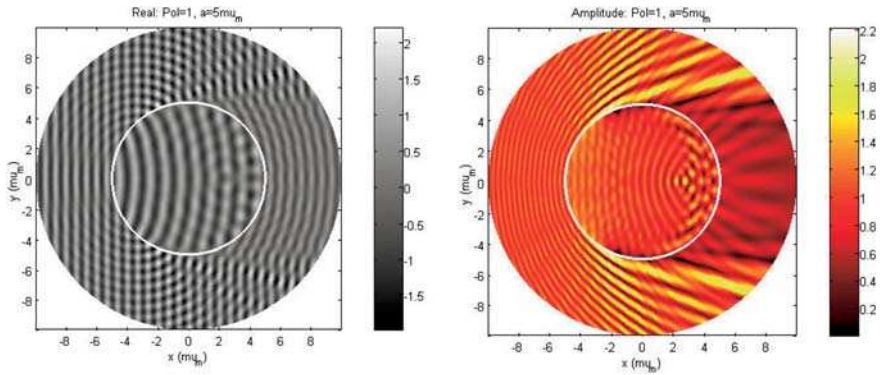


Figure 27-24: Diffraction at a dielectric air-cylinder ($n=1$) in glass ($n=1.5$) and in TE-polarization $\lambda = 1 \mu\text{m}$, $10 \mu\text{m}$ diameter.

27.4.3

Mie Scattering

Solving the Maxwell equations in spherical coordinates, G. Mie [27-13] rigorously computed the scattered field of spheres for excitation with plane waves. He used it to describe the optical properties of colloidal metallic solutions, in particular, polarization and angular distribution of the scattered field. Extensive and excellent presentations of the theory can be found in [27-15] and [27-16]. The method for obtaining the internal and external electromagnetic field is again a separation approach, as for the plane and the cylinder. \vec{E} and \vec{H} are represented in polar coordinates

$$\begin{aligned} \vec{E} &= E_r \vec{e}_r + E_\phi \vec{e}_\phi + E_\theta \vec{e}_\theta, \\ \vec{H} &= H_r \vec{e}_r + H_\phi \vec{e}_\phi + H_\theta \vec{e}_\theta. \end{aligned} \tag{27-53}$$

Here we present the results for the scattered far-field only. The scattering geometry is described in spherical coordinates. It is sketched in Figure 27-25.

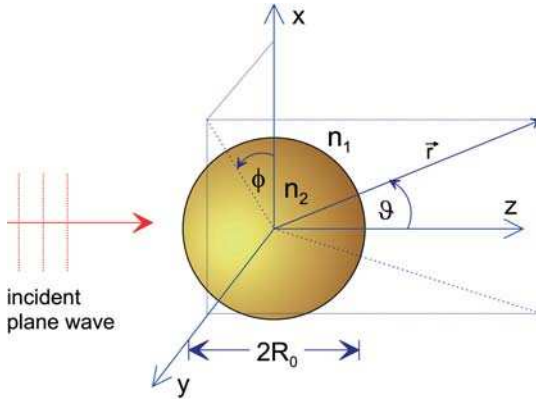


Figure 27-25: Geometry for diffraction of a plane wave by a sphere.

The result depends on two parameters: The ratio n of the refractive index n_2 of the sphere to the refractive index n_1 of the surroundings and the relation of the circumference of the sphere to the wavelength

$$q = \frac{2\pi R_0}{\lambda} . \tag{27-54}$$

The scattered field is given by

$$\begin{aligned} E_\phi &= i \frac{e^{ikr}}{kr} S_1(\vartheta) \sin \phi , \\ E_\vartheta &= i \frac{e^{ikr}}{kr} S_2(\vartheta) \cos \phi , \end{aligned} \tag{27-55}$$

with the scattering amplitudes

$$\begin{aligned} S_1(\vartheta) &= \sum_{m=1}^{\infty} \frac{2m+1}{m(m+1)} \{ a_m(n, q) \pi_m(\cos \vartheta) + b_m(n, q) \tau_m(\cos \vartheta) \} , \\ S_2(\vartheta) &= \sum_{j=1}^{\infty} \frac{2m+1}{m(m+1)} \{ a_m(n, q) \tau_m(\cos \vartheta) + b_m(n, q) \pi_m(\cos \vartheta) \} . \end{aligned} \tag{27-56}$$

The functions $\pi_m(\cos \vartheta)$ and $\tau_m(\cos \vartheta)$ denote Legendre functions of the first kind and their first derivative

$$\begin{aligned} \pi_m(z) &= P_m(z) , \\ \tau_m(z) &= \frac{d}{dz} P_m(z) . \end{aligned} \tag{27-57}$$

Their evaluation is conveniently performed using the following recurrence relations [27-17]

$$\begin{aligned}
 (m+1)P_{m+1}(z) &= (2m+1)zP_m^1(z) - mP_{m-1}^1(z), \\
 (z^2-1)\frac{d}{dz}P_m^1(z) &= mzP_m(z) - mP_{m-1}(z).
 \end{aligned}
 \tag{27-58}$$

The Mie scattering coefficients are [27-16] for a non-magnetic medium ($\mu_r=1$)

$$a_m = \frac{n^2 j_m(nq) \frac{d}{dq} [q j_m(q)] - j_m(q) \frac{d}{dq} [nq j_m(nq)]}{n^2 j_m(nq) \frac{d}{dq} [qh_m^{(1)}(q)] - h_m^{(1)}(q) \frac{d}{dq} [nq j_m(nq)]},
 \tag{27-59}$$

$$b_m = \frac{j_m(nq) \frac{d}{dq} [q j_m(q)] - j_m(q) \frac{d}{dq} [nq j_m(nq)]}{j_m(nq) \frac{d}{dq} [qh_m^{(1)}(q)] - h_m^{(1)}(q) \frac{d}{dq} [nq j_m(nq)]},$$

with the spherical Hankel function [27-17]

$$h_m^{(1)}(z) = j_m(z) + iy_m(z) = \sqrt{\frac{\pi}{2z}} H_m^{(1)}(z)
 \tag{27-60}$$

and the spherical Bessel functions of the first and second kind following from the Bessel functions of the first and second kind according to [27-17]

$$\begin{aligned}
 j_m(z) &= \sqrt{\frac{\pi}{2z}} J_{m+1/2}(z), \\
 y_m(z) &= \sqrt{\frac{\pi}{2z}} Y_{m+1/2}(z).
 \end{aligned}
 \tag{27-61}$$

Mie scattering routines have been implemented on various platforms. For the Mie scattering examples shown here we have generated the data with Matlab™ routines published by C. Mätzler [27-42].

Figure 27-26 shows the angular distribution of scattered intensities at spheres for varying q -parameters and real refractive indices.

In accordance with the general properties of diffracted fields, the scattering distribution becomes more directional with increasing size. The scattered light which is polarized perpendicular to the scattering plane (the plane, in which the angular distribution of the scattered light is observed) has a more uniform angular distribution than the component which is polarized parallel to the scattering plane. This is in accordance with the polarization characteristics of dipole radiation to which the Mie scattering for small q reduces.

The influence of the imaginary part of the refractive index on the scattering distribution is shown in figure 27.27.

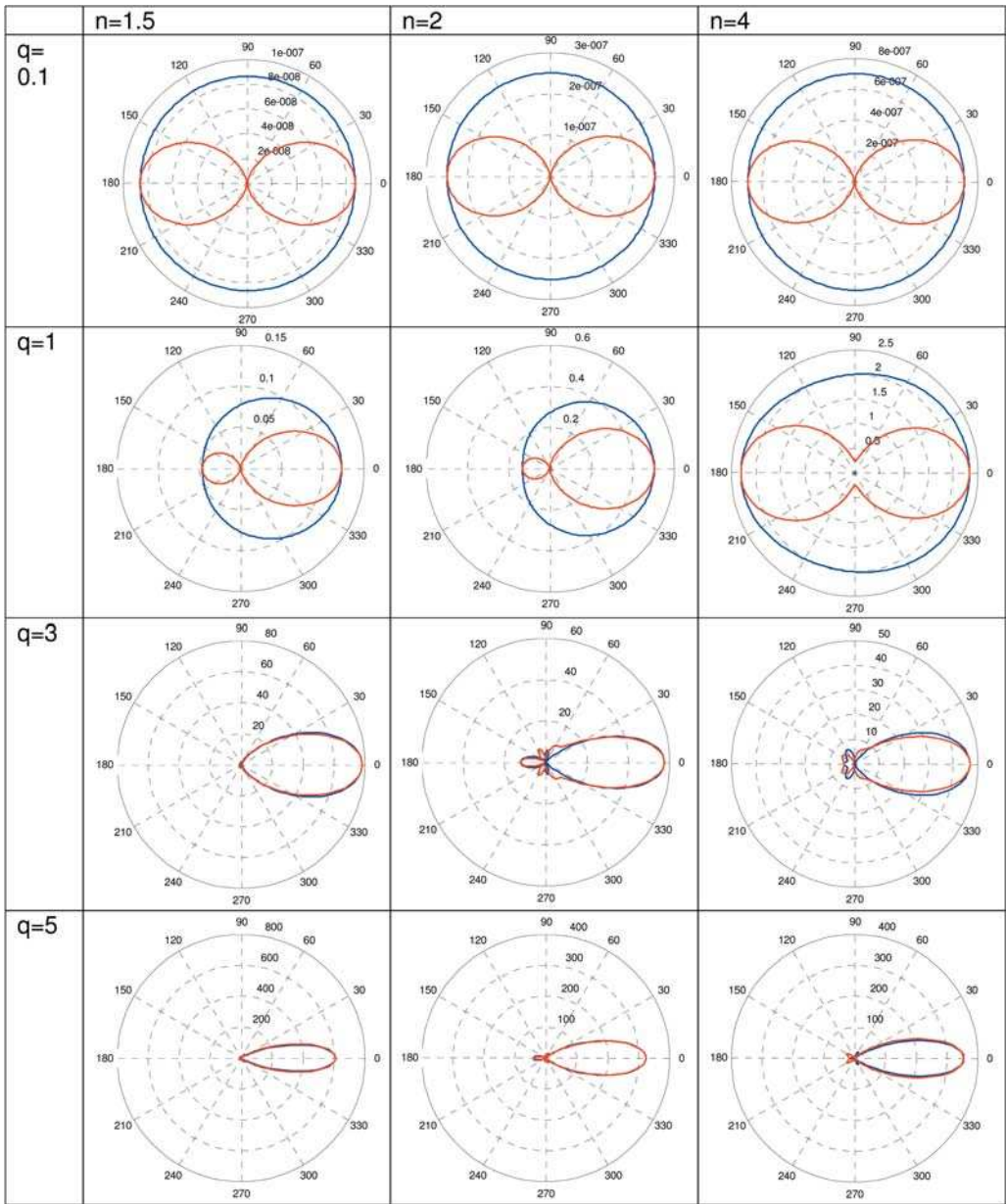


Figure 27-26: Angular distribution of scattered intensities at spheres for varying q parameters and real refractive indices. Blue: $|S_1|^2$, red: $|S_2|^2$

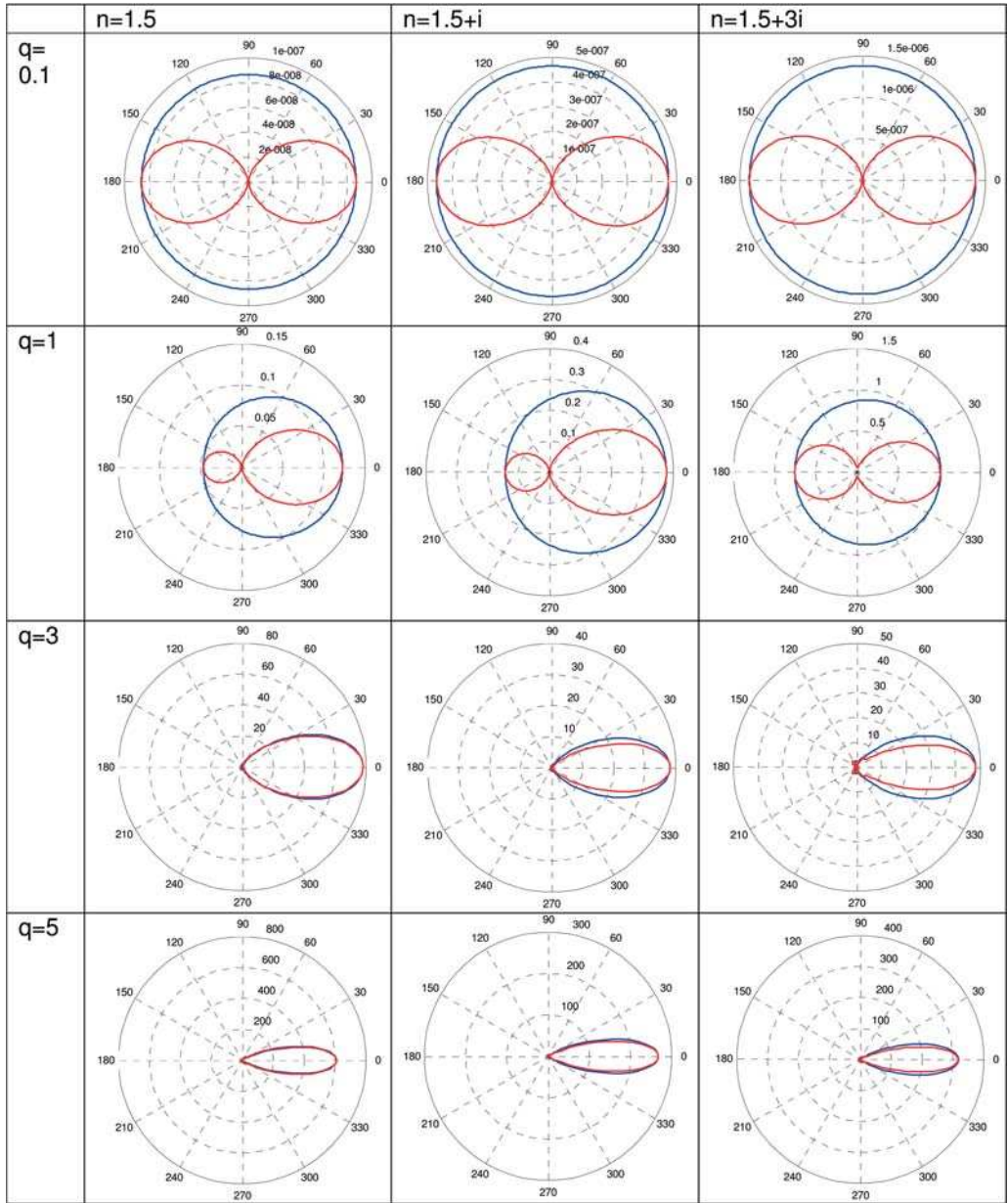


Figure 27-27: Angular distribution of scattered intensities at spheres for varying q parameters and refractive indices of increasing imaginary part. Blue: $|S_1|^2$, red: $|S_2|^2$.

For stray-light computations, the scattered efficiency defined as [27-16]

$$Q_{\text{sca}} = \frac{2}{q^2} \sum_{m=1}^{\infty} (2m+1)(|a_m|^2 + |b_m|^2) \quad (27-62)$$

and the backscattering efficiency [27-16]

$$Q_{\text{back}} = \frac{1}{q^2} \left| \sum_{m=1}^{\infty} (2m+1)(-1)^m (a_m - b_m) \right|^2 \quad (27-63)$$

are of particular interest. The scattered efficiency and the ratio of backscatter efficiency $Q_{\text{back}}/Q_{\text{sca}}$ are plotted below. The upper row was computed for pure dielectric spheres of increasing size-parameter q and refractive index n in air ($n = 1$). In the lower row the real part of the refractive index was kept constant ($n_r = 1.5$) and the imaginary part n_i was increased from 0.01 to 4.

The difference between the scattering properties of dielectric and conducting spheres is significant. Dielectric spheres show a pronounced fine structure stemming from internal resonances.

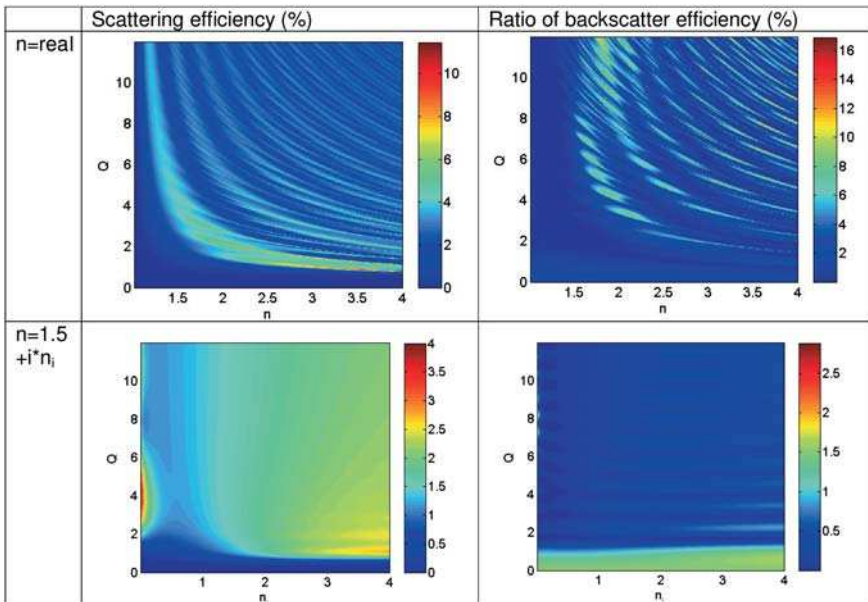


Figure 27-28: Scattering efficiency and ratio of backscattered light for an increasing real part (top) and imaginary part of the refractive index (bottom).

27.5 Numerical Methods for Arbitrary Structures

An analytical computation of diffraction fields is possible for a very limited number of structures only. For technical applications, the diffraction field of more general shapes is of importance. To this end, numerical methods have been developed that solve the Maxwell equations by numerical means for arbitrary structures. Although the number of different structure types, for which a particular method can be used to compute the diffraction field, is quite large, it is at least up to now, not possible to rely on one method only. Each method has its strengths and weaknesses that should be known in order to choose the correct one for a given problem. In the following we give a brief account of several rigorous methods.

27.6 Coupled Dipole Method

The coupled dipole method (CDM) was originally developed to compute the scattering of light at intergalactic dust clouds [27-18]. Mie scattering is not sufficient in this case because the dust particles cannot be considered as spherical.

The basic idea of the CDM is to subdivide the scattering article into an ensemble of small cells considered as elemental scatterers. In CDM an incident wave induces a dipole into each cell so that the elemental scatterers are dipoles and the scattering structure is represented by a dipole grid. The distance of the dipoles should be considerably smaller than the wavelength.

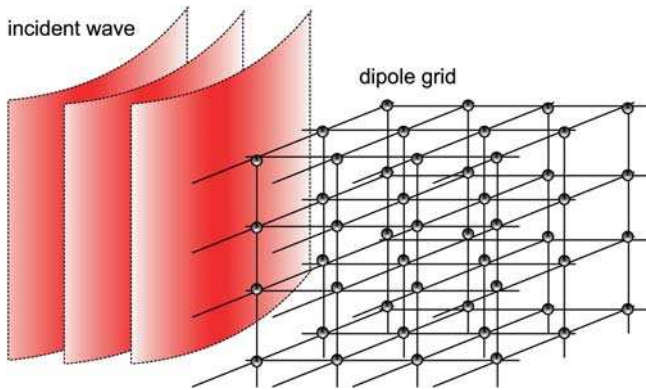


Figure 27-29: Representation of a scatterer as a dipole grid in the coupled dipole method.

The dipole-moment \vec{p} induced at a dipole at position \vec{r}_j depends on the complex amplitude of the electric field component \vec{E} and the polarizability α of the dipole.

$$\vec{p}(\vec{r}_j) = \alpha_j \vec{E}(\vec{r}_j) . \quad (27-64)$$

While α is generally a second-order tensor, it becomes scalar for isotropic particles. Its value can be computed with the help of the relation of Clausius–Mosotti [27-19]

$$\alpha = 3\epsilon_0 \frac{n^2 - 1}{n^2 + 2} V \quad (27-65)$$

where V denotes the volume of a single elemental scatterer. For complex refractive indices the polarizability can be computed from the first term of the Mie scattering expansion.

The electric field component of the field of a radiating dipole of moment \vec{p} at position \vec{r}_j observed at position \vec{r}_i is

$$\vec{E}(\vec{r}_i) = \left[k^2 \frac{(\vec{R}_{ij} \times \vec{p}_j) \times \vec{R}_{ij}}{R_{ij}^2} + \left(\frac{3\vec{R}_{ij}(\vec{R}_{ij} \cdot \vec{p}_j)}{R_{ij}^2} - \vec{p}_j \right) \left(\frac{1}{R_{ij}^2} - \frac{ik}{R_{ij}} \right) \right] \frac{e^{ikR_{ij}}}{R_{ij}} \quad (27-66)$$

with $\vec{R}_{ij} = \vec{r}_j - \vec{r}_i$ and $\vec{p}_j = \vec{p}(\vec{r}_j)$. Because this is a linear equation in p we can insert the polarizability α and the electric field \vec{E} according to eq. (27-64) to define a scattering matrix according to

$$\vec{E}(\vec{r}_i) = \mathbf{S}_{ij}(\alpha) \vec{E}(\vec{r}_j) . \quad (27-67)$$

The field at a given position \vec{r} is now the sum of the incident field \vec{E}_0 plus the contribution of all dipoles

$$\vec{E}(\vec{r}_j) = \vec{E}_0(\vec{r}) + \sum_{\substack{i=1 \\ i \neq j}}^N \mathbf{S}_{ij} \vec{E}(\vec{r}_i) . \quad (27-68)$$

To compute the complete field outside the scatterer it is necessary to know the correct field inside it. Two simple methods are available to do this computation.

- a) A self-consistency approach as originally proposed by Purcell and Penny-packer [27-18]

$$\sum_{i=1}^N (\delta_{ij} - \mathbf{S}_{ij}) \vec{E}(\vec{r}_i) = \vec{E}_0(\vec{r}_j) . \quad (27-69)$$

The equation can be solved by a self-consistency method or as a linear system of equations.

- b) A multiple-scattering approach. Here the scattering orders are introduced successively into the evaluation to obtain the field at each scatterer

$$\vec{E}(\vec{r}_j) = \vec{E}_0(\vec{r}_j) + \underbrace{\sum_{i=1}^N \mathbf{S}_{ij} \vec{E}_0(\vec{r}_i)}_{\vec{E}_1} + \underbrace{\sum_{i=1}^N \mathbf{S}_{ij} \vec{E}_1(\vec{r}_i)}_{\vec{E}_2} + \sum_{i=1}^N \mathbf{S}_{ij} \vec{E}_2(\vec{r}_i) + \dots \quad (27-70)$$

27.7
Integral Equation Approach and Moment Method

The integral equation approach generally solves the same problem as the coupled dipole method. But it assumes a more general approach based on the radiated vector potential of a current distribution.

The formulation of an integral equation for the free-space scattering problem and its solution using the moment method, dates back to Richmond [27-20]. It has been extended to general 3D structures [27-21] and surface structures [27-22], [27-23], and it was applied to the simulation of polarization-dependent interference microscopy [27-24].

27.7.1
The Moment Method

Consider an arbitrary scatterer illuminated by an incident wave (27-30).

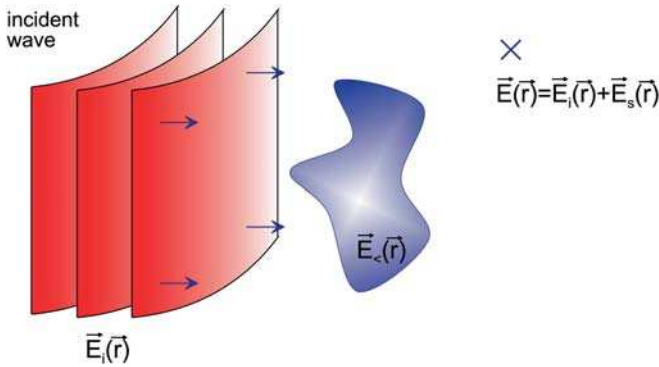


Figure 27-30: General scattering configuration.

The scattered field outside the scatterer is the sum of the incident field and the scattered field:

$$\vec{E}(\vec{r}) = \vec{E}_0(\vec{r}) + \vec{E}_s(\vec{r}) . \tag{27-71}$$

We assume that the scattered field is the result of a linear operator acting on the field in the scatterer ($\vec{r}_<$ denotes a point inside the scatterer)

$$\vec{E}_s(\vec{r}) = \mathbf{S} \vec{E}(\vec{r}_<) . \tag{27-72}$$

In the framework of the moment method, the electric field (or the current) inside the body is developed into basis functions

$$\vec{E}(\vec{r}_<) = \sum_i E_i \vec{F}_i(\vec{r}_<) . \tag{27-73}$$

With excitation coefficients E_i and an inner product

$$\langle \vec{\mathcal{W}}, \vec{E}(\vec{r}_{<}) \rangle \quad (27-74)$$

is defined, resulting in a linear system of equations

$$\langle \vec{\mathcal{W}}_j, \vec{E}(\vec{r}_{<}) \rangle = \sum_i E_i \langle \vec{\mathcal{W}}_j, \vec{E}(\vec{r}_{<}) \rangle. \quad (27-75)$$

27.7.2

Form of Scattering Operator

To obtain a specific expression for the scattering operator, we use the radiated field of a harmonic current distribution (cf. section 17.3) in a homogeneous medium

$$\vec{E}(\vec{r}) = \frac{i\omega\mu_0}{4\pi} \iint_V \vec{j}(\vec{r}') \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|} d^3\vec{r}', \quad (27-76)$$

where the current distribution is obtained from the polarization current

$$\vec{j} = \frac{\partial \vec{P}}{\partial t} = -i\omega\epsilon_0(\epsilon_r - 1)\vec{E}. \quad (27-77)$$

Resulting in

$$\vec{E}(\vec{r}) = \frac{ik^2}{4\pi} \iint_V (\epsilon(\vec{r}') - 1)\vec{E}(\vec{r}') \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|} d^3\vec{r}' \quad (27-78)$$

i.e., \mathbf{S} is a linear integral operator

$$\mathbf{S}(\vec{E}) = \frac{ik^2}{4\pi} \iint_V (\epsilon(\vec{r}') - 1)\vec{E}(\vec{r}') \frac{e^{ik|\vec{r}-\vec{r}'|}}{|\vec{r}-\vec{r}'|} d^3\vec{r}'. \quad (27-79)$$

The scattering operator yields the electric field outside a volume V as a function of its dielectric constant and the electric field inside it. An arbitrary medium in which the scattering occurs can be considered by inserting the appropriate Green's tensor $G(r, r')$, resulting in

$$\mathbf{S}(\vec{E}) = \frac{ik^2}{4\pi} \iint_V (\epsilon(\vec{r}') - 1)\vec{E}(\vec{r}')\mathbf{G}(\vec{r}, \vec{r}')d^3\vec{r}'. \quad (27-80)$$

27.7.3

Scattering in Three-layer Medium

In optical microscopy, structures are situated on planar substrates. Because of this the cover-film-substrate configuration is of some interest. To do imaging simulations of this kind of object [27-24], consider an arbitrary structure that is situated in a stratified medium consisting of three layers parallel to the x -axis (Figure 27-31).

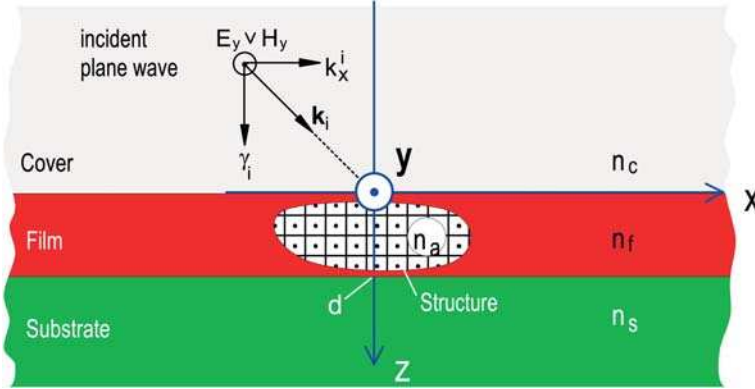


Figure 27-31: Geometry for scattering calculation according to the integral equation approach.

A film of thickness d and refractive index n_f lies on a substrate of refractive index n_s ; it is covered by a medium of refractive index n_c . The structure of refractive index distribution $n_a(\vec{r})$ is situated entirely inside the film. We assume non-magnetic materials ($\mu_r=1$) and a monochromatic, linearly polarized, obliquely incident electromagnetic wave of frequency ω with the field components

$$\begin{aligned}\vec{E}_i(\vec{r}) &= \vec{E}_i e^{i\vec{k}_i \cdot \vec{r}}, \\ \vec{H}_i(\vec{r}) &= \frac{1}{k_0 z_0} (\vec{k}_i \times \vec{E}_i) e^{i\vec{k}_i \cdot \vec{r}}\end{aligned}\quad (27-81)$$

and the 2D coordinate $\vec{r} = (x, 0, z)$. k_i denotes the propagation constant and z_0 is the free-space impedance. Two linear polarizations are considered: In TM-polarization \vec{E}_i has a y component only and in TE-polarization \vec{H}_i is parallel to the y -axis. Arbitrary incident fields can be dealt with by a Fourier decomposition.

The diffraction near-field (E, H) of the incident plane wave in the considered system is a superposition of a background field (E_0, H_0) that is due to the undisturbed three-layer medium without a scattering structure and the scattered field (E_s, H_s).

The electric field distribution at an arbitrary point is obtained by a solution of the integral equation

$$\vec{E}(\vec{r}) = \vec{E}_0(\vec{r}) + \int_{\text{structure}} \int \mathbf{G}(\vec{r}, \vec{r}') \Delta\epsilon(\vec{r}') \vec{E}(\vec{r}') d^2 \vec{r}' \quad (27-82)$$

where $\Delta\varepsilon(\vec{r})$ is the normalized material contrast due to the scatterer. \vec{E}_0 denotes the background field that is obtained for the undisturbed stratified medium. $\mathbf{G}(\vec{r}, \vec{r}')$ is Green's tensor (dyadic Green's function) of the three-layer medium. In the following we construct the three-layer media Green's tensor from the free-space Green's tensor of which, in a source-free region, the Fourier expansion becomes

$$\begin{aligned}\mathbf{G}(\vec{r}, \vec{r}') &= \frac{i}{4} \left[\nabla \nabla \cdot + k_f^2 \mathbf{I} \right] \vec{H}_0(k_f |\vec{r} - \vec{r}'|) \\ &= \frac{i}{4\pi} \left[\nabla \nabla \cdot + k_f^2 \mathbf{I} \right] \int_{-\infty}^{\infty} \frac{1}{\gamma_f} \exp\{i\gamma_f |z - z'|\} e^{ik_x(x-x')} dk_x\end{aligned}\quad (27-83)$$

where \mathbf{I} denotes the unity dyade, $\nabla \nabla \cdot$ denotes the grad-div operation, and γ_f denotes the z component of the propagation vector in the film. Performing the derivatives under the integral we obtain

$$\mathbf{G}(\vec{r}, \vec{r}') = \frac{i}{4\pi} \int_{-\infty}^{\infty} \left[k_f^2 \mathbf{I} - \vec{s}(\vec{s}) \right] \frac{1}{\gamma_f} \exp\{i\gamma_f |z - z'|\} e^{ik_x(x-x')} dk_x \quad (27-84)$$

with

$$\vec{s} = \begin{pmatrix} k_x \\ 0 \\ \text{sgn}(z - z')\gamma_f \end{pmatrix} \quad (27-85)$$

where $\text{sgn}(z)$ denotes the signum-function ($\text{sgn}(z) = 1, 0, -1$ for $z > 0, = 0, < 0$, respectively).

Now, the kernel of the integral (27-84) describes the propagation of the angular spectrum of the radiation field of a point-like source from the plane $z = z'$ to the plane z in free space. In order to describe wave propagation in a three-layer stack, we have to modify the propagation operator $\exp\{i\gamma_f |z - z'|\}$. Including the multiple reflected fields between both interfaces and the propagation in the different media we obtain for the Green's tensor

$$\mathbf{G}(\vec{r}, \vec{r}') = \int_{-\infty}^{\infty} \mathbf{g}(k_x, z, z') e^{ik_x(x-x')} dk_x \quad (27-86)$$

with

$$\mathbf{g}(k_x, z, z') = \begin{pmatrix} \gamma_f P_{11} & 0 & -k_x \text{sgn}(z - z') P_{-11} \\ 0 & k_f^2 \gamma_f^{-1} P_{11} & 0 \\ -k_x \text{sgn}(z - z') P_{-1-1} & 0 & k_x^2 \gamma_f^{-1} P_{1-1} \end{pmatrix} \quad (27-87)$$

and

$$P_{f\bar{g}}(k_x, z, z_m) = \begin{cases} [S_f e^{i\gamma_f |z_m|} + R_{f\bar{g}}(k_x, 0, z_m)] e^{i\gamma_c |z_m|} & : \text{cover} \\ S_f e^{i\gamma_f z - z_m} + R_{f\bar{g}}(k_x, z, z_m) & : \text{film} \\ [S_f e^{i\gamma_f |d - z_m|} + R_{f\bar{g}}(k_x, d, z_m)] e^{i\gamma_s |z_m|} & : \text{substrate} \end{cases} \quad (27-88)$$

where γ_f and γ_s are the propagation constants in the film and substrate defined analogously to γ_c . s_f is due to the signum function with

$$s_f = \begin{cases} 1 & : f = 1 \\ \text{sgn}(z - z_m) & : f = -1. \end{cases} \quad (27-89)$$

The reflection term $R_{f\bar{g}}$ becomes according to an elementary calculation

$$R_{f\bar{g}}(k_x, z, z_m) = \frac{1}{1 - r_c r_s e^{i\gamma_f 2d}} [f g r_c e^{i\gamma_f (z_m + z)} + g r_s e^{i\gamma_f (2d - z_m - z)} + r_c r_s (e^{i\gamma_f (2d - z_m + z)} + f e^{i\gamma_f (2d + z_m - z)})] \quad (27-90)$$

r_c and r_s are the polarization-dependent Fresnel reflection coefficients of the film-cover and the film-substrate interface. The factor g takes into account that the reflection coefficients have to be replaced by their negative values for E_x and H_x .

Having the appropriate Green's tensor, it remains to solve the integral equation (27-82) in order to obtain the electric field distribution inside the structure. For this purpose (27-82) is converted into a linear system of equations using the puls-function point-matching version of the moment method: The structure is divided into N quadratic cells with a cross-section small enough that the electric field within a single cell can be considered constant. The integral over the structure is then replaced by a sum over the cells. The fields are equated for the center of each cell and the integral of the single cell is evaluated in a closed form by replacing the square cells with circular cells of equal area [27-20].

Special care has to be taken of the diagonal elements. They describe the contribution of the scattered field of a single cell to the field in the center of the cell itself. This field consists of two parts: The free-space contribution from the principal-value representation of the scattered field integral [27-25] and a reflected field contribution. The linear system of equations has the structure [27-24]

$$\sum_{\nu=1}^N S_{\nu\mu} E_y(\vec{r}_\nu) = E_y^0(\vec{r}_\mu), \quad \mu = 1, \dots, N \quad (27-91)$$

for TM-polarization and

$$\sum_{\nu=1}^N [a_{\nu\mu} E_x(\vec{r}_\nu) + b_{\nu\mu} E_z(\vec{r}_\nu)] = E_x^0(\vec{r}_\mu) \quad (27-92)$$

$$\sum_{\nu=1}^N [c_{\nu\mu} E_x(\vec{r}_\nu) + d_{\nu\mu} E_z(\vec{r}_\nu)] = E_z^0(\vec{r}_\mu)$$

for TE-polarization.

In order to circumvent problems with a large condition number of the moment method matrix in TE-polarization, it is useful to perform a direct integration over a single cell for cell-pairs that are three cells or less apart [27-26]. After calculation of the electric field within the structure the scattered E_y component (TE-pol.) or H_y component (TM-pol.) is calculated outside.

The numerical properties of the computation method are mainly determined by the applied algorithm for solution of the linear system of equations. The LU-decomposition [27-27] is a well known, stable algorithm that is particular useful if we wish to calculate the near-fields for different angles of incidence simultaneously.

In particular, for large problems, it is advantageous to use an iterative solution method like the conjugate gradient method (CGM) [27-27] and to store only a part of the matrix and derive the rest from the symmetry properties – or even to compute the matrix elements for each iteration anew. Although iterative methods may drastically reduce the memory requirements, they do, of course, increase the computation time. A survey and comparison of appropriate numerical methods for large systems is given in ref. [27-28].

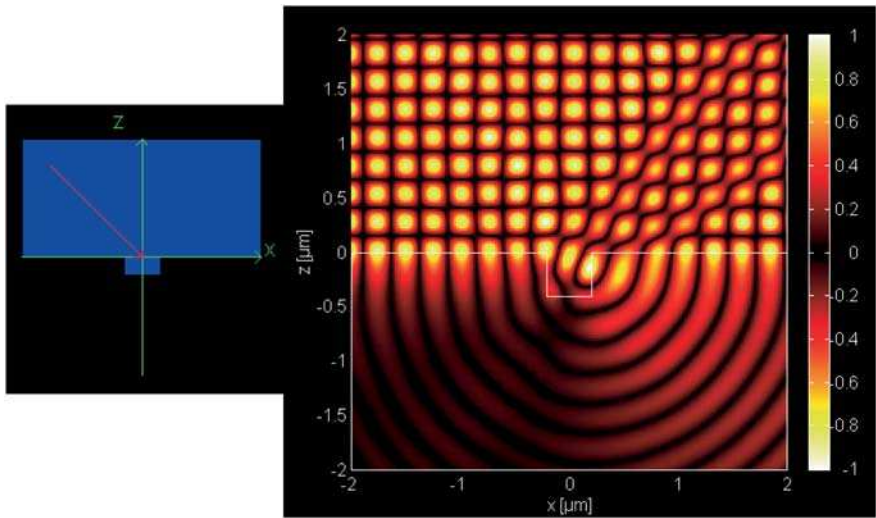


Figure 27-32: Scattering of an evanescent wave by a 400×400 nm protrusion on a quartz surface. $\lambda = 549$ nm. Angle of incidence = 45° . TE-polarization. The amplitude of the real part of E_y is shown.

As an example, Figure 27-32 shows the real part of the near-field for the scattering of an evanescent wave at a glass-air interface by a 400×400 nm² obstacle. The inten-

sity above the interface basically shows the intensity nodes due to the interference of the incident and reflected wave, below the evanescent wave and the free-space scattered field due to the obstacle. Below, a standing wave pattern has formed parallel to the interface, consisting of the evanescent field and the free-space field scattered by the obstacle.

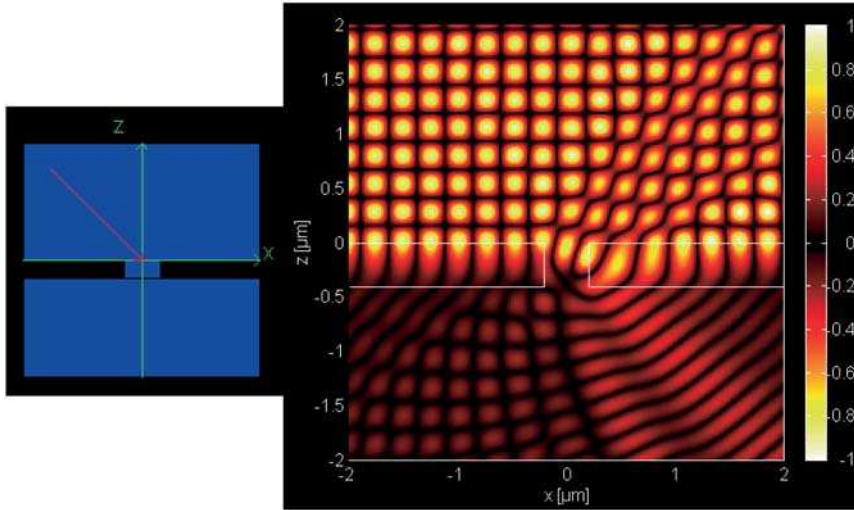


Figure 27-33: Scattering and frustrated total reflection. $\lambda = 549$ nm. TE-polarization angle of incidence = 45° . The amplitude of the real part of E_y is shown.

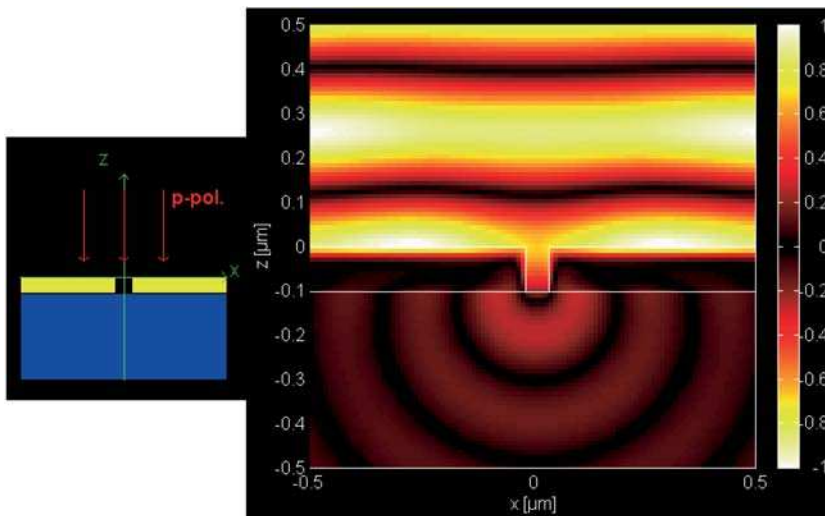


Figure 27-34: Transmission through 50 nm slit in a chromium layer on quartz in TM-polarization. $\lambda = 549$ nm. Angle of incidence = 45° . The amplitude of the real part of H_y is shown. For TE-polarization, no transmission is observed.

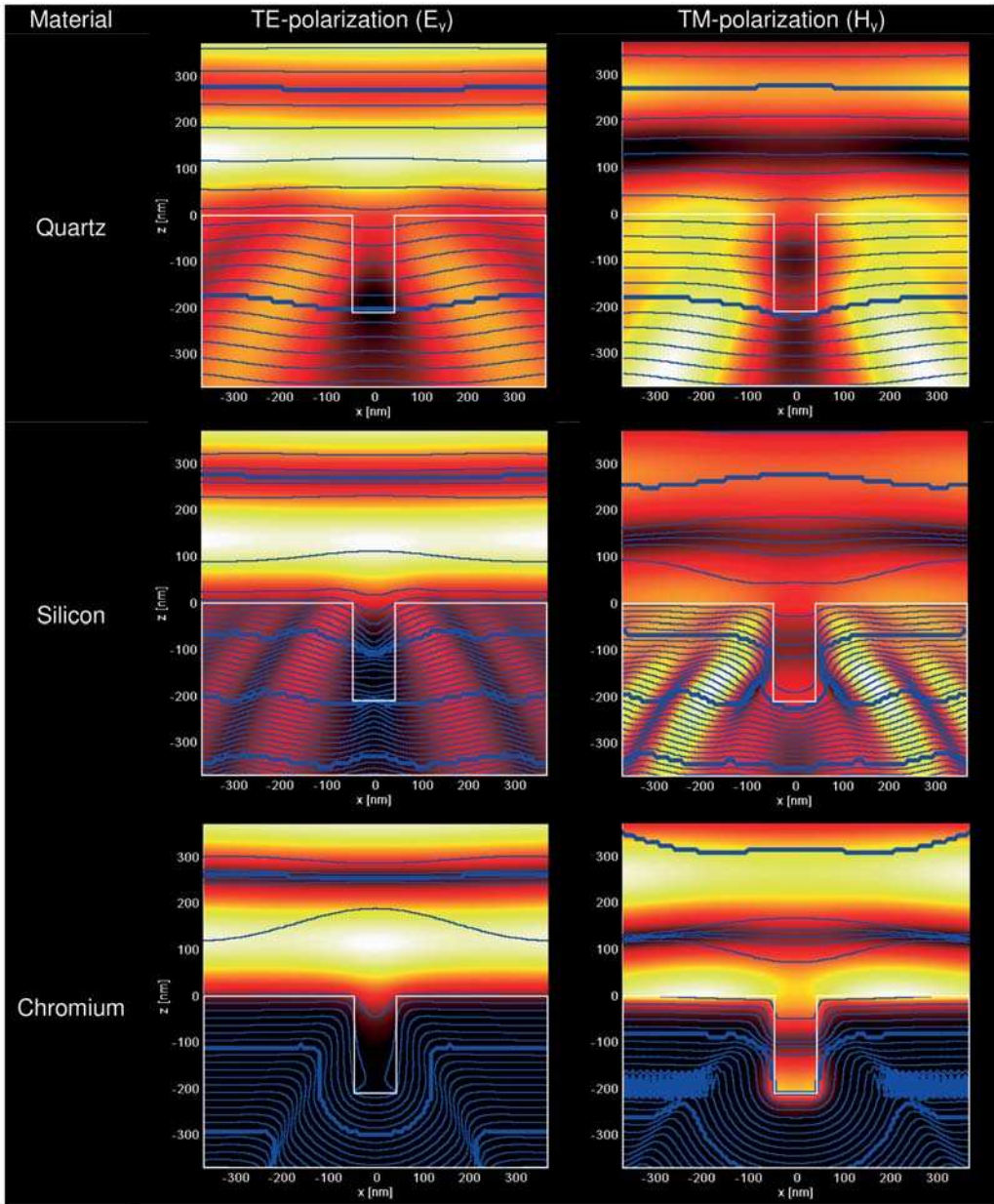


Figure 27-35: Comparison of near-field plots for a 100 nm wide and 200 nm deep groove in a bulk material for TE (left) and TM (right)-polarization. $\lambda = 549$ nm. Perpendicular incident plane wave. Color-map: amplitude; blue lines: phase.

A slightly altered configuration is shown in Figure 27-33. Now a second air-quartz interface is placed below the structure. The distance to the first interface is smaller than the wavelength and frustrated total reflection occurs. You will see that, on the left side, a part of the evanescent wave is actually coupled into the second quartz-block.

The last image in this series (Figure 27-34) shows the transmission of a perpendicularly incident plane wave through a sub-wavelength slit (width= 50 nm at $\lambda= 549$ nm). You can see the standing wave that has formed on top of the chromium layer, the penetration of the light into the chromium due to its finite skin depth (stemming from the finite conductivity), and the cylindrical wave that is radiated by the slit below it.

An example of the near-field representation by amplitude and phase is shown in Figure 27-35.

27.8 Fourier Modal Method

The Fourier modal method (FMM), also called rigorous coupled wave analysis (RCWA), is a rigorous method of computing the diffraction of electromagnetic waves at arbitrary gratings [27-29][27-30].

27.8.1

Theory

The theoretical basis of the RCWA is the Floquet theorem, according to which the propagation vector of electromagnetic waves in periodical structures cannot take arbitrary values along the grating vector. It can take only values obtained from the harmonics of the grating's periodicity; i.e., for the grating depicted in Figure 27-36, we obtain for an incident plane wave with the propagation vector k_i (conical diffraction)

$$k_{xv} = k_{xi} + v \frac{2\pi}{\Lambda} . \quad (27-93)$$

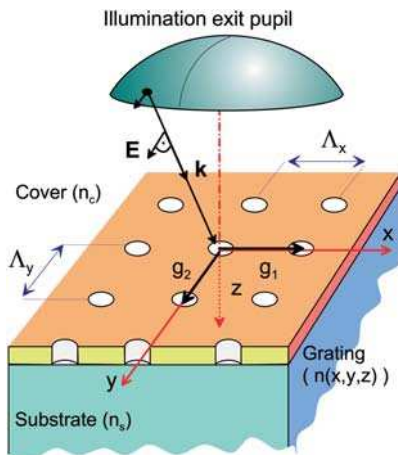


Figure 27-36: Grating geometry for a single layer.

For a two-dimensional grating we have accordingly

$$\vec{k}_{\nu\mu} = \begin{pmatrix} k_{x\nu} \\ k_{y\mu} \end{pmatrix} = \vec{k}_i + \begin{pmatrix} \nu \frac{2\pi}{\Lambda_x} \\ \mu \frac{2\pi}{\Lambda_y} \end{pmatrix}. \quad (27-94)$$

A condensed form is obtained by defining the harmonics of the grating vector as

$$\vec{k}_{\nu\mu} = \begin{pmatrix} k_{x\nu} \\ k_{y\mu} \end{pmatrix} = \vec{k}_i + \begin{pmatrix} \nu \frac{2\pi}{\Lambda_x} \\ \mu \frac{2\pi}{\Lambda_y} \end{pmatrix} \quad (27-95)$$

yielding the allowed propagation vectors

$$\vec{k}_{\nu\mu} = \vec{k}_i + \vec{g}_{\nu\mu}. \quad (27-96)$$

The complete propagation vector is obtained from the refractive index of the medium in which the wave is propagating

$$\vec{k}_{\nu\mu} = \begin{pmatrix} k_{\nu\mu}^{x\perp} \\ k_{\nu\mu}^{y\perp} \\ \sqrt{n^2 - |\vec{k}_{\nu\mu}^\perp|^2} \end{pmatrix}. \quad (27-97)$$

While the propagation directions of the plane waves are fixed by geometry, the computation of the complex amplitude and polarization of the diffracted orders is more complex.

The grating region of periodicity (Λ_x, Λ_y) is divided into films with the dielectric constant $\varepsilon(z) = \text{const}$. The dielectric constant and fields are expanded into space harmonics for each film. The expansion into a Fourier series is truncated at a certain maximum index M .

$$\begin{aligned} \varepsilon(\vec{r}_\perp) &= \sum_{\nu,\mu=-M}^M \varepsilon_{\nu\mu} \exp\{i2\pi\vec{g}_{\nu\mu} \cdot \vec{r}_\perp\} \\ E_x(\vec{r}_\perp, z) &= \sum_{\nu,\mu=-M}^M S_{\nu\mu}^x(z) \exp\{i\vec{k}_{\nu\mu}^\perp \cdot \vec{r}_\perp\} \\ H_x(\vec{r}_\perp, z) &= \frac{i}{Z_0} \sum_{\nu,\mu=-M}^M U_{\nu\mu}^x(z) \exp\{i\vec{k}_{\nu\mu}^\perp \cdot \vec{r}_\perp\} \end{aligned} \quad (27-98)$$

where the free-space impedance is Z_0 . For E_y and H_y similar expansions hold. The z components are obtained from the Maxwell equations.

Using the correct Fourier factorization [27-33], Maxwell's equations yield the following $(4M \times 4M)$ matrix equation for the components,

$$\begin{pmatrix} \frac{\partial \vec{S}_y}{\partial z} \\ \frac{\partial \vec{S}_x}{\partial z} \\ \frac{\partial \vec{U}_y}{\partial z} \\ \frac{\partial \vec{U}_x}{\partial z} \end{pmatrix} = k_0 \begin{pmatrix} 0 & 0 & \mathbf{K}_y \mathbf{E}^{-1} \mathbf{K}_x & \mathbf{I} - \mathbf{K}_y \mathbf{E}^{-1} \mathbf{K}_y \\ 0 & 0 & \mathbf{K}_x \mathbf{E}^{-1} \mathbf{K}_x - \mathbf{I} & -\mathbf{K}_x \mathbf{E}^{-1} \mathbf{K}_y \\ \mathbf{K}_x \mathbf{K}_y & \mathbf{A}^{-1} - \mathbf{K}_y^2 & 0 & 0 \\ \mathbf{K}_x^2 - \mathbf{E} & -\mathbf{K}_x \mathbf{K}_y & 0 & 0 \end{pmatrix} \begin{pmatrix} \vec{S}_y \\ \vec{S}_x \\ \vec{U}_y \\ \vec{U}_x \end{pmatrix} \quad (27-99)$$

where \vec{S}_x , \vec{S}_y , \vec{U}_x , and \vec{U}_y are vectors formed from the coefficients of the Fourier series, \mathbf{K}_x is the diagonal matrix with elements $k_{x\nu\mu}$, \mathbf{I} is the unity-matrix, \mathbf{E} is the matrix formed from $\varepsilon_{\nu-\nu'}$, $\mu-\mu'$ and \mathbf{A} the matrix formed of $\varepsilon^{-1}_{\nu-\nu'}$, $\mu-\mu'$.

Forming the second derivation with respect to z and inserting the resulting first derivations yields

$$\begin{pmatrix} \frac{\partial^2 \vec{S}_y}{\partial z^2} \\ \frac{\partial^2 \vec{S}_x}{\partial z^2} \end{pmatrix} = k_0^2 \begin{pmatrix} \mathbf{K}_x^2 + \mathbf{D}\mathbf{E} & \mathbf{K}_y (\mathbf{E}^{-1} \mathbf{K}_x \mathbf{A}^{-1} - \mathbf{K}_x) \\ \mathbf{K}_x (\mathbf{E}^{-1} \mathbf{K}_y \mathbf{E} - \mathbf{K}_y) & \mathbf{K}_y^2 + \mathbf{B}\mathbf{A}^{-1} \end{pmatrix} \begin{pmatrix} \vec{S}_y \\ \vec{S}_x \end{pmatrix}, \quad (27-100)$$

$$\begin{pmatrix} \frac{\partial^2 \vec{U}_y}{\partial z^2} \\ \frac{\partial^2 \vec{U}_x}{\partial z^2} \end{pmatrix} = k_0^2 \begin{pmatrix} \mathbf{K}_y^2 + \mathbf{A}^{-1} \mathbf{B} & (\mathbf{K}_x - \mathbf{A}^{-1} \mathbf{K}_x \mathbf{E}^{-1}) \mathbf{K}_y \\ (\mathbf{K}_y - \mathbf{E} \mathbf{K}_y \mathbf{E}^{-1}) \mathbf{K}_x & \mathbf{K}_x^2 + \mathbf{E}\mathbf{D} \end{pmatrix} \begin{pmatrix} \vec{U}_y \\ \vec{U}_x \end{pmatrix}.$$

with

$$\begin{aligned} \mathbf{B} &= \mathbf{K}_x \mathbf{E}^{-1} \mathbf{K}_x - \mathbf{I}, \\ \mathbf{D} &= \mathbf{K}_y \mathbf{E}^{-1} \mathbf{K}_y - \mathbf{I}. \end{aligned} \quad (27-101)$$

It is sufficient to solve one of the matrix equations above and to obtain the other component by application of a linear differential equation. Solving (27-100) for \vec{S} yields the eigenvalues λ_m ($m = -N \dots N$) and to each eigenvalue the eigenvector \vec{W}_m with

$$\begin{pmatrix} S_{y\nu} \\ S_{x\nu} \end{pmatrix} = \sum_{m=-N}^N W_{\nu m} \{ c_m^+ e^{-k_0 q_m z} + c_m^- e^{k_0 q_m (z-d)} \}, \quad (27-102)$$

where q_m denotes the positive square roots of the eigenvalues

$$q_m = +\sqrt{\lambda_m} \quad (27-103)$$

and c_m^{+-} are the excitation coefficients of the modes in the grating region.

In a compact matrix notation, the electric field components are

$$\begin{pmatrix} \vec{S}_y \\ \vec{S}_x \end{pmatrix} = \mathbf{W}(\phi^- \vec{c}^+ + \phi^+ \vec{c}^-) \quad (27-104)$$

with the matrices

$$\begin{aligned} \mathbf{W} &= [w_{vm}], \\ \vec{c}^+ &= [c_m^+], \quad \vec{c}^- = [c_m^-], \\ \phi^+ &= e^{k_0 Q z}, \quad \phi^- = e^{k_0 Q(z-d)}, \\ \mathbf{Q} &= [q_m]. \end{aligned} \quad (27-105)$$

U follows from the solution of

$$\begin{pmatrix} \frac{\partial \vec{S}_y}{\partial z} \\ \frac{\partial \vec{S}_x}{\partial z} \end{pmatrix} = k_0 \begin{pmatrix} \mathbf{K}_y \mathbf{E}^{-1} \mathbf{K}_x & -\mathbf{D} \\ \mathbf{B} & -\mathbf{K}_x \mathbf{E}^{-1} \mathbf{K}_y \end{pmatrix} \begin{pmatrix} \vec{U}_y \\ \vec{U}_x \end{pmatrix} = \Omega \begin{pmatrix} \vec{U}_y \\ \vec{U}_x \end{pmatrix} \quad (27-106)$$

to

$$\begin{pmatrix} \vec{U}_y \\ \vec{U}_x \end{pmatrix} = \Omega^{-1} \begin{pmatrix} \frac{\partial \vec{S}_y}{\partial z} \\ \frac{\partial \vec{S}_x}{\partial z} \end{pmatrix} \quad (27-107)$$

i.e.,

$$\begin{pmatrix} \vec{U}_y \\ \vec{U}_x \end{pmatrix} = \Omega^{-1} \mathbf{W}(-\mathbf{Q} \phi^- \vec{c}^+ + \mathbf{Q} \phi^+ \vec{c}^-). \quad (27-108)$$

Depending on the geometry of the diffracting structure and the illumination, different simplified eigen equations can be derived from (27-99). For a linear grating (grating vector in the x direction) with conical illumination, for instance, $\mathbf{K}_y = k_y \mathbf{I}$, and (27-99) reduces to two $(M \times M)$ second-order differential equations

$$\begin{aligned} \frac{\partial^2 \vec{U}_x}{\partial z^2} &= k_0^2 (\mathbf{K}_y^2 + \mathbf{K}_x^2 - \mathbf{E}) \vec{U}_x, \\ \frac{\partial^2 \vec{S}_x}{\partial z^2} &= k_0^2 (\mathbf{K}_y^2 + \mathbf{K}_x \mathbf{E}^{-1} \mathbf{K}_x \mathbf{A}^{-1} - \mathbf{A}^{-1}) \vec{S}_x. \end{aligned} \quad (27-109)$$

For a line grating in non-conical illumination, i.e., purely TE or TM-polarized, a further simplification is obtained. For TE-polarization ($\vec{E} = [0, E_y, 0]$) it is sufficient to solve

$$\frac{\partial^2 \vec{S}_y}{\partial z^2} = k_0^2 (\mathbf{K}_x^2 - \mathbf{E}) \vec{S}_y \quad (27-110)$$

and for TM-polarization ($\vec{H} = [0, H_y, 0]$)

$$\frac{\partial^2 \vec{U}_y}{\partial z^2} = k_0^2 \mathbf{A}^{-1} (\mathbf{K}_x \mathbf{E}^{-1} \mathbf{K}_x - \mathbf{I}) \vec{U}_y. \quad (27-111)$$

The various levels of simplification are sketched in Figure 27-37.

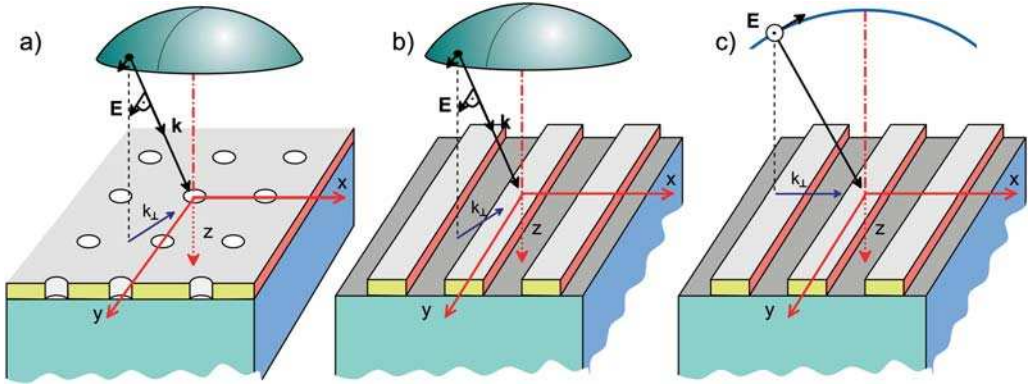


Figure 27-37: Increasing levels of simplification for diffraction computations.

a) Conical diffraction at 2D grating; b) conical diffraction at 1D grating; c) non-conical diffraction (s- and p-polarization parallel to TE and TM-polarization) at 1D-grating.

For a complete solution of the diffraction problem, the modes of the films must be coupled together and to the free-space modes. Basis for the coupling is the continuity of the tangential fields, i.e., the projection of the field vectors onto the interfaces. The enhanced transmittance matrix approach [27-30] is easy to implement and avoids numerical instabilities. For the computation of the tangential fields, a normalized vector perpendicular to the plane of incidence is needed

$$\vec{\kappa} = \frac{\vec{k} \times \vec{k}_\perp}{|\vec{k} \times \vec{k}_\perp|} = \frac{1}{\sqrt{k_x^2 + k_y^2}} \begin{pmatrix} -k_y \\ k_x \\ 0 \end{pmatrix}. \quad (27-112)$$

Now the tangential field components are just the projection of the actual field onto κ

$$\vec{E}_\perp = (\vec{\kappa} \cdot \vec{E}) \vec{\kappa}, \quad \vec{H}_\perp = (\vec{\kappa} \cdot \vec{H}) \vec{\kappa}. \quad (27-113)$$

The diffracted orders obtained by this means are $\vec{E}_\perp(\vec{k}_{\nu\mu}^{\perp})$ and $\vec{H}_\perp(\vec{k}_{\nu\mu}^{\perp})$. Because both are defined relative to the propagation vector, a transform into global coordinates is needed to sum up the fields of several propagation directions as they occur in near-field and image computations. The required transform is obtained from three constituting equations (the indices $\nu \mu$ are omitted for clarity)

$$\begin{aligned}
E_{\perp} &= \vec{\kappa} \cdot \vec{E} \\
H_{\perp} &= \vec{\kappa} \cdot \left(Z_0^{-1} k_0^{-1} \vec{\kappa} \times \vec{E} \right) \\
0 &= \vec{k} \cdot \vec{E}
\end{aligned} \tag{27-114}$$

which provide, after a straightforward arrangement of terms,

$$\begin{pmatrix} -\frac{k_y}{|\vec{k}_{\perp}|} & \frac{k_x}{|\vec{k}_{\perp}|} & 0 \\ \frac{k_x \gamma}{Z_0 k_0 |\vec{k}_{\perp}|} & \frac{k_y \gamma}{Z_0 k_0 |\vec{k}_{\perp}|} & -\frac{|\vec{k}_{\perp}|}{Z_0 k_0} \\ k_x & k_y & \gamma \end{pmatrix} \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} = \begin{pmatrix} E_{\perp} \\ H_{\perp} \\ 0 \end{pmatrix} \tag{27-115}$$

from which the electric field distribution $E(\rho_x, \rho_y)$ is computed by inversion.

The corresponding magnetic field is

$$\vec{H} = (k_0 Z_0)^{-1} \vec{k} \times \vec{E}. \tag{27-116}$$

27.8.2

Diffraction Efficiency

The rate of the incident energy that is diffracted into a certain order is the diffraction efficiency of the grating for this order. It depends on the angles of incidence and diffraction. For conical diffraction it is obtained from

$$\eta = |E_{\perp}|^2 \operatorname{Re} \left\{ \frac{\gamma}{\gamma_i} \right\} + |H_{\perp}|^2 \operatorname{Re} \left\{ \frac{n_1^2 \gamma}{n_{12}^2 \gamma_i} \right\} \tag{27-117}$$

with $n_{12} = n_1$ for reflection and $n_{12} = n_2$ for transmission. The diffraction efficiency takes into account that the flow of energy into and from the planar structure region depends on the angle of propagation and the change in the refractive index. The energy of a diffracted component is equal to the diffraction efficiency times the energy of the incident wave.

27.9

Finite-difference Method

In the wave equation, the propagation of the field amplitude is determined by the corresponding derivatives of the field after the coordinates. A direct method of solution approximates these derivatives by finite differences on a discrete grid. The remaining problem is then to solve a system of linear equations for the field components on all the grid points, taking the boundary conditions into account.

A typical discretization scheme is the approximation of the derivatives by symmetric differences according to the equation [27-36][27-37][27-38]

$$u_x = \frac{u_{i+1} - u_{i-1}}{2 \Delta x} \quad (27-118)$$

for the first, and

$$u_{xx} = \frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta x^2} \quad (27-119)$$

for the second derivative with the field component u and the transverse discretization step Δx . The terms in the wave equation couple the field components in the transverse and in the axial direction. The order of approximation of the derivatives determines the number of coupled field points and therefore the bandwidth of the resulting system matrices. In most cases, the approximation of a slowly-varying envelope is made and the derivative of second order to the propagation coordinate z is neglected. Then it is possible to calculate the field at the coordinate $z + \Delta z$ for the transverse coordinate x from the fields of the positions at $x - \Delta x$, x and $x + \Delta x$ for the forgoing propagation step z . This is the easiest way to calculate the field and is called the explicit difference scheme. It is illustrated in figure 27-38 on the upper side. As is well known from the theory of numerical solution of partial differential equations, an explicit scheme accumulates the errors at every z -step and therefore the propagation over large distances cannot be calculated with sufficient accuracy. The so called implicit scheme on the other side is always stable, if the parameters are adjusted well. For this method, the calculation uses the inverse coupling of three values of the step at $z + \Delta z$ to one field at the position z . If the corresponding equations are put together, a system of linear equations results and these can be solved by linear algebra with the inversion of the system matrix.

If the coordinate system is chosen correctly, the wavefront is approximately oriented tangential to the lateral coordinate x . The direction of propagation is then nearly oriented in the z direction. In this case, the dominant part of the field at a new grid location point results from the field of the corresponding point with the same coordinate x or index j . The influence of the neighboring locations with the indices $j - 1$ and $j + 1$ is small. The influence of the more distant point with indices $j - 1$, $j + 2$ and so on, can then be neglected. The characteristic quantity to describe an elementary grid cell is the Fresnel number of this small area and is defined by

$$N_F = \frac{\Delta x^2}{\lambda \cdot \Delta z} \quad (27-120)$$

The calculation of the numerical beam propagation is stable and shows a good conditioning, if the Fresnel number has a large value in comparison to one. The corresponding system of linear equations is then diagonal dominant and performs a stable inversion. The stability of the chosen method can always be influenced by changing the transverse grid size Δx of the propagation step-width Δz .

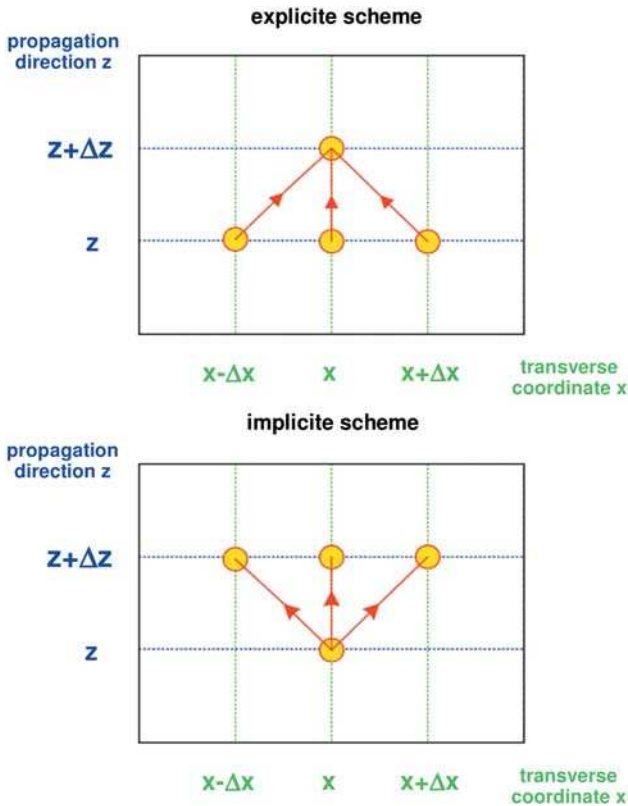


Figure 27-38: Explicit and implicit differencing scheme for the calculation of the field at the propagation from z to $z + \Delta z$.

So the most critical aspects for calculation schemes with finite difference are field distributions with steep gradients or steps, and the behaviour at the boundary.

27.9.1

Boundary Conditions

The interaction of the field with the boundary has to be handled with care. For a free-space propagation without limiting stops or interfaces, it is reasonable to expand the calculation grid, so that the field strength at the outer points is absolutely negligible. Alternatively, special boundary conditions can be formulated, which prevent reflections and interferences [27-39].

1. Absorbing Boundaries

Absorbing boundaries force the field to decrease to zero at the border of the calculation region. This prevents the reflection of outgoing field components back into the propagation area. A usual approach to generate this behavior is

to add an imaginary part to the refractive index in the outer region of the calculation grid with a smooth growth of the amount. If j indicates the transverse grid index with the maximal value N , a parameter β to describe the gradient of the index wall and a strength parameter γ , this setup can be modeled by the formula

$$\Delta(n^2) = \frac{n_o \gamma}{ik_o} \left[1 - \cos^\beta \left(\frac{\pi(N+1-j)}{N} \right) \right]. \quad (27-121)$$

In reality, the exact condition of zero reflection can only be fulfilled for one value of the incidence angle. A further drawback of absorbing boundary conditions is that the energy is not constant during the propagation of the field.

Figure 27-39 shows the profile of the real and the imaginary part of the refractive index in the case of an absorbing boundary.

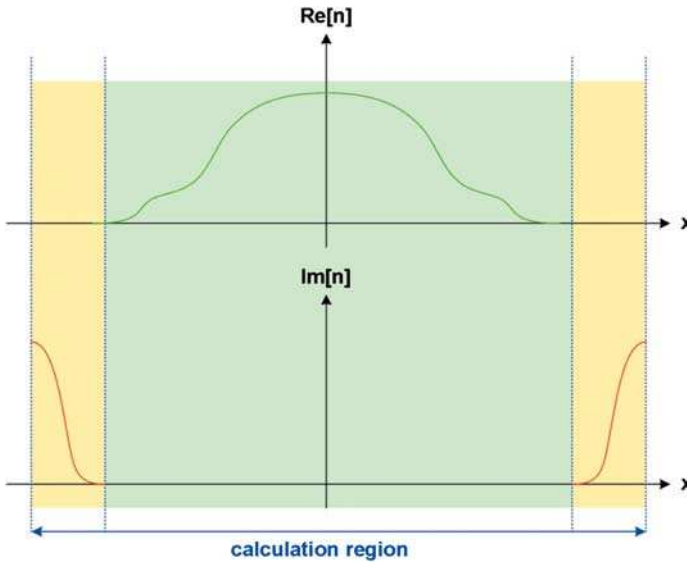


Figure 27-39: Real and imaginary part of the profile of the refractive index for an absorbing boundary.

2. Transparent Boundary

Another way to be sure that the propagation in the interior range of the calculation grid is not disturbed by boundary effects, is to construct transparent boundaries. If the field is analyzed regarding the local propagation direction, the boundary is constructed to be transparent for an incident wave in the corresponding direction.

27.9.2

Implicit Paraxial Wave Equation in Two Dimensions

If the wave equation of Helmholtz is approximated paraxially, the term with the z-derivative of the second order vanishes and one gets the following form for the field strength E (E is here one scalar field component)

$$2ik_0 n_0 \frac{\partial E}{\partial z} = \frac{\partial^2 E}{\partial x^2} + k_0^2 \cdot [n^2(x, z) - n_0^2] E . \quad (27-122)$$

If the derivatives are discretized and approximated by finite differences, the following equation results for the implicit method

$$-aE_{j-1,m+1} + bE_{j,m+1} - aE_{j+1,m+1} = aE_{j-1,m} + cE_{j,m} + aE_{j+1,m} \quad (27-123)$$

with the following coefficients

$$a = \frac{\Delta z}{2 \Delta x^2} , \quad (27-124)$$

$$b = \frac{\Delta z}{\Delta x^2} - \frac{\Delta z}{2} (n_{j,m+1}^2 - n_0^2) + 2ik_0 n_0 , \quad (27-125)$$

$$c = -\frac{\Delta z}{\Delta x^2} + \frac{\Delta z}{2} (n_{j,m}^2 - n_0^2) + 2ik_0 n_0 . \quad (27-126)$$

27.9.3

Paraxial Wave Equation in Cylindrical Coordinates

In this special case, setups with rotational symmetry can be calculated. This has the advantage of providing a fast calculation, but the problems with the singularity at the axis $r=0$ make this attempt less attractive. The paraxial approximated wave equation takes the form [27-40]

$$2ik_0 n_0 \frac{\partial E}{\partial z} = \frac{\partial^2 E}{\partial r^2} + \frac{1}{r} \frac{\partial E}{\partial r} + k_0^2 [n^2(x, z) - n_0^2] E . \quad (27-127)$$

With the transformation

$$\Phi = r E \quad (27-128)$$

we then get the new equation for the function Φ

$$2ik_0 n_0 \frac{\partial \Phi}{\partial z} = \frac{\partial^2 \Phi}{\partial r^2} - \frac{1}{r} \frac{\partial \Phi}{\partial r} + \frac{1}{r^2} + k_0^2 [n^2(x, z) - n_0^2] E \quad (27-129)$$

which always obeys the boundary condition

$$\Phi(r = 0) = 0. \quad (27-130)$$

If the coordinate along the propagating direction is scaled by

$$t = \frac{z}{2k} \quad (27-131)$$

and the abbreviation

$$\gamma = \frac{2 \Delta r^2}{\Delta z} \quad (27-132)$$

is used, we get the following equation for the implicit scheme

$$a_j \Phi_{j+1, n+1} + b_j \Phi_{j, n+1} + c_j \Phi_{j-1, n+1} = d_j \quad (27-133)$$

where the coefficients are defined by

$$a_j = 1 - \frac{1}{2j}, \quad (27-134)$$

$$c_j = 1 + \frac{1}{2j}, \quad (27-135)$$

$$b_j = -2 + i\gamma + \frac{1}{j^2} + \Delta r^2 (k_0^2 n^2 - k_0^2), \quad (27-136)$$

$$d_j = -a_j \Phi_{j+1, n} - (b_j - 2i\gamma) \Phi_{j, n} - c_j \Phi_{j-1, n}. \quad (27-137)$$

The calculation near the axis can be supported by a direct insertion of a Taylor expansion of the field E , which cannot be recalculated from the function Φ because of the vanishing factor $r = 0$ for this point. The expansion results in the relation

$$E_1 = E_2 - \frac{3}{7} E_3 + \frac{1}{14} E_4. \quad (27-138)$$

The figures 27-40 to 27-42 show some calculation examples for this special discretization method.

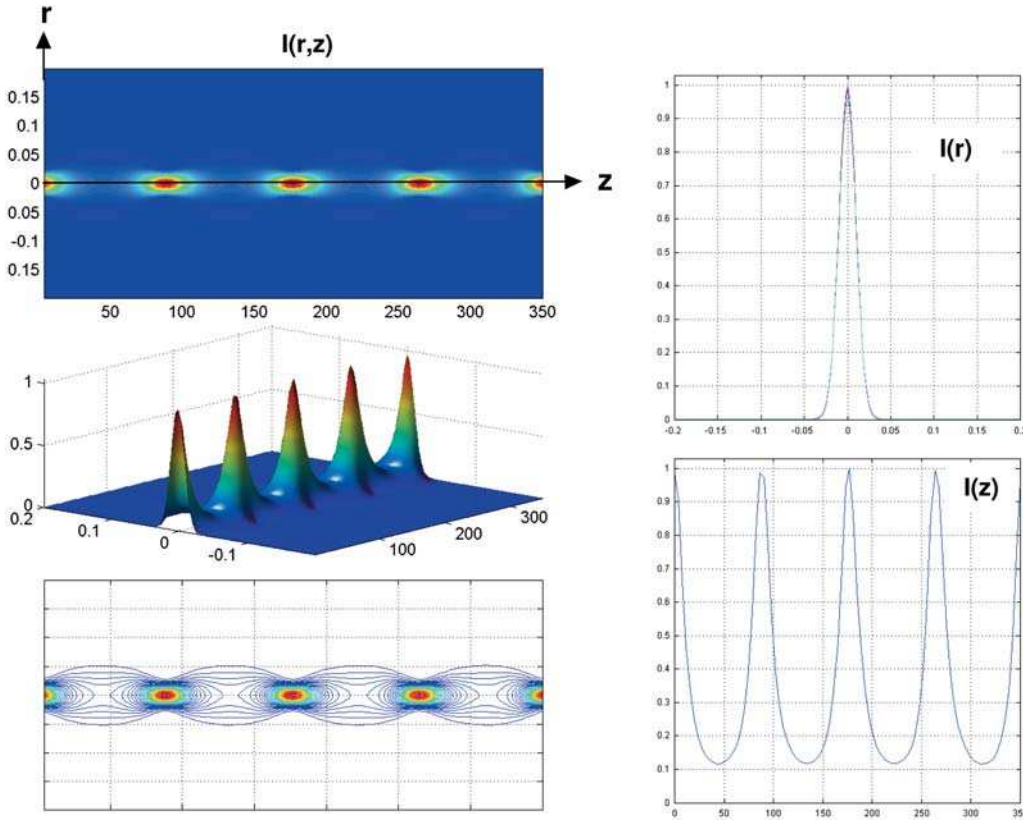


Figure 27-40: Propagation of a Gaussian beam in a gradient index medium with a parabolic distribution of the refractive index. The well known periodical refocussing after the pitch length is seen clearly. Left side: intensity profile as a function of the radial coordinate r and the axial coordinate z . Right side, top: transverse intensity profile; right side, bottom: axial intensity distribution.

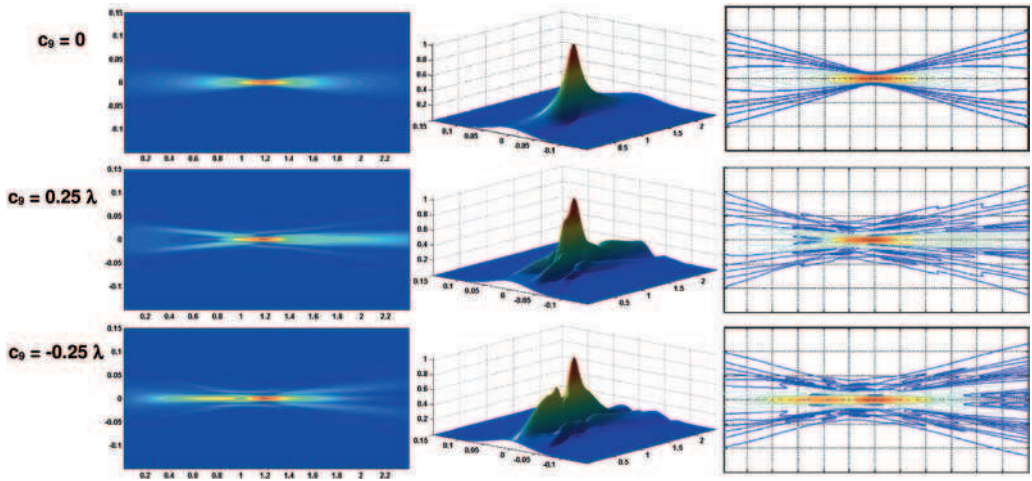


Figure 27-41: Focussing of a Gaussian beam in the ideal case (top row), with a spherical aberration of $\lambda/4$ (middle row) and with $-\lambda/4$ spherical aberration (bottom row). The asymmetrical intensity distribution around the image plane is indicated very clearly.

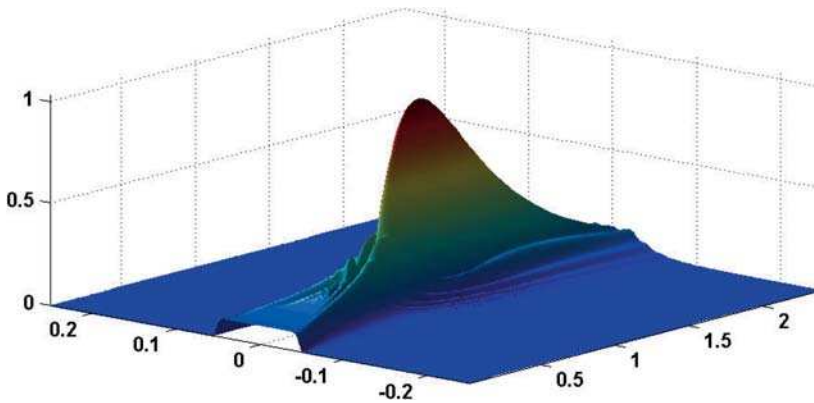


Figure 27-42: Propagation of a beam with a super Gaussian intensity profile with an exponent of $m = 20$. The translation invariance of the beam profile, the hot spots and the Fresnel diffraction ripple can be seen, occurring during the propagation.

27.9.4

ADI-formulation of the Paraxial Wave Equation in Three Dimensions

If the full geometry with two transverse dimensions is taken into account, the discretization of the paraxial wave equation becomes a little more complicated. If the operator with the transverse derivatives is written in the form

$$\hat{H} = \frac{1}{2k} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right) = \hat{H}_x + \hat{H}_y, \quad (27-139)$$

the derivative of the first order along the propagation direction delivers the representation

$$\left(1 - \frac{\hat{H}_x}{2} - \frac{\hat{H}_y}{2} \right) E_{i,j,n+1} = \left(1 + \frac{\hat{H}_x}{2} + \frac{\hat{H}_y}{2} \right) E_{i,j,n} \quad (27-140)$$

where the field strength E is discretized in the three coordinates' direction with the indices i, j and n . In the approximation of decoupled operators in the two coordinate directions, the expressions in the brackets can be factorized and we get the equation

$$\left(1 - \frac{\hat{H}_x}{2} \right) \left(1 - \frac{\hat{H}_y}{2} \right) E_{i,j,n+1} = \left(1 + \frac{\hat{H}_x}{2} \right) \left(1 + \frac{\hat{H}_y}{2} \right) E_{i,j,n}. \quad (27-141)$$

According to this formula, the process of a z -step can be separated into two steps with alternating influence of the derivatives. This idea is called Peaceman–Rachford scheme, the equation can be split into two single equations with first-order operators

$$\left(1 - \frac{\hat{H}_x}{2} \right) E_{i,j,n+\frac{1}{2}} = \left(1 + \frac{\hat{H}_y}{2} \right) E_{i,j,n}, \quad (27-142)$$

$$\left(1 - \frac{\hat{H}_y}{2} \right) E_{i,j,n+1} = \left(1 + \frac{\hat{H}_x}{2} \right) E_{i,j,n+\frac{1}{2}}. \quad (27-143)$$

27.9.5

Split-step-beam Propagation Method

If the paraxial approximated wave equation

$$2ik \frac{\partial E}{\partial z} = \nabla_{\perp}^2 E + k^2 \left[\frac{n^2(x, y)}{n_0^2} - 1 \right] E \quad (27-144)$$

is written in the operator notation

$$\frac{\partial E}{\partial z} = \frac{1}{2ik} \nabla_{\perp}^2 E + \frac{k}{2i} \cdot \left[\frac{n^2(x, y)}{n_0^2} - 1 \right] \cdot E, \quad (27-145)$$

$$\frac{\partial E}{\partial z} = \hat{H}[E] + \hat{G}[E] \quad (27-146)$$

with the differential operator

$$\hat{H}[E] = \frac{1}{2ik} \nabla_{\perp}^2 E \quad (27-147)$$

and the non-linear operator G , which describes the inhomogeneities of the refractive index profile

$$\hat{G}[E] = \frac{k}{2i} \cdot \left[\frac{n^2(x, y)}{n_0^2} - 1 \right] \cdot E \quad (27-148)$$

the solution of the differential equation can be written in the symbolic form

$$E(z) = \int (\hat{H} + \hat{G}) E dz, \quad (27-149)$$

$$E(z + \Delta z) = e^{(\hat{H} + \hat{G}) \cdot \Delta z} E(z). \quad (27-150)$$

It should be noticed that the two operators are not commutative, so the sequence of the application of the exponential factors cannot be interchanged. The non-commutativity results from the spatial variation of the refractive index profile. In a quadratic approximation, the above expression can be written in the separated form

$$E(z + \Delta z) = e^{\hat{H} \cdot \Delta z} \cdot e^{\hat{G} \cdot \Delta z} E(z). \quad (27-151)$$

The solution of the propagation problem therefore can be obtained by propagating the field in the Fourier space by an expansion into plane waves and then correcting the phase of the field caused by the changed refractive index. The single steps of the algorithm of the so called split-step-beam propagation are therefore as follows:

1. Fourier transformation of the field

$$E(k_x, z) = \hat{F}[E(z)]. \quad (27-152)$$

2. Propagation of the plane waves in the spatial frequency domain

$$E(k_x, z + \Delta z) = E(k_x, z) \cdot e^{\frac{ik_x^2 \cdot \Delta z}{2k_0}}. \quad (27-153)$$

3. Back transform of the spectrum in the spatial domain

$$E'(x, z + \Delta z) = \hat{F}^{-1}[E(k_x, z + \Delta z)]. \quad (27-154)$$

4. Correcting the field according to the change in the refractive index

$$E(x, z + \Delta z) = E'(x, z + \Delta z) \cdot e^{\hat{G} \cdot \Delta z}. \quad (27-155)$$

The method described above can be performed in a more accurate way by splitting the propagation step into two steps of the half-step width before and after the index correction.

The figures 27-43 and 27-44 show an application of the classical beam propagation algorithm for the scattering of a Gaussian beam by a centered and an off-axis bubble of the refractive index respectively.

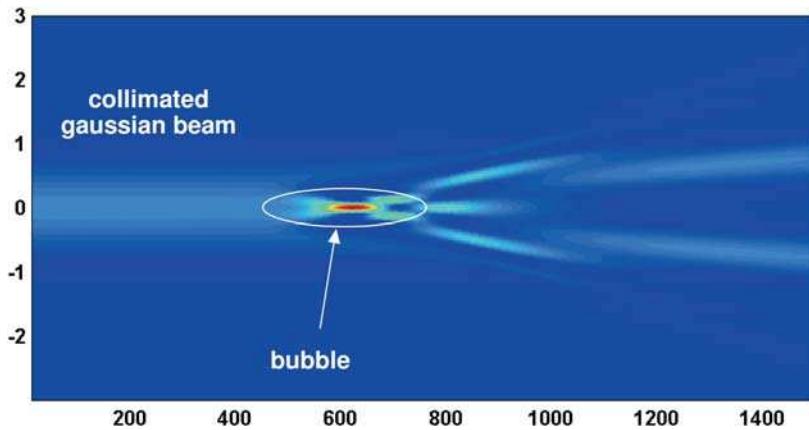


Figure 27-43: Propagation of a collimated Gaussian beam through a centered disturbance of the refractive index by a bubble.

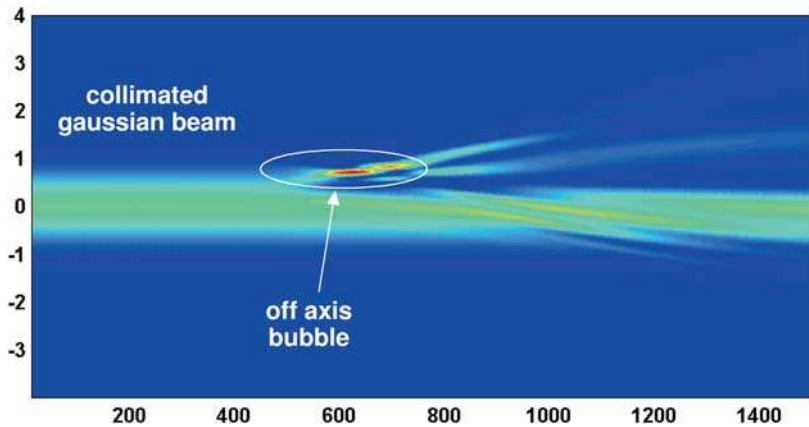


Figure 27-44: Propagation of a collimated Gaussian beam through an off-axis disturbance of the refractive index by a bubble.

27.10

Rigorous Diffraction in Optical Imaging

The interpretation of microstructure images in the nanometer region cannot be done without considering near-field effects resulting from an interaction of the electromagnetic wave with the structure topography. If not taken into account, these effects spoil the optical micro metrology at wavelength sized structures.

In principle, it should be possible to localize an edge with nm precision, provided that the next edge is far enough away to neglect any proximity effects, i.e., intensity variations due to the overlap of images. However, considerably higher deviations occur in practical linewidth measurements, i.e., the attainable accuracy is reduced. The reasons are due to the edge localization criteria: the basis is usually a geometrical-optics image interpretation according to which the reflected/transmitted field is computed by geometrical optics (“optical footprint”).

But the interaction of light with a structure is more complicated:

- Electromagnetic boundary conditions yield polarization dependencies.
- Lateral scattered fields and possible excited surface waves yield multiple scattering between neighboring structure elements.
- Structures with high aspect ratios act as waveguides.
- For complex structures, phase shifts due to material changes and multiple-beam interferences occur.

Optical imaging gains in resolution if the wavelength is reduced. The reason for this is the wavelength dependence of the optical resolution limit which is, in coordinates, normalized to the wavelength

$$\frac{\Delta x}{\lambda} = \frac{\kappa}{NA}. \quad (27-156)$$

In fact, the scalar diffraction scales linearly with the wavelength. With $\zeta = x/\lambda$, etc., the scalar wave equation becomes

$$\Delta U + \left(\frac{2\pi}{\lambda}\right) U = 0 \quad \mapsto \quad \left(\frac{\partial^2}{\partial \xi^2} + \frac{\partial^2}{\partial \eta^2} + \frac{\partial^2}{\partial \zeta^2}\right) U(\xi, \eta, \zeta) + (2\pi)^2 U(\xi, \eta, \zeta) = 0. \quad (27-157)$$

27.10.1

Dielectrics and Metals

Optical imaging is performed on real materials with electromagnetic waves. The image of a structure depends on the material composition of the object: in particular whether it is dielectric or metallic. The difference between dielectrics and metals is somewhat arbitrary. This is particular true for optical frequencies. Drude’s oscillator model provides for the frequency-dependent dielectric function approximately

$$\frac{\varepsilon - 1}{\varepsilon + 2} = \frac{Ne^2}{3\varepsilon_0 m_e} \sum_j \frac{f_j}{\omega_j^2 - \omega^2 - i\gamma_j \omega} \quad (27-158)$$

with the molecular density N , the elementary charge e , the mass m_e of the electron, the oscillator strength f_j of resonant frequency ω_j and the attenuation constant γ_j . If the first resonance frequency is zero, the material contains free electrons and is a conductor. If it is not zero, the material is dielectric. The main electromagnetic properties of the materials can be classified in terms of the dielectric constant as follows:

Type	Dielectric constant	Refractive index $n = \varepsilon^{1/2}$	limit
Dielectric	$\text{real}(\varepsilon_r) > 1, \text{imag}(\varepsilon_r) \rightarrow 0$	$\text{real}(n) \gg \text{imag}(n)$	$\text{imag}(n) = 0$
Metal	$\text{real}(\varepsilon_r) \rightarrow 0, \text{imag}(\varepsilon_r) \gg 1$	$\text{real}(n) \approx \text{imag}(n) \gg 1$	$\text{real}(n) = \text{imag}(n) \rightarrow \infty$
Surface plasmons ($\omega > \omega_p$)	$\text{real}(\varepsilon_r) < 0, \text{imag}(\varepsilon_r) \rightarrow 0$	$\text{real}(n) \rightarrow 0, \text{imag}(n) \gg 1$	

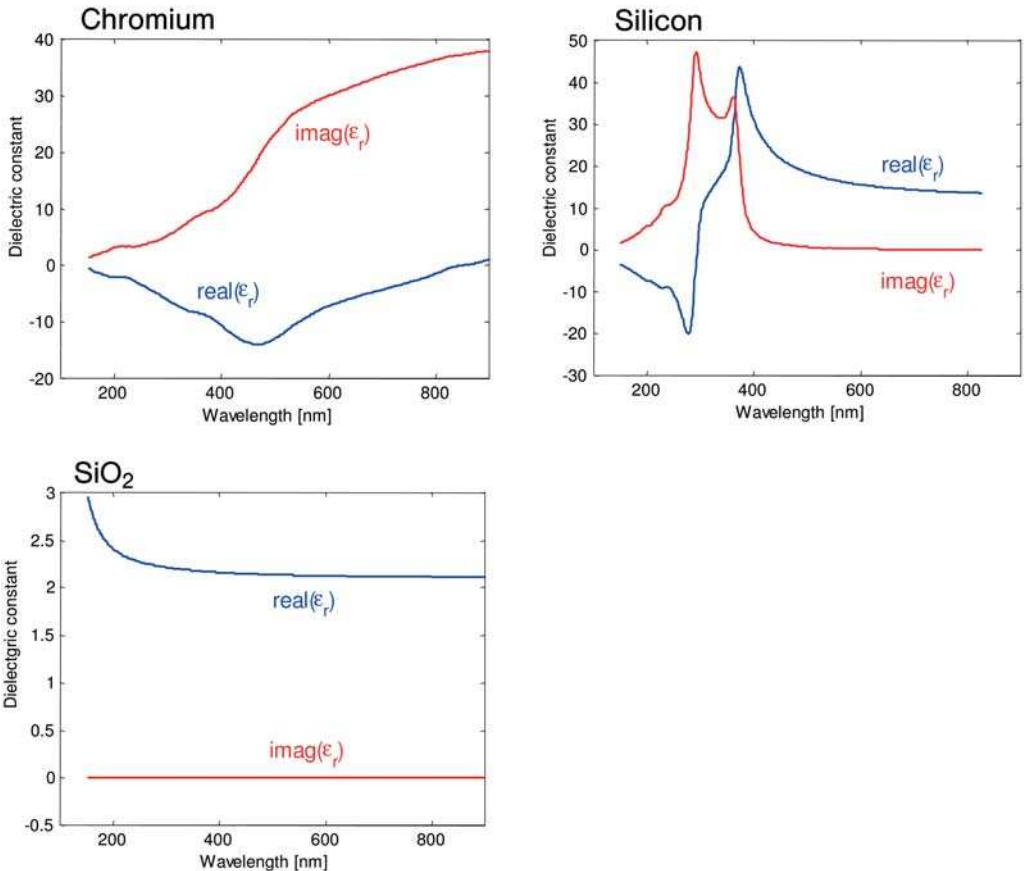


Figure 27-45: Dielectric function of chromium, silicon and SiO₂ [27-41].

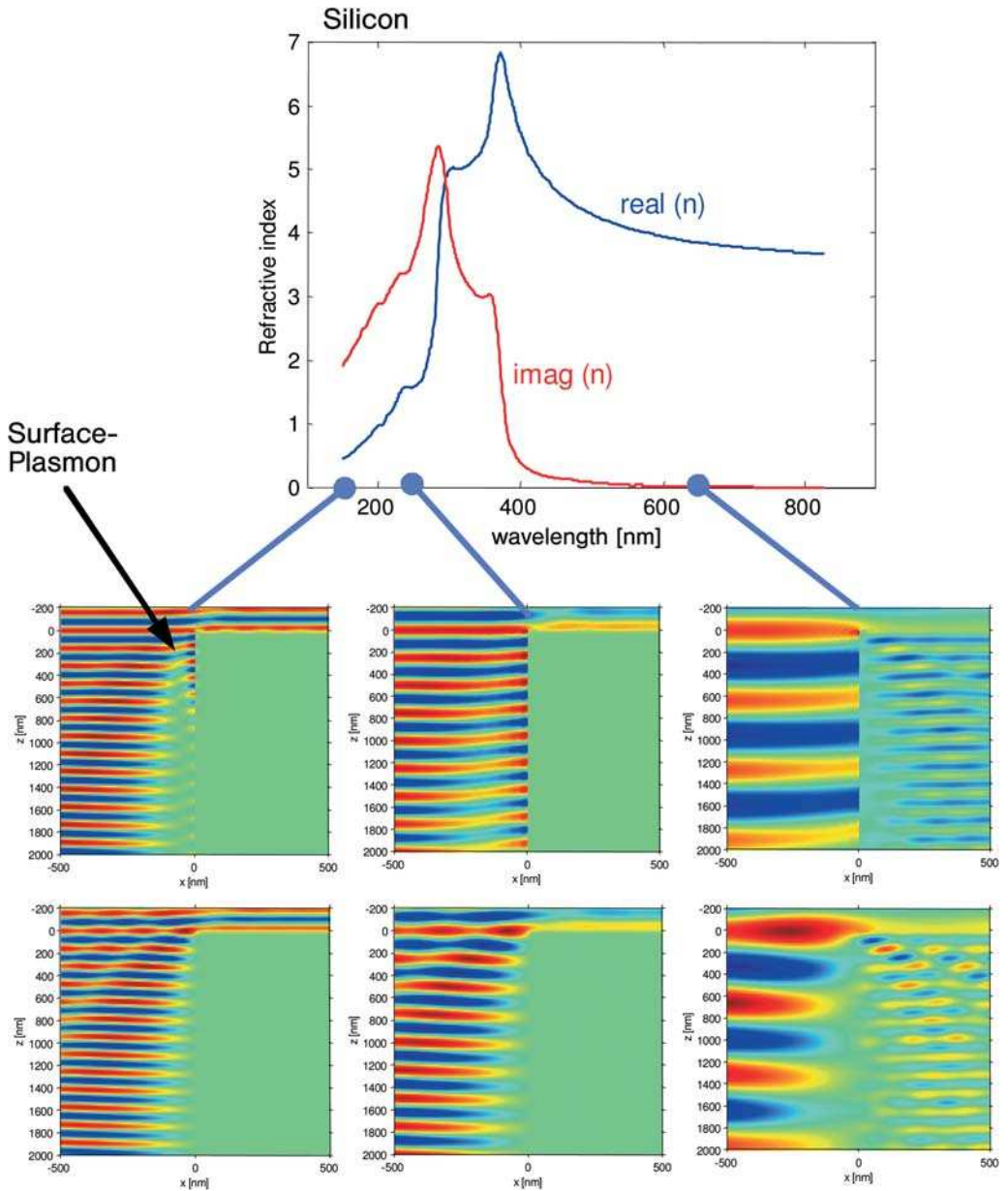


Figure 27-46: Top: Dielectric function of silicon; Middle: Corresponding edge near-fields for TM polarization; Bottom: Edge near-fields for TE polarization (incident plane wave from top).

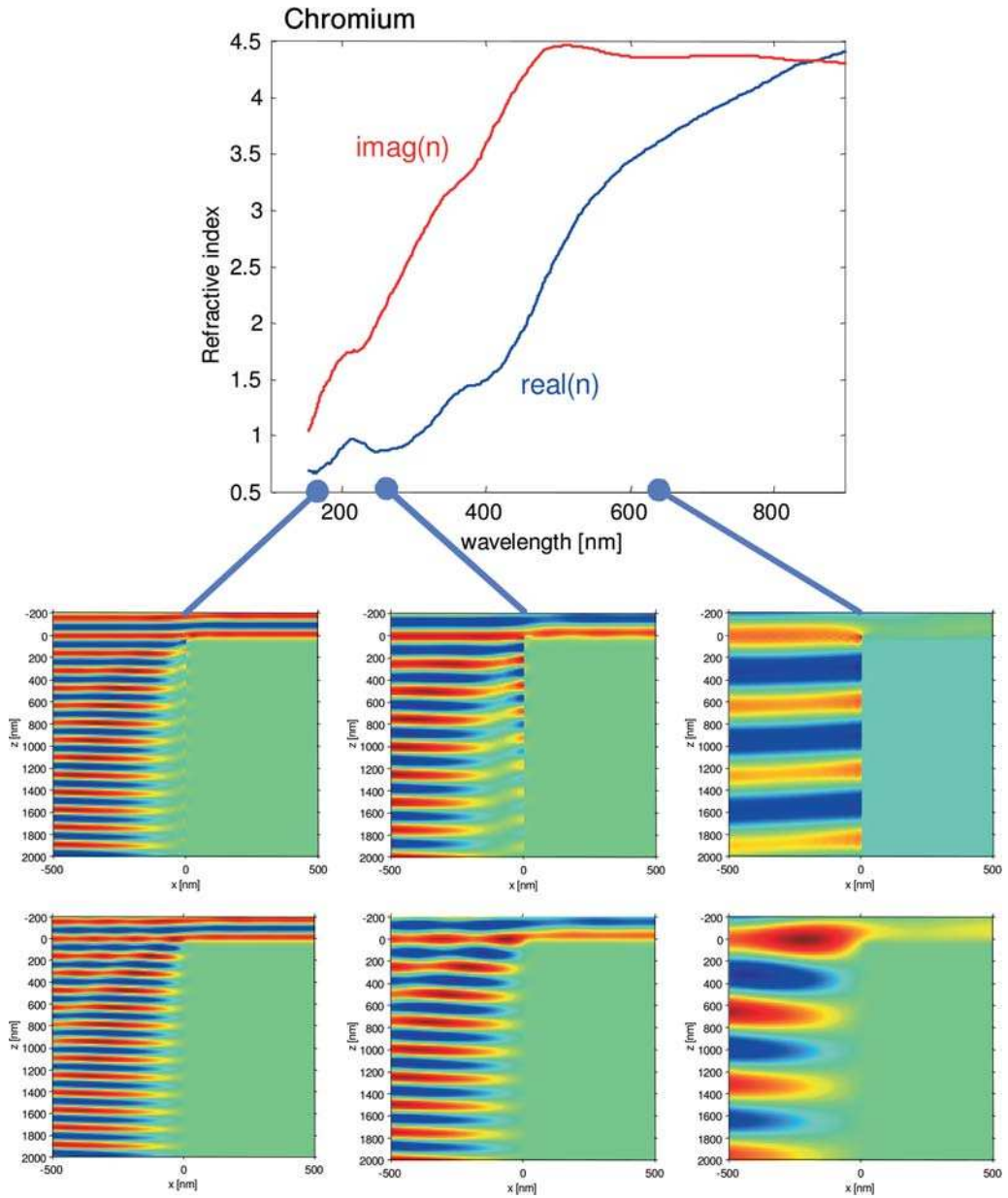


Figure 27-47: Top: Dielectric function of chromium; Middle: Corresponding edge near- fields for TM polarization; Bottom: Edge near-fields for TE polarization (incident plane wave from top).

The different kinds of materials yield different fields at the boundaries. The electromagnetic field is transmitted by the dielectric but absorbed by the metal. In the third case, surface plasmons may be excited, yielding long-range interactions.

The material properties change significantly near the resonance regions of the dielectric function. This effect is pronounced for chromium and silicon when traversing from the visible to the DUV region. (Data from [27-41]).

An elemental entity of the wave-structure interaction is a plane wave incident on an isolated edge. Using the FMM the corresponding near-fields were simulated for wavelengths of 157 nm, 248 nm and 633 nm. In order to avoid reflections from the bottom, a gradual transition from the refractive index of the material to one was implemented. The two images below show the near-fields for silicon and chromium. For easy comparison the dispersion diagram is shown in the same figure for both.

With decreasing wavelength, silicon traverses two resonances (cf. 27-45) and becomes metallic in this region. For TE-polarization an indication for the excitation of surface plasmons is visible for $\lambda=157$ nm and, with larger wavelength at $\lambda=248$ nm. For $\lambda=633$ nm silicon is mainly dielectric with a significant field inside the material near the edge. The theoretical surface plasmon wavelengths λ_p and damping constants α_p are shown in the table.

Material	Wavelength	λ_p/λ_0	α_p
Silicon/Air	157 nm	0.90	0.07
	248 nm	0.98	0.03
Chromium/Air	157 nm	0.99	0.32
	248 nm	0.93	0.09
	633 nm	0.99	0.01

27.11

Simulation of Polarized Imaging by use of Rigorous Diffraction

The image of a structure for a prescribed illumination is computed by a two-step process. First the spectrum of plane waves necessary to synthesize the desired illumination spectrum is determined. For each spectral component the diffracted orders of the structure are computed and stored.

A detailed description of polarized imaging will be given in chapter 28. Here we give only a brief account of a simple method for simulating polarized reflection microscopy as it is used for microstructure metrology [27-35]. Consider a single incident plane wave from the illuminating spectrum which is diffracted by the structure. To compute the corresponding electric field distribution in the image plane, all diffraction orders within the numerical aperture NA are summed

$$\vec{E}(\lambda, \vec{a}, x', y', z) =$$

$$Q(\lambda, \vec{p}'') \sum_{\substack{\nu, \mu \\ k_{\nu\mu}^{\perp} \leq N A k_0}} \psi(\vec{p}'_{\nu\mu}) \mathbf{J}(\vec{p}_{\nu\mu}) \vec{E}_{\nu\mu}^{\perp}(\vec{k}_{\nu\mu}^{\perp}) d\sigma \exp\left\{i\left(\vec{k}'_{\nu\mu} \cdot \vec{r}' + \gamma z + w(\vec{p}'_{\nu\mu})\right)\right\}. \quad (27-159)$$

$Q(\lambda, \vec{a})$ is the complex amplitude of the incident wave in the exit pupil of the illumination system. $w(\vec{p}'_{\nu\mu})$ are the wavefront aberrations. $\mathbf{J}(\vec{p}_{\nu\mu})$ is a Jones matrix for consideration of polarization aberrations and polarization-affecting components in the imaging pupil, and z is the defocus of the object. $E_{\nu\mu}^{\perp}$ is obtained by a transformation of the diffracted wave into the pupil P taking energy conservation in the object plane into account

$$\vec{E}_{nm} = \frac{1}{|\vec{k}_{\perp}|} \begin{pmatrix} -k_y & k_x \\ \pm Z_0^{-1} k_x & \pm Z_0^{-1} k_y \end{pmatrix} \begin{pmatrix} \text{Re}^{1/2}\{\gamma/\gamma_i\} E_{\perp} \\ \text{Re}^{1/2}\{n_i \gamma/n_{12}^2 \gamma_i\} H_{\perp} \end{pmatrix}. \quad (27-160)$$

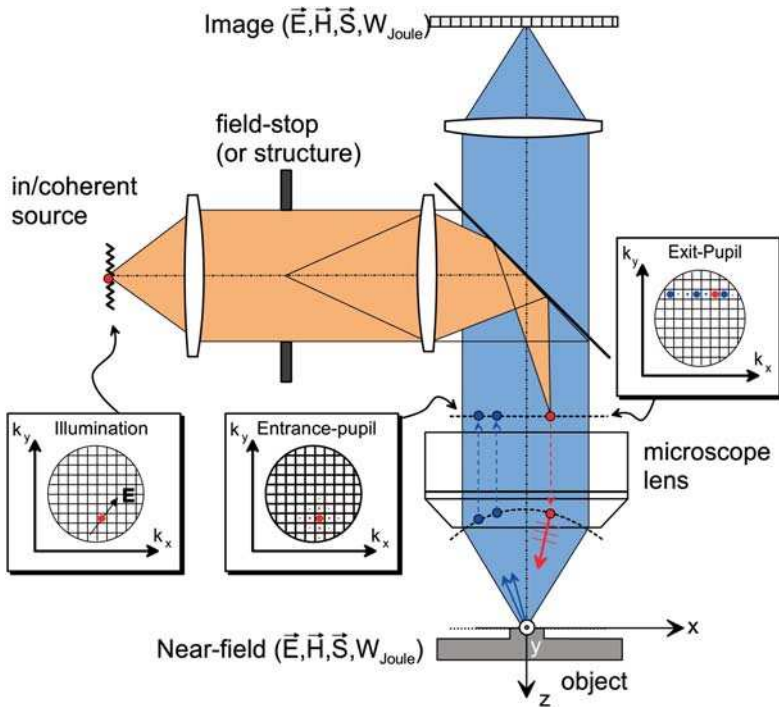


Figure 27-48: Simulating reflection microscopy.

Matrix ψ (see chapter 28) performs the transformation from the exit pupil of the imaging system into image space. The propagation vector of the $\nu\mu^{\text{th}}$ diffracted order behind the exit pupil of the imaging system with magnification m is

$$\vec{k}'_{\nu\mu} = m^{-1} \vec{k}_{\nu\mu}^{\perp}. \quad (27-161)$$

The computation of the image due to complete illumination follows from an integral, approximated as a sum, over the wavelength and illumination pupil where the computed quantity depends on the coherence of the effective source in the illumination entrance pupil. For a coherent, and therefore monochromatic, source the complex electric field in the image plane is computed by

$$\vec{E}_\lambda(x', y', z) = \sum_{\nu\mu} \vec{E}(\vec{p}_{\nu\mu}, \lambda, x', y', z) d\sigma \quad (27-162)$$

where $d\sigma$ denotes the size of a patch in the pupil plane. For an incoherent and possibly polychromatic source, the image polarization matrix is computed as

$$\mathbf{P}(x', y', z) = \sum_j \sum_{\nu\mu} \mathbf{P}(\vec{p}_{\nu\mu}, \lambda_j, x', y') d\sigma \Delta\lambda \quad (27-163)$$

where $\Delta\lambda$ is the sampling distance of the spectrum. \mathbf{P} is the polarization matrix (coherency matrix), i.e., the time mean value of the dyadic product of the electric field with its conjugate

$$\mathbf{P} = \begin{pmatrix} \langle E_x E_x^* \rangle & \langle E_x E_y^* \rangle & \langle E_x E_z^* \rangle \\ \langle E_y E_x^* \rangle & \langle E_y E_y^* \rangle & \langle E_y E_z^* \rangle \\ \langle E_z E_x^* \rangle & \langle E_z E_y^* \rangle & \langle E_z E_z^* \rangle \end{pmatrix}. \quad (27-164)$$

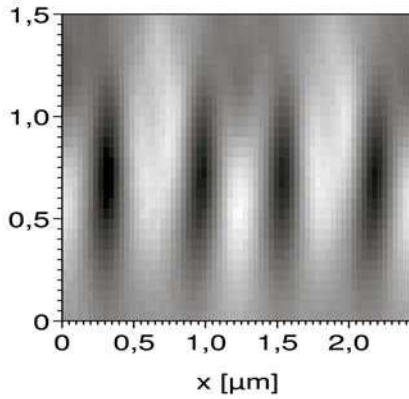
It is a straightforward matter to show the linearity of the polarization matrices for incoherent fields. The intensity is the trace of \mathbf{P} . The non-diagonal elements are a measure of the correlation of the components. They correspond to quantities measured by polarization interferometry [27-34]

Example: Polarization-dependent image of a topographical grating in Si, computed with the FMM [27-35].

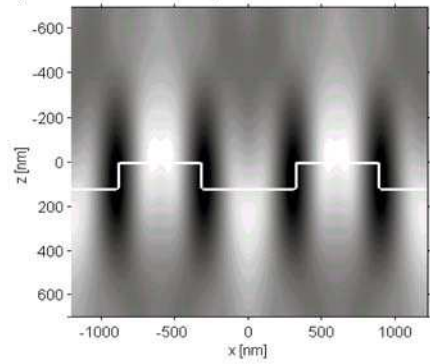
As an example, we consider the polarization effect in the image of a Si grating of grooves of 121 nm depth 630 nm width and 1200 nm pitch, i.e., a lateral aspect ratio of 1:1.9. The illumination wavelength is 440 nm, the illumination NA is 0.75, the imaging NA is 0.95, and the magnification is 250 (Leica DMR equipped with a 250/0.95 lens). The illumination is linearly polarized (TE: E parallel to grooves, TM: E perpendicular to grooves). The focus position was increased from -700 nm to $+700$ nm. For the computation the illumination pupil was divided into patches of $0.1 \times 0.1 NA^2$ and the refractive index of Si was taken from [27-41] as $n = 4.787 + i0.169$.

For this example, both linear polarizations provide clearly distinct image patterns. In TM-polarization (Figure 27-49) the intensity minimum is vertically oriented, while it is inclined in TE-polarization (Figure 27-50). In terms of accuracy for metrology, the first situation is certainly preferred. The vertical position of the intensity maxima is shifted between TE and TM-polarization, in both measurement and computation.

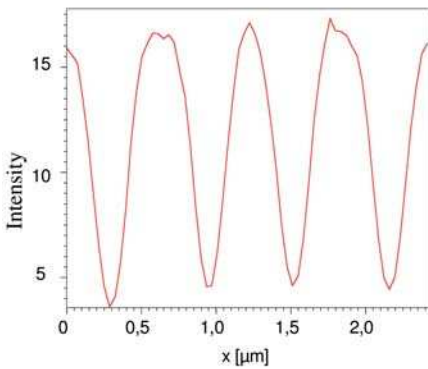
a) focus-series measurement



b) focus series computation



c) Intensity in focus, measurement



d) intensity in focus, computation

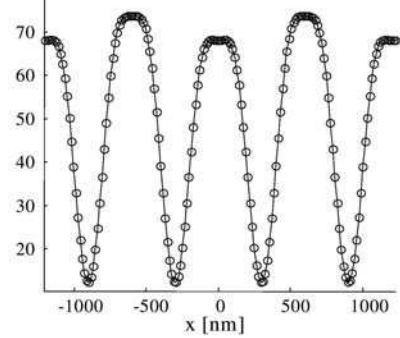


Figure 27-49: Intensity in the image of a Si grating of 1200 nm pitch and 630 nm grooves of 121 nm depth. Illumination: TM-polarization, $\lambda = 440$ nm, $NA = 0.7$. Imaging- $NA = 0.95$.

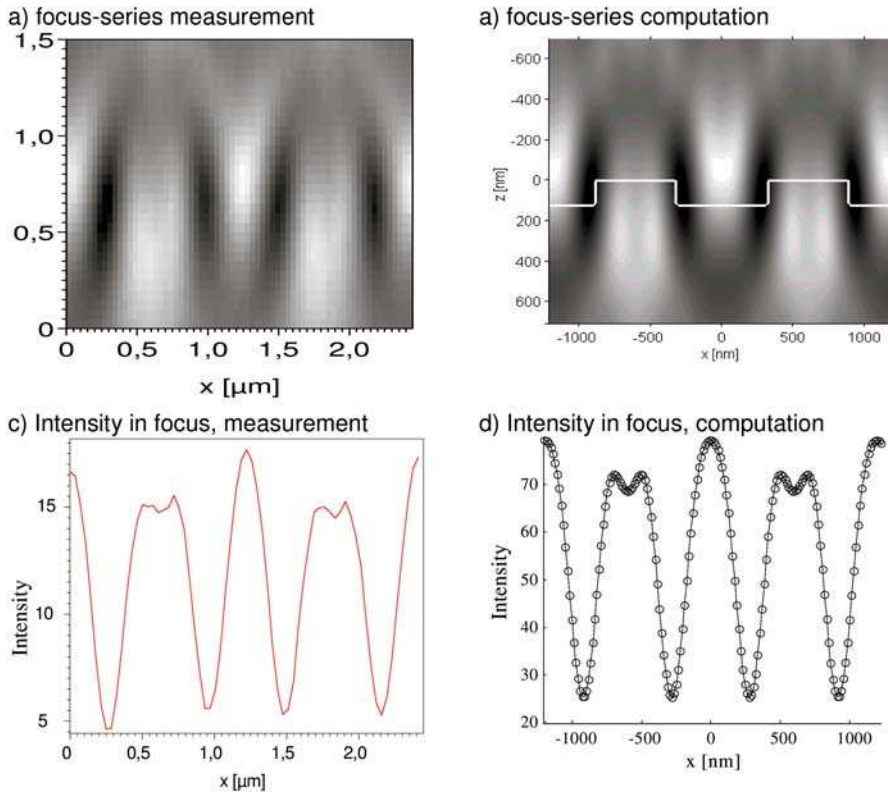


Figure 27-50: As Figure 27.49 but with TE-polarization.

27.12

Literature

- 27-1 E. Wolf, Electromagnetic diffraction in optical systems: I. An integral representation of the image field, *Proc. R. Soc. London Ser. A* **253**, 349–357 (1959).
- 27-2 B. Richards, E. Wolf, Electromagnetic diffraction in optical systems: II. Structure of the image field in an aplanatic system, *Proc. R. Soc. London Ser. A* **253**, 358–379 (1959).
- 27-3 C. J. R. Sheppard, A. Choudhury, “Annular pupils, radial polarization, and superresolution”, *Applied Optics* **43**, 4322–4327 (2004).
- 27-4 R. Dorn, S. Quabis, G. Leuchs, Sharper focus for a radially polarized light beam, *Phys. Rev. Lett.* **91** (2003) 233901/1–4.
- 27-5 J. A. Stratton, L. J. Chu, Diffraction theory of electromagnetic waves, *Phys. Rev.* **56**, 99–107 (1939).
- 27-6 S. Silver (ed.), *Microwave Antenna Theory and Design* (IEE Electromagnetic Wave Series 19, London, Peregrinus).
- 27-7 A. Sommerfeld, *Vorlesungen über Theoretische Physik*, Vol. 3 “Optik”, 3rd ed. (Leipzig, 1960), p. 173, Chap. 34E.
- 27-8 H.A. Bethe, Theory of Diffraction by Small Holes, *Phys. Rev.* **66**, 163–182, (1944).
- 27-9 R.C. McPhedran, G.H. Derrick, L.C. Botten, Theory of Crossed Gratings, in *Electromagnetic Theory of Gratings* R. Petit (Hrsg.), Topics in Current Physics Bd. 22 (Springer, Berlin 1980).
- 27-10 M. Born, E. Wolf, *Principles of Optics*, Pergamon Press, 6th ed., 1980.
- 27-11 Lord Rayleigh, *Phil. Mag.* **36**, 365 (1918).

- 27-12** J.R. Wait, Scattering of a plane wave from a circular dielectric cylinder at oblique incidence, *Can. J. Phys.* **33**, 189 (1955).
- 27-13** G. Mie, Beiträge zur Optik trüber Medien, speziell kolloidaler Metallösungen, *Ann. Phys.* **25**, 377 (1908).
- 27-14** C. Yeh, Scattering of Obliquely Incident Light Waves by Elliptical Fibers, *J. Opt. Soc. Am.* **54**, 1227 (1964).
- 27-15** H. C. van de Hulst, *Light Scattering by Small Particles* (Dover, New York, 1981).
- 27-16** C.F. Bohren, D.R. Huffman, *Absorption and Scattering of Light by Small Particles*, (John Wiley, New York, 1983).
- 27-17** M. Abramowitz, I.A. Stegun *Handbook of Mathematical Functions* (Dover, New York, 1965).
- 27-18** E.M. Purcell, C.R. Pennypacker, Scattering and Absorption of Light by Nonspherical Dielectric Grains, *The Astrophysical Journal* **186**, 705–714 (1973).
- 27-19** M. Born, *Optik*, 3rd ed., (Springer, Heidelberg, 1972).
- 27-20** J.H. Richmond, Scattering by a dielectric cylinder of arbitrary cross section shape, *IEEE Transact. Ant. Prop.* **13** 334 (1965), and TE-wave scattering of a cylinder of arbitrary cross section shape, *IEEE Transact. Ant. Prop.* **14**, 460 (1966).
- 27-21** C.C. Su, Electromagnetic Scattering by a Dielectric Body with Arbitrary Inhomogeneity and Anisotropy, *IEEE Transact. Ant. Prop.* **37**, 384–389 (1989).
- 27-22** M.A. Taubenblatt, Light scattering from cylindrical structures on surfaces, *Opt. Lett.* **15**, 255 (1990).
- 27-23** A. Sentenac, J.-J. Greffet, Scattering by deep inhomogeneous gratings, *J. Opt. Soc. Am. A* **9**, 996 (1992).
- 27-24** M. Totzeck, H.J. Tiziani, Interference-microscopy of sub- λ structures: A rigorous computation method and measurements, *Optics Communications* **136**, 61–74 (1997).
- 27-25** C.C. Su, A simple evaluation of some principal value integrals for Dyadic Greens function using symmetry property, *IEEE Transact. Ant. Prop.* **35**, 1306 (1987).
- 27-26** C.C. Su, Calculation of electromagnetic scattering from a dielectric cylinder using the Conjugate gradient method and FFT, *IEEE Transact. Ant. Prop.* **35**, 1418 (1987).
- 27-27** W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, *Numerical Recipes in C*, 2nd ed. (Cambridge 1992).
- 27-28** T.K. Sarkar, K.R. Siarkiewicz, R.F. Stratton, Survey of numerical methods for solution of Large Systems of linear equations for electromagnetic field problems, *IEEE Transact. Ant. Prop.* **29**, 847 (1981).
- 27-29** M.G. Moharam, E.B. Grann, D.A. Pommet, T.K. Gaylord, Formulation for stable and efficient implementation of the rigorous coupled-wave analysis of binary gratings, *J. Opt. Soc. A* **12**, 1068–1076 (1995).
- 27-30** M.G. Moharam, E.B. Grann, D.A. Pommet, T.K. Gaylord, Stable implementation of the rigorous coupled-wave analysis for surface-relief gratings: enhanced transmittance matrix approach, *J. Opt. Soc. A* **12**, 1077–1086 (1995).
- 27-31** P. Lalanne, G.M. Morris, Highly improved convergence of the coupled-wave method for TM-polarization, *J. Opt. Soc. A* **13**, 779–784 (1996).
- 27-32** Lifeng Li, New formulation of the Fourier modal method for crossed surface relief gratings, *J. Opt. Soc. Am. A* **14**, 2758–2767 (1997).
- 27-33** Lifeng Li, Use of Fourier series in the analysis of discontinuous periodic structures, *J. Opt. Soc. A* **13**, 1870–1876 (1996).
- 27-34** M. Totzeck, H. Jacobsen, H.J. Tiziani, Edge localization of sub-wavelength structures using polarization interferometry and extreme-value criteria, *Appl. Opt.* **39**, 6295–6305 (2000).
- 27-35** M. Totzeck, Numerical simulation of high-NA quantitative polarization microscopy and corresponding near-fields, *Optik* **112**, 399–406 (2001).
- 27-36** G. D. Smith, *Numerical Solution of Partial differential Equations: Finite Difference Methods* (Clarendon Press, Oxford, 1985).
- 27-37** J. C. Strikwerda, *Finite Difference Schemes and Partial Differential Equations* (Wadsworth & Brooks, Pacific Grove, 1989).
- 27-38** W. F. Ames, *Numerical Methods for Partial Differential Equations* (Academic Press, New York, 1992).
- 27-39** P. P. Banerjee, *Nonlinear Optics* (Marcel Dekker, New York, 2004).
- 27-40** S. T. Hendow and S. A. Shakir, Recursive numerical solution for nonlinear wave propagation in fibers and cylindrically symmetric systems, *Appl. Opt.* **25**, 1759 (1986).
- 27-41** E.D. Palik, *Handbook of Optical Constants of Solids* (Academic Press, Orlando, 1985).
- 27-42** See the internet page of Thomas Wriedt for a list of available programs and reports: <http://www.iwt-bremen.de/vt/laser/wriedt>.

28

Polarization and Optical Imaging

- 28.1 Introduction 590
- 28.2 The Image-forming Field 590
- 28.3 Interference of Electromagnetic Waves 592
 - 28.3.1 Two-beam Vector Interference 592
 - 28.3.2 Contrast for High-NA, s- and p-polarization 593
 - 28.3.3 Influence of Recording Medium 594
 - 28.3.4 Vector Effect in Optical Microscopy 595
 - 28.3.5 Vector Effect in Optical Lithography 595
- 28.4 Polarized Ray Trace 596
 - 28.4.1 Definition of Ray, Beam and Path 597
 - 28.4.2 Ray-splitting at Anisotropic Elements 597
 - 28.4.3 Refraction and Reflection at Birefringent Interfaces 598
 - 28.4.4 The Single-path Approximation 599
- 28.5 Optical Systems with Polarization Effects 604
- 28.6 Polarized Imaging Model 605
 - 28.6.1 Scalar Image 606
 - 28.6.2 Vector Image for Completely Polarized Illumination 607
 - 28.6.3 Vector Image for Partially Polarized Illumination 609
- 28.7 Vector Point-spread Function 610
 - 28.7.1 VPSF for Complete Polarization 610
 - 28.7.2 VPSF for Unpolarized Illumination 611
- 28.8 Polarized Optical Transfer Function 612
 - 28.8.1 Polarized Illumination 612
 - 28.8.2 Unpolarized Illumination 612
- 28.9 Jones Matrix Pupil 612
 - 28.9.1 Definition for Completely Polarized Illumination 613
 - 28.9.2 Separation of a Scalar Factor 614
 - 28.9.3 Decomposition into Retardance and Diattenuation 615
 - 28.9.4 Example 616
- 28.10 Jones Matrix Pupils in the Polarization Matrix Calculus 617
- 28.11 Jones-matrix-based System Optimization 619
- 28.12 Aberrations of the Transmitted Wavefront 620
- 28.13 Jones–Zernike Wavefront Aberrations 621

28.13.1	Principle of the Modal Characterization of a Jones Pupil	621
28.13.2	Jones–Zernike Expansion	621
28.13.3	Properties of the Jones–Zernike Polynomials	623
28.14	Literature	625

28.1

Introduction

Any optical imaging is based on the interference of light. Because light is an electromagnetic wave the interference contrast is polarization-dependent. So the polarization of the waves has a direct impact on the image quality. While identical polarization states may interfere perfectly, orthogonal states do not interfere at all. This is of particular importance for high-numerical aperture systems – in particular lithography lenses where the image is formed with high NA – and for optical systems that make use of polarization, such as, for instance, projection devices using liquid crystal spatial-light modulators or polarization microscopy. Furthermore, polarization can be used to encode information into electromagnetic waves and to retrieve the information from them. Liquid crystal displays and digital projectors are well known examples.

Chapter 26 considered the mathematical description of polarization states and their change due to various optical components. In the current chapter we will concentrate on the system aspect: How do we describe image formation for polarization? How do we describe and optimize optical systems that have an influence on the state of polarization?

Before we can proceed to a mathematical description of image formation for polarized light, we must clarify which quantity of the electromagnetic field is actually measured by optical detectors like photographic emulsions, photoresists or CCD chips. Methods for a polarized ray trace are presented thereafter. The subjects covered in the following sections are optical systems where polarization effects are of importance. We have chosen a microscope lens and a lens for optical lithography as examples. After this we discuss the interference of electromagnetic waves in contrast to scalar waves. In the subsequent chapter we define and discuss the Jones pupil and present methods for its visualization. Finally, we proceed to the vector point-spread function, the polarization transfer function and the optimization of optical systems with respect to polarization.

28.2

The Image-forming Field

Generally speaking, optical imaging is just the superposition of electromagnetic waves. However, any image has to be recorded, and before we are able to compute polarized images, we must address the question of which quantity in the field produces the final image, i.e., to which component is the recording medium susceptible? Is it the electric field component, or the magnetic field component or perhaps the magnitude of the Poynting vector? The first answer to this question can be derived from the fact that, for the detection of light, electrons have to be excited. This is true for photoemulsion, the eye, fluorescence and charge coupled devices (CCDs). Being charged particles, the electrons are expected to be excited much more efficiently by the electric field component than by the magnetic field.

The experimental proof that it is the electrical component which is detected at optical frequencies, was given in 1890 by Otto Wiener [28-1]. He used a standing wave in front of a silver mirror. The highly conducting mirror enforces a node of the electric field at the surface and subsequent nodes at half a wavelength distance from each other. As can be easily deduced from Maxwell's equations, the magnetic field, in contrast, has a maximum at the surface. The first node lies a quarter of a wavelength above the surface. The electric and magnetic energy density of the standing wave is

$$\begin{aligned} \rho_E &= \frac{1}{2} |E_0 e^{ik_0 z} - E_0 e^{-ik_0 z}|^2 = |E_0|^2 \sin(k_0 z), \\ \rho_H &= \frac{1}{2} |H_0 e^{ik_0 z} + H_0 e^{-ik_0 z}|^2 = |H_0|^2 \cos(k_0 z), \end{aligned} \tag{28-1}$$

where z denotes the axial coordinate (cf. figure 28-1) and $k_0 = 2\pi/\lambda$. In the standing wave Wiener introduced a glass plate with a very thin (here "very" means: smaller than the wavelength λ , in his experiment $\approx \lambda/30$) photographic emulsion composed of collodion added with silver chloride. The glass plate was laid on the mirror at an extremely shallow angle α . After development, the pellicle showed a regular pattern of dark stripes at distances $\lambda/2\sin\alpha$.

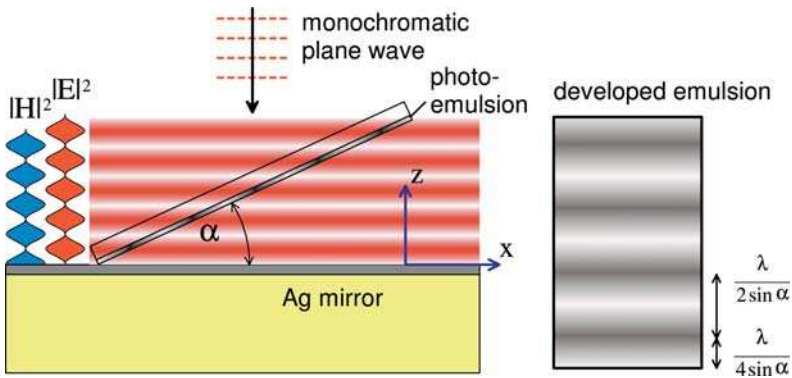


Figure 28-1: Wiener's experiment proving that photoemulsion is sensitive to the electric rather than to the magnetic field.

At the surface of the mirror, the pellicle was bright, i.e., no field was detected there. The first dark stripe appeared at a distance of $\lambda/4\sin\alpha$ above the silver surface. This is in accordance with the expectation that the electric field excites the silver chloride. Recasting the experiment with a fluorescent dye as detector yielded the same result [28-2].

Therefore the image is computed from the squared modulus of the electrical component of the electromagnetic field. Furthermore, because here we are not concerned with magneto-optical effects ($\mu_r = 1$) it is sufficient to consider solely the electrical field component.

28.3

Interference of Electromagnetic Waves

As already stated, the interference of light is a necessary prerequisite for image formation. In the following, the influence of polarization on the interference pattern of electromagnetic waves is discussed.

28.3.1

Two-beam Vector Interference

In linear media, electromagnetic fields satisfy the superposition principle, i.e., at each instant of time the complete electric and magnetic field is the sum of the respective fields of the interfering waves. The superposition of two monochromatic plane electromagnetic waves of frequency $\omega/2\pi$ yields the electric field

$$\vec{E}(\vec{r})e^{-i\omega t} = \left(\vec{E}_1(\vec{r}) + \vec{E}_2(\vec{r}) \right) e^{-i\omega t}. \quad (28-2)$$

The considerations developed for the partial coherence of scalar waves are valid here, too. The additional aspect is that only parallel field components can interfere. The field strength of orthogonal components is not added; their intensity superposes independently.

The energy density becomes

$$\rho_E = \frac{n}{2} \left\| \vec{E}_1(\vec{r}) + \vec{E}_2(\vec{r}) \right\|^2 = I_0(\vec{r})(1 + \gamma(\vec{r}) \cos \Delta\varphi(\vec{r})), \quad (28-3)$$

where I_0 denotes the mean energy density

$$I_0 = \frac{n}{2} \left(\left\| \vec{E}_1 \right\|^2 + \left\| \vec{E}_2 \right\|^2 \right), \quad (28-4)$$

$\Delta\varphi$ the phase difference

$$\Delta\varphi = \arg(\vec{E}_2) - \arg(\vec{E}_1), \quad (28-5)$$

and γ the contrast

$$\gamma = \frac{2 \left| \vec{E}_1 \cdot \vec{E}_2^* \right|}{\left\| \vec{E}_1 \right\|^2 + \left\| \vec{E}_2 \right\|^2}. \quad (28-6)$$

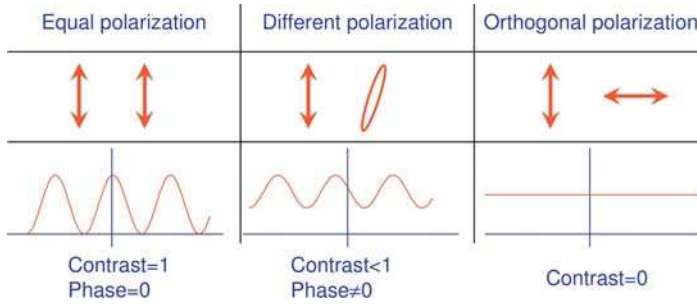


Figure 28-2: Illustration showing that the interference of polarized light “picks” equal polarization states, i.e., the contrast and phase depend on the polarization difference.

The contrast is proportional to the inner product of the field vectors, i.e., to the projection of one polarization state onto another. Coherent waves of equal polarization and intensity yield a perfect contrast (figure 28-2 left). The lateral position of the interference pattern depends on the relative phase difference of both waves. If the polarization state of one wave is changed, the contrast reduces (figure 28-2 center). Now, a contrast of one is no longer feasible. In addition, the phase may change. For an image this may result in a blur and distortion. If the waves are orthogonally polarized the interference pattern vanishes completely (figure 28-2 right). There is no contrast, because the inner product of orthogonal polarization states is zero.

28.3.2

Contrast for High-NA, s- and p-polarization

Of particular importance is the difference between the s and p-polarized interference of electromagnetic plane waves. Consider two linearly polarized plane waves enclosing an angle θ . Their propagation vectors describe a plane. The s-polarized component is oriented perpendicular and the p-polarized component is oriented parallel to this plane (figure 28-3).

Depending on the enclosed angle θ , the interference contrast becomes

$$\gamma = \begin{cases} 1 & : \text{ s-Pol.} \\ \cos(\theta) & : \text{ p-Pol.} \end{cases} \quad (28-7)$$

The dependence of the contrast on the enclosed angle is sketched in figure 28-4. The s-polarized contrast is independent of the angle θ , while the p-polarized contrast decreases to zero for $\theta = 90^\circ$. For larger angles, the magnitude of the contrast increases again but with a changed sign, i.e., the contrast has reversed, that is the position of interference maxima and minima is exchanged.

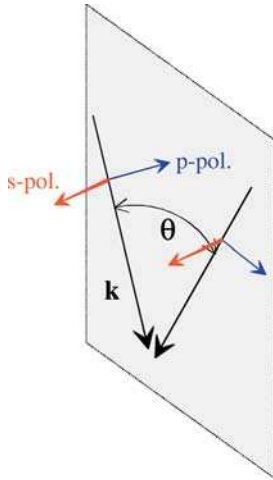


Figure 28-3: Interference of two plane electromagnetic waves. Definition of s- and p-polarization.

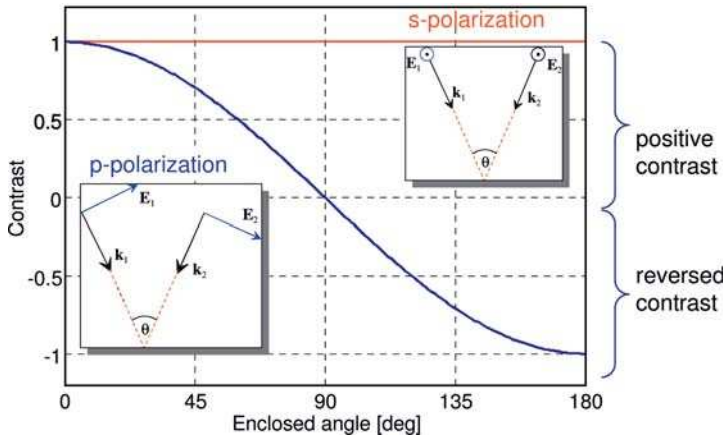


Figure 28-4: Contrast of the interference pattern for two plane waves of equal amplitude, depending on the enclosed angle.

This effect is called the “vector effect”. It introduces a polarization-dependent contrast into the interference of plane waves at high mutual propagation angles.

28.3.3

Influence of Recording Medium

The enclosed angle between the interfering waves is measured *inside* the recording medium. Consider a given enclosed angle θ_i of the propagating plane waves before the recording medium. The contrast for p-polarization becomes less for a recording medium of high refractive index because the refraction at the interface reduces the enclosed angle (cf. figure 28-5) according to

$$\theta = \arcsin \left(\frac{n_i \sin \theta_i}{n_R} \right). \quad (28-8)$$

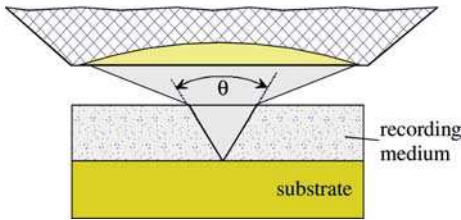


Figure 28-5: Reduction of enclosed angle θ , due to refraction at the interface with the recording medium.

28.3.4

Vector Effect in Optical Microscopy

This effect is usually of no importance for image formation in optical microscopy, not even for high-resolution microscopy. This is because in microscopy the high angles occur in the object space, but not in the image space. Consider a microscope lens of (object side) numerical aperture NA , magnification M and a refractive index n_R of the recording medium. The maximum enclosed angle occurring in the recording medium is

$$\theta_{\max} = 2 \arcsin \frac{NA}{Mn_R}. \quad (28-9)$$

Table 1 shows some typical values.

Table 1: Vector effect for microscopy lenses.

NA	M	n_R	θ [°]	p-pol. contrast ($\cos\theta$)
0.2	10	1.5	1.4	0.9997
0.6	50	1.5	0.9	0.9998
0.9	100	1.5	0.7	0.99993

28.3.5

Vector Effect in Optical Lithography

In high-NA optical lithography a de-magnified image of the object is produced and the vector effect is important. As an example, consider a point-symmetrical configuration where both propagation directions enclose the same angle with the normal to the surface of the recording medium. We are interested in the contrast we can

achieve for a certain pitch p : The enclosed angle θ depends on the pitch p , the wavelength λ , and the refractive index n_R of the recording medium according to

$$\theta = 2 \arcsin \frac{\lambda}{2pn_R}. \quad (28-10)$$

Combination of eqs (28-10) and (28-7) yields the dependence of the contrast on the half-pitch. It is shown in figure 28-6 for a resist refractive index of 1.7 and a recording wavelength of 193 nm (ArF laser). The s-polarized contrast is independent of the pitch, while the p-polarized one drops down to zero at a pitch of 80 nm. For still smaller half-pitches the contrast increases again, but with reversed sign, meaning that bright and dark regions are exchanged. The contrast of the unpolarized interference is just the mean value of the s and p-polarized contrast. It becomes zero for antiparallel propagation within the resist.

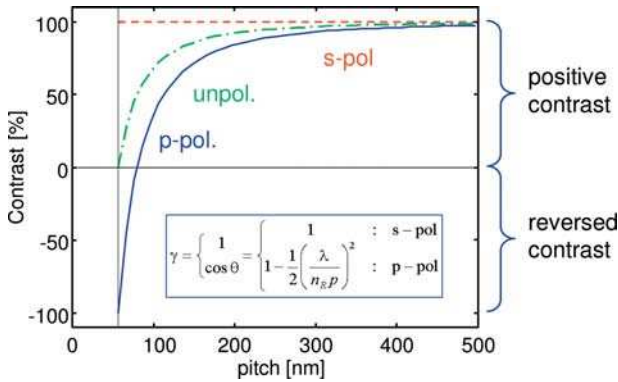


Figure 28-6: Contrast of the interference pattern of two plane waves of equal amplitude depending on the pitch in a medium of $n_R = 1.7$ and with a wavelength of $\lambda = 193$ nm.

28.4

Polarized Ray Trace

To evaluate and optimize the performance of optical systems that contain polarization-changing elements such as, for instance, birefringent lenses, it is necessary to perform a polarization optical ray trace. A ray in a scalar ray trace is characterized by its direction and amplitude. The phase of the corresponding plane wave is obtained from the path.

In a polarization optical ray trace, a ray is also characterized by its polarization state. In a purely scalar optical system, the polarization state remains constant during propagation. However, in a polarization optical system it may change. These changes are due to retardance and diattenuation resulting from:

- a) interfaces (in general between anisotropic media [28-3]–[28-6]);
- b) thin film coatings (in general anisotropic coatings [28-7]);
- c) propagation in anisotropic materials (see chapter 26);
- d) microstructures because of form birefringence (see chapter 27).

28.4.1

Definition of Ray, Beam and Path

We need to distinguish clearly between the items ray, beam and path when used in the following.

- A *ray* represents a single monochromatic plane wave with distinct properties such as, for example, a distinct propagation direction. The propagation of a ray follows the laws of geometrical optics.
- A *beam* is a monochromatic bundle composed of plane waves with a spectrum of propagation directions. The propagation of a beam then results from the propagation of all the contributing plane waves due to a superposition.
- A *path* is the geometrical pathway of a ray.

28.4.2

Ray-splitting at Anisotropic Elements

A common prerequisite of ray-tracing routines is the “single path model”, stating that a single incident ray generates precisely one output ray. This assumption is violated if anisotropic elements are traversed by the ray. At each interface to an anisotropic medium a ray is split into two rays (figure 28-7).

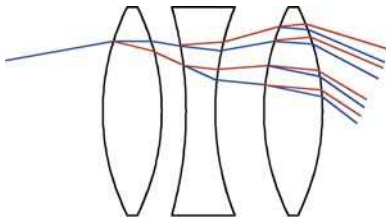


Figure 28-7: Ray-splitting due to double refraction in a compound lens.

In order to compute the optical properties for a particular incident ray, in principle all rays have to be traced through the system. This is precisely what is implemented in some ray-tracing software. However, this method is only feasible for a small number of anisotropic elements because the number N_B of bifurcations increases exponentially with the number N_E of elements

$$N_B = 2^{N_E}. \quad (28-11)$$

28.4.3

Refraction and Reflection at Birefringent Interfaces

The correct ray bifurcation must be computed for correct polarized raytracing through anisotropic elements. Depending on the constellation of anisotropic and birefringent media, up to five rays (one incident, two reflected and two refracted) may occur at a single interface. The various constellations for a single incident plane wave are depicted in figure 28-8 [28-6].

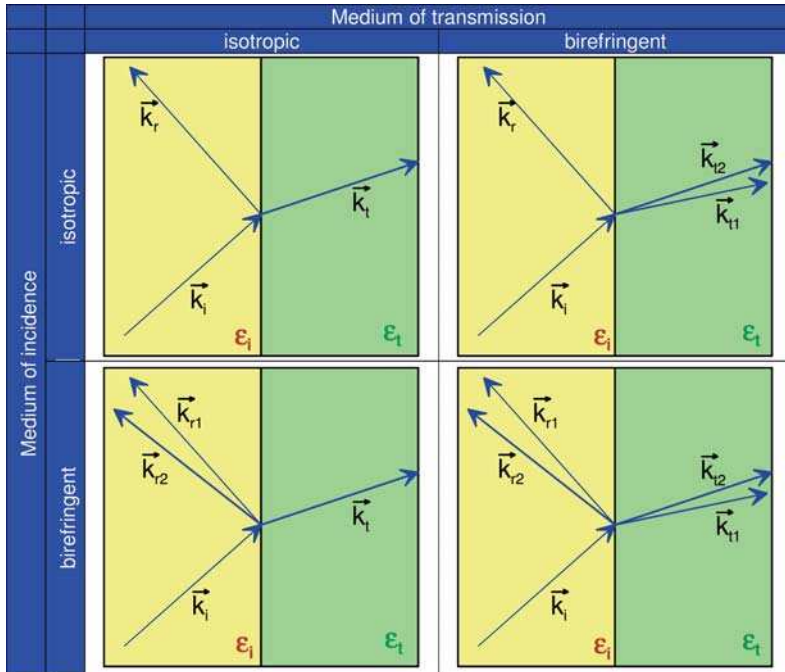


Figure 28-8: Ray-splitting due to double refraction at an interface.

To perform the ray-trace, the direction, polarization amplitude and phase of the reflected and transmitted rays have to be determined for a given incident polarized ray. This is a three-step process, because the direction and polarization of the transmitted and the reflected rays are already determined by the direction of the incident ray and the dielectric tensor of the materials. Knowing the direction and the polarization, the amplitude and phase then have to be determined in a third step.

- 1) The tangential component of the propagation vector of the incident, transmitted and reflected field must be identical, because the continuity of the tangential field components is required for every point on the interface. Because of this the incident, transmitted and reflected propagation vectors must be coplanar. From Fresnel's quartic equation we can compute, for every propagation direction, the corresponding two refractive indices $n_1(\vartheta)$, $n_2(\vartheta)$.

The actual angles of refraction are obtained from a solution of the refraction law

$$\begin{aligned} n_i \sin \vartheta_i &= n_1(\vartheta_1) \sin \vartheta_1 \\ n_i \sin \vartheta_i &= n_2(\vartheta_2) \sin \vartheta_2 \end{aligned} \quad (28-12)$$

for ϑ_1 and ϑ_2 . A similar procedure is applied for the reflected rays.

- 2) With the direction of the refracted rays their polarization is then also fixed. It is the eigenpolarizations, which correspond to the refractive index as eigenvalues of the refracted and reflected rays.
- 3) The polarization of the incident ray determines the amplitude and phase of both the refracted and both the reflected rays. As in the derivation of Fresnel's equations, the continuity of the tangential electric and magnetic field at the interface is applied in order to derive the refraction and reflection coefficient. As an extension to the scalar case, two of these coefficients exist, if the corresponding medium is birefringent (see for instance [28-6]).

28.4.4

The Single-path Approximation

For weakly birefringent media the phenomenon of geometrical ray splitting due to double refraction can be ignored in the same way as a *single-path approximation* is applied ("small birefringence approximation" in [28-5]): Two orthogonal polarized rays traverse the same single path with identical propagation directions from the front to the rear surface of a lens (figure 28-9).

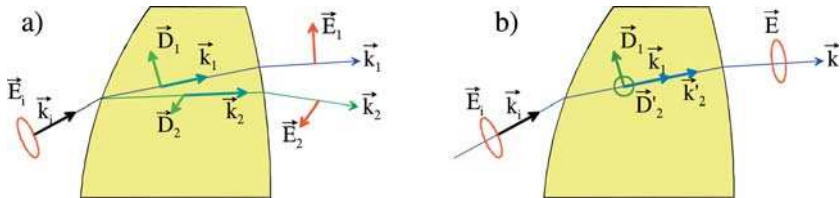


Figure 28-9: Refraction at a birefringent lens: a) exact; b) in a single-path approximation.

This path can be chosen to be either the path of the ray with the greater refractive index or the path of the ray with the smaller one. Along the chosen path both eigenpolarizations, together with their corresponding refractive indices, are computed. The polarization state at the rear surface of the lens then results from the phase difference of both eigenpolarizations (and also from the transitions at the interfaces due to the Fresnel formulas). This approach provides a significant simplification of the polarized optical ray trace. Because of this its range of validity is discussed in detail below.

Validity of the Single-path Approximation for a Single Interface

In the single-path approximation the first refracted ray (\vec{k}_1) follows the same path as the second ray (\vec{k}_2), i.e., the second propagation vector is rotated into the propagation vector of the first ray (figure 28-10).

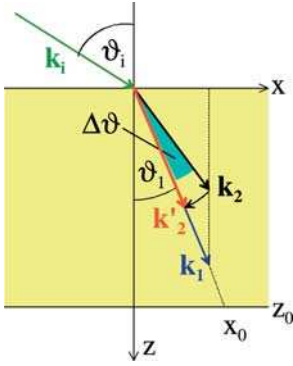


Figure 28-10: Illustration of the single-path approximation.

Consider a point (x_0, z_0) a distance L from the point of incidence, i.e.,

$$\vec{r} = L \frac{\vec{k}_1}{|\vec{k}_1|}. \quad (28-13)$$

While the correct phase difference, including the angular split on the ordinary ray, is

$$\begin{aligned} \Delta\varphi_{\text{exact}} &= (\vec{k}_2 - \vec{k}_1) \cdot \vec{r} = \frac{\vec{k}_2 \cdot \vec{k}_1 - |\vec{k}_1|^2}{|\vec{k}_1|} L \\ &= \left(|\vec{k}_2| \cos(\Delta\vartheta) - |\vec{k}_1| \right) L, \end{aligned} \quad (28-14)$$

the phase difference in the single-path approximation becomes

$$\Delta\varphi_{\text{single ray}} = \left(|\vec{k}_2| - |\vec{k}_1| \right) L. \quad (28-15)$$

The introduced phase error is simply the difference of the correct phase and the approximated phase, i.e.,

$$\begin{aligned} \delta\varphi &= \Delta\varphi_{\text{single ray}} - \Delta\varphi_{\text{exact}} = \left(|\vec{k}_2| - \cos(\Delta\vartheta) |\vec{k}_2| \right) L \\ &= (1 - \cos(\Delta\vartheta)) |\vec{k}_2| L \approx \frac{1}{2} \Delta\vartheta^2 |\vec{k}_2| L \approx -\frac{\pi}{\lambda} \tan^2(\vartheta) \frac{\Delta n^2}{n} L. \end{aligned} \quad (28-16)$$

The amount of the error depends linearly on the traversed length and quadratically on the birefringence. It is small if

$$\frac{\Delta n^2}{\lambda} L \ll 1 \quad (28-17)$$

assuming a wavelength of 200 nm and a maximum tolerable phase difference of 0.001 rad, the regions of validity as sketched in figure 28-11 are obtained. As a rule of thumb: the single-path approximation is valid for $\Delta n < 10^{-4}$.

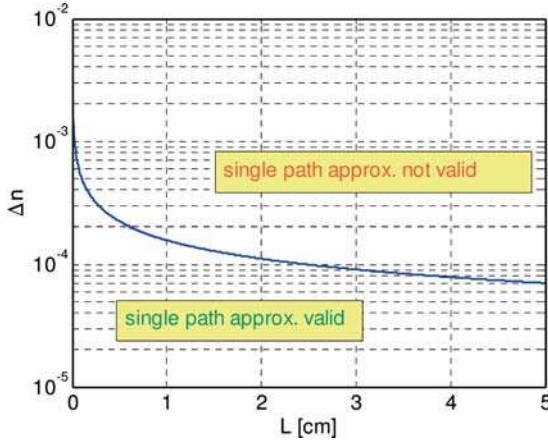


Figure 28-11: Validity of the single-path approximation for a single refraction at $\lambda = 200$ nm and $\delta\varphi_{\max} = 0.001$.

Validity of the Single-path Approximation for an Optical Element

A wedge additionally introduces an angular divergence of the rays as is shown schematically in figure 28-12. Therefore, besides the retardance, we have to consider also the angular divergence in order to judge the validity of the single-path approximation for the wedge.

Consider the incidence of a ray onto a wedge, which is embedded in a homogeneous medium. The angular divergence $\Delta\beta$ between the two rays leaving the wedge, follows from the continuity of the tangential components of the propagation vectors at the second interface. Considering only terms linear in the birefringence Δn , we obtain

$$\Delta\beta \approx W(\vartheta, \gamma) \cdot \Delta n \quad (28-18)$$

with the abbreviation

$$W(\vartheta, \lambda) = \frac{\sin(\vartheta - \gamma) - \cos(\vartheta - \gamma) \tan(\vartheta)}{\sqrt{1 - n^2 \sin^2(\vartheta - \gamma)}}. \quad (28-19)$$

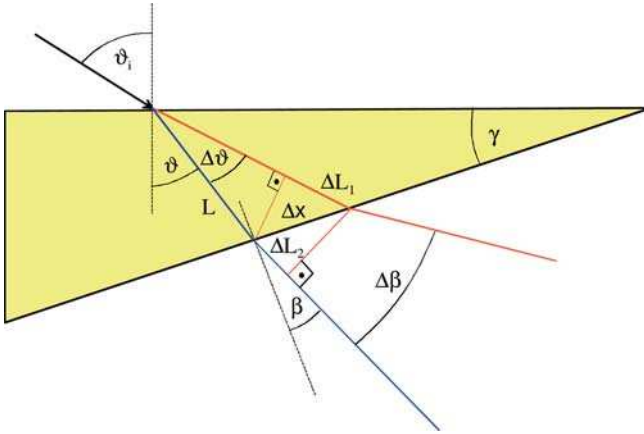


Figure 28-12: Angular divergence at a wedge.

This equation exhibits a critical angle where total reflection occurs and where the deviation $\Delta\beta$ becomes arbitrarily high. However, a few degrees away from this limit, W settles to a magnitude of about 1.0.

For the wedge, an estimation of the phase error in the single-path approach is given by

$$\delta\varphi \approx -2\pi \frac{L}{\lambda} \frac{\Delta n^2}{n} \frac{\tan \vartheta \tan(\vartheta - \gamma)}{\cos(\vartheta)}. \quad (28-20)$$

Because of the factor $\tan(\vartheta - \gamma)$ the phase error for the wedge increases dramatically for $|\vartheta - \gamma| \rightarrow 90^\circ$ i.e. near total reflection. However, some degrees away from this critical regime a rough estimation for the absolute value of the last factor in eq. (29) is 1.0, and we obtain a typical maximum phase error of

$$|\delta\varphi_{\text{typ,max}}| \approx 2\pi \frac{d}{\lambda} \frac{\Delta n^2}{n}, \quad (28-21)$$

i.e., for $d \approx 1 \text{ m} = 10^7 \lambda$ and $\Delta n \approx 10^{-6}$ ($n \approx 1.5$) the maximum phase error is about $5 \cdot 10^{-5}$ rad.

While, for a plane parallel plate, the phase error behind the plate is independent of the distance from its rear surface because the transmitted rays are parallel, this is no longer true for the wedge. In the case of the wedge the two emerging rays enclose an angle $\Delta\beta$ with the consequence that the phase error increases with the distance from the wedge. An estimation for this phase error is roughly given by

$$\delta\varphi_{\text{trav}} = \frac{2\pi}{\lambda} L_{\text{trav}} [1 - \cos(\Delta\beta)] \approx \pi \frac{L_{\text{trav}}}{\lambda} \Delta\beta^2 \approx \pi \frac{L_{\text{trav}}}{\lambda} \Delta n^2, \quad (28-22)$$

i.e., this phase error is proportional to $L_{\text{trav}} \Delta n^2 / \lambda$. For $\lambda = 157 \text{ nm}$, $\Delta\beta \approx \Delta n = 10^{-6}$ and a distance of $L_{\text{trav}} = 1 \text{ m}$ we obtain $\delta\varphi_{\text{trav}} \approx 2.0 \cdot 10^{-5} \text{ rad} \approx 1.1 \cdot 10^{-3}^\circ$.

Taking the sum of the equations (28-21) and (28-22), a single lens introduces a phase error of approximately

$$\delta\varphi_{\text{single lens}} = k_0 L \Delta n^2 \quad (28-23)$$

with $k_0 = 2\pi/\lambda$ and $L \approx L_{\text{trav}} + d$.

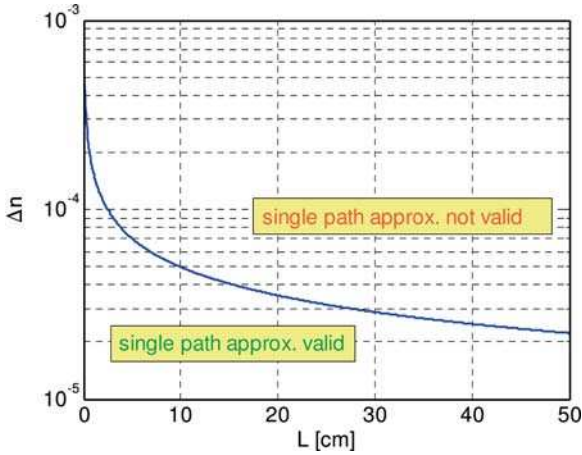


Figure 28-13: Validity of the single-path approximation for a single element at $\lambda = 200$ nm and $\delta\varphi_{\text{max}} = 0.001$ beyond the critical angle.

Validity of the Single-path Approximation for an Optical System

For several lenses in succession a crude estimation for the resulting phase error is a simple superposition of the phase of all individual lenses. If we assume N equally spaced lenses separated by the same distance L , and if we assume that each lens increases the angular divergence additionally by the same $\Delta\vartheta \approx \Delta n$, we obtain for the phase error of the whole system

$$\delta\varphi_{\text{System}} = k_0 \sum_{j=1}^N (j\Delta n_j)^2 L_j \approx k_0 \Delta n^2 L \sum_{j=1}^N j^2 = k_0 \Delta n^2 L \frac{N(N+1)(2N+1)}{6}. \quad (28-24)$$

For $N = 30$ lenses, separated by the distance $L = 5$ cm, $\lambda = 157$ nm and $\Delta n = 10^{-6}$, we obtain $\delta\varphi_{\text{Lenses}} \approx 0.02$ rad $\approx 1.0^\circ$, which corresponds to an upper bound of $3 \cdot 10^{-3} \lambda \approx 0.5$ nm at $\lambda = 157$ nm.

For $N = 10$ lenses, the error is about $8 \cdot 10^{-4}$ rad, which corresponds to 0.02 nm.

28.5

Optical Systems with Polarization Effects

Polarization has to be taken into account for any high-accuracy optical design and simulation. Of course this is particularly true for optical systems that make use of the polarization of light such as, for instance, projectors based on liquid crystals or lenses for polarization microscopy. But it applies also to optical systems with an extremely high level of aberration-correction like lenses for optical projection lithography.

Optical lenses for polarization microscopy have to be made of stress-free glasses where stress-induced birefringence is minimized. Furthermore, the anti-reflection coatings have to be optimized for their polarization performance. Figure 28-14 shows as an example a microscope lens for polarization microscopy. The anti-reflection coatings are critical at the rim of the lenses where high angles of incidence occur. In particular, near the Brewster angle of the lens material strong polarization effects occur which have to be neutralized by a proper coating design.

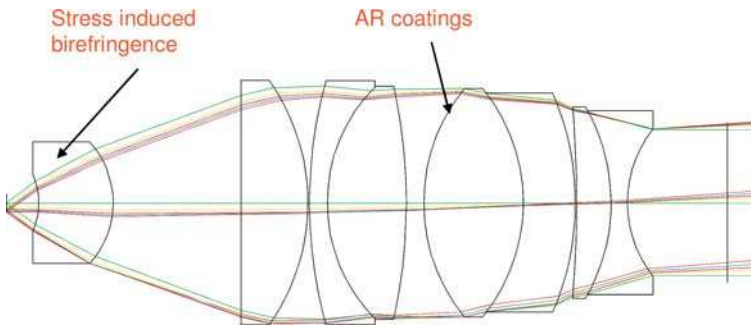


Figure 28-14: Examples of an optical system where polarization effects are of importance: Lens for polarization microscopy (Zeiss lens, Patent WO2004092802).

Figure 28-15 shows the drawing of a projection lens for optical lithography at 193 nm. The lenses are made of quartz (amorphous SiO_2) and calcium fluoride (crystalline CaF_2). The height of the system is 1000 mm, the image field is approximately 20 mm in diameter and the scalar aberrations of the wavefront are corrected down to an RMS of five thousandth of a wavelength ($5 \text{ m}\lambda$).

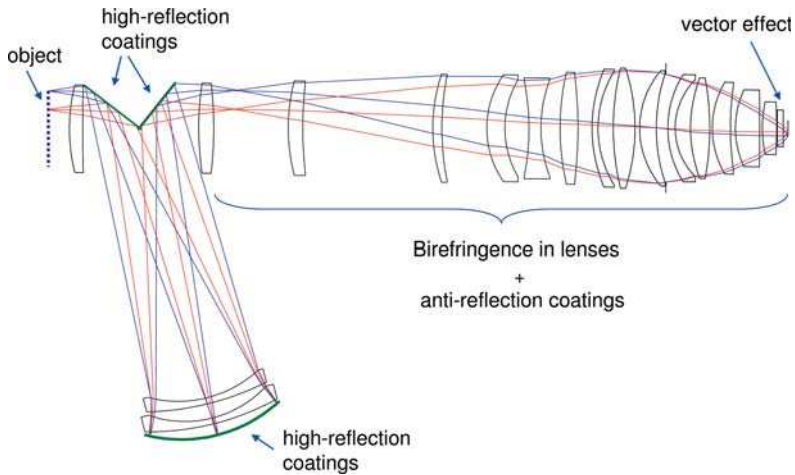


Figure 28-15: Example of optical system where polarization effects are of importance: Lens for VUV projection lithography. Zeiss catadioptrical lens, for illustration only, US 6,717,746 B2.

The sources for polarization effects are:

- Intrinsic birefringence of the CaF_2 lenses (cf. chapter 18).
- Stress-induced birefringence in quartz and CaF_2 .
- Polarization-dependent transmission of the anti-reflection (AR) coatings on the lens surfaces.
- Polarization-dependent reflection at the high-reflection (HR) coatings of the mirrors.
- Polarization-dependent object transmission – in particular for structure sizes that approach the wavelength.
- Polarization dependence of the interference contrast for high numerical apertures.

28.6 Polarized Imaging Model

For simulation and also even for the description of polarization effects on imaging we need an imaging model that we will briefly describe here. The starting point is an extension of the pupil-based model of scalar imaging to polarization, following the lines of ref. [28-8]. The geometry is sketched in figure 28-16. The illumination pupil, entrance pupil, and exit pupil are spherical due to the precise fulfilment of the sine condition. According to the illumination setting, it is assumed that the illumination pupil is filled in a sigma region with an incoherent source (Hopkins effective source). To produce an image we start with a single point in this region. It radiates a plane wave onto the object.

The illuminating plane wave is diffracted by the object, producing an angular spectrum of plane waves, i.e., a number of mutually coherent plane waves, that propagate in different directions. Within the numerical aperture (NA) of the object side of the lens (usually $NA/4$) they are captured by the entrance pupil and enter the imaging system (the lens). The lens transfers the diffraction orders with a linear scaling factor of the pupil coordinate (= demagnification) to the exit pupil. There, each diffraction order generates a plane wave that radiates into the image plane, where it produces the structure image by interference with the plane waves stemming from the other diffraction orders.

Any aberrations are assigned to the transfer from the entrance pupil to the exit pupil via the pupil plane. This procedure is repeated for the completely filled section (sigma-setting) of the illumination pupil.

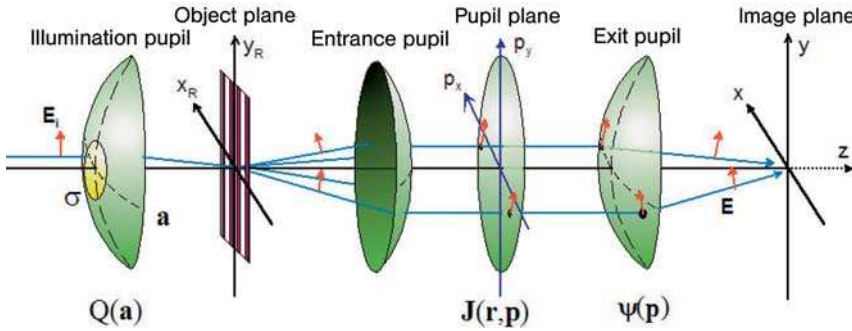


Figure 28-16: Geometry of the imaging model.

28.6.1

Scalar Image

In accordance with the description above, the intensity of the coherent scalar image is computed as a plane wave superposition integral of the diffraction spectrum $T(p,a)$ of the object, times the scalar pupil function $L(p)$ describing the wavefront aberrations and pupil limits [28-9]:

$$I_{\text{coh}}(\vec{r}, \vec{a}) = |F\{T(\vec{p}, \vec{a})L(\vec{p})\}|^2. \quad (28-25)$$

\vec{p} denotes the pupil coordinates and \vec{a} the illumination pupil coordinates. The plane wave superposition integral $F\{\}$ is just a 2D Fourier transform integral, where the kernel contains the z -propagation operator

$$F(f(\vec{p})) = \iint_{-\infty}^{\infty} f(p_x, p_y) \cdot e^{i\vec{k} \cdot \vec{r}} dp_x dp_y. \quad (28-26)$$

The pupil coordinates (p_x, p_y) are related to the propagation vector \vec{k} of the corresponding plane wave in the object or image space according to the formula

$$\vec{k} = k_0 \begin{pmatrix} -p_x \\ -p_y \\ \sqrt{n^2 - p_x^2 - p_y^2} \end{pmatrix}, \quad (28-27)$$

where n is the refractive index of the medium in the image or object space and $k_0 = 2\pi/\lambda$. Equation (28-27) describes the rotation of the propagation vector from a ray parallel to the optical axis to a ray that traverses the ideal image point as if refracted by a perfect collimator. The negative sign of the x- and y-component stems from the fact that the coordinate system in the pupil is chosen to describe the wavefront. But the k-vectors describe the corresponding plane wave spectrum. Both are just the negative of each other. Because the optical system is assumed to fulfil the sine condition, the relationship between the pupil coordinates of the entrance pupil (p_x^R, p_y^R) and exit pupil (p_x, p_y) is given with the magnification M according to

$$p_x = \frac{p_x^R}{M}, \quad p_y = \frac{p_y^R}{M}. \quad (28-28)$$

Note that M is a signed quantity. For microscope objectives, for instance, the magnification M might be 100 with an image inversion, i.e., $M = -100$. Without image inversion, M is positive. The pupil coordinates of the exit pupil are the coordinates in which the systems pupil is described, i.e., they will be termed “pupil coordinates”.

Eventually, the complete image for a given setting is obtained from the superposition of the intensities of the coherent images

$$I(\vec{r}) = \int \int_{\sigma} Q(\vec{a}) I_{\text{coh}}(\vec{r}, \vec{a}) d^2 \vec{a} \quad (28-29)$$

where Q denotes the sigma-setting of the illumination.

28.6.2

Vector Image for Completely Polarized Illumination

For the computation of the vector image we have to deal with the vector diffraction pattern of the object and its propagation. In vector diffraction, a diffraction order is not characterized by the scalar quantities of amplitude and phase, but by its Jones vector. The Jones vector denotes the polarization state of light by a complex two-component vector (two components are sufficient because far-fields are transverse). It contains the state of polarization, the amplitude, and the phase of the wave. To derive the coherent vector image from the scalar one we have to make the following replacements:

The scalar angular spectrum $T(\vec{p}, \vec{a})$ of the mask diffraction field has to be replaced by an angular spectrum of the Jones vectors

$$\vec{E}_{\text{Mask}}(\vec{p}, \vec{a}, \vec{E}_i) = \begin{pmatrix} E_x(\vec{p}, \vec{a}, \vec{E}_i) \\ E_y(\vec{p}, \vec{a}, \vec{E}_i) \end{pmatrix}. \quad (28-30)$$

This depends on the pupil coordinate \vec{p} , on the direction \vec{a} , and the polarization \vec{E}_i of the illuminating plane wave.

The scalar pupil function $L(\vec{p})$ has to be replaced by a Jones-matrix pupil function

$$\mathbf{J}(\vec{p}) = \begin{pmatrix} J_{xx}(\vec{p}) & J_{xy}(\vec{p}) \\ J_{yx}(\vec{p}) & J_{yy}(\vec{p}) \end{pmatrix} \quad (28-31)$$

describing a Jones matrix for each pupil coordinate. This will be discussed in more detail in the next section.

Because of the vector interference (cf. section 28.3) the actual 3D electric field in the resist has to be computed from the Jones-vector distribution in the pupil

$$\begin{pmatrix} E'_x(p_x, p_y) \\ E'_y(p_x, p_y) \\ E'_z(p_x, p_y) \end{pmatrix} = \boldsymbol{\psi}(\vec{p}) \begin{pmatrix} E_x(p_x, p_y) \\ E_y(p_x, p_y) \end{pmatrix}, \quad (28-32)$$

where the 3×2 transfer matrix $\boldsymbol{\psi}(\vec{p})$

$$\boldsymbol{\psi}(\vec{p}) = \begin{pmatrix} 1 - \frac{p_x^2}{n(n + \sqrt{n^2 - p_x^2 - p_y^2})} & -\frac{p_x p_y}{n(n + \sqrt{n^2 - p_x^2 - p_y^2})} \\ -\frac{p_x p_y}{n(n + \sqrt{n^2 - p_x^2 - p_y^2})} & 1 - \frac{p_y^2}{n(n + \sqrt{n^2 - p_x^2 - p_y^2})} \\ \frac{p_x}{n} & \frac{p_y}{n} \end{pmatrix}, \quad (28-33)$$

of the Jones vector in the pupil, to the 3D E -field in image space, follows from the change in direction of the propagation vector. It is a succession of three rotational matrices which provides this change of direction

$$\begin{aligned} \vec{E} &= \begin{pmatrix} \cos \varphi & -\sin \varphi & 0 \\ \sin \varphi & \cos \varphi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos \vartheta & 0 & \sin \vartheta \\ 0 & 1 & 0 \\ -\sin \vartheta & 0 & \cos \vartheta \end{pmatrix} \\ &\times \begin{pmatrix} \cos \varphi & \sin \varphi & 0 \\ -\sin \varphi & \cos \varphi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} E_x \\ E_y \\ 0 \end{pmatrix} \end{aligned} \quad (28-34)$$

with (cf. figure 27-2)

$$\sin \vartheta = \frac{1}{n} \sqrt{p_x^2 + p_y^2}, \quad \sin \varphi = \frac{p_y}{\sqrt{p_x^2 + p_y^2}}, \quad \cos \varphi = \frac{p_x}{\sqrt{p_x^2 + p_y^2}}. \quad (28-35)$$

Putting all this together we obtain the coherent vector image intensity (which is not a vector in the mathematical sense)

$$\begin{pmatrix} I_x \\ I_y \\ I_z \end{pmatrix}_{\text{coh}}(\vec{r}, \vec{a}) = \left| F_\alpha \left\{ (\vec{p})\mathbf{J}(\vec{p})\vec{E}_{\text{Mask}}(\vec{a}, \vec{p}, \vec{E}_i) \right\} \right|^2 \quad (28-36)$$

where the plane wave superposition integral is applied to each component separately and

$$I_{\text{coh}} = I_x + I_y + I_z. \quad (28-37)$$

These definitions allow the separation of the impact of the imaging system, described by the Jones pupil \mathbf{J} , from the high-NA effects described by the 2D–3D transformation matrix ψ . This makes sense because the Jones pupil describes a lens property and can be optimized in lens design, whereas the high-NA effects stem from the 3D geometry in the image plane.

28.6.3

Vector Image for Partially Polarized Illumination

So far we have described polarized imaging for completely polarized illumination only. However, the degree of polarization g , defined as the intensity in a polarized state I_{pol} divided by the overall intensity

$$g = \frac{I_{\text{pol}}}{I_{\text{pol}} + I_{\text{unpol}}} \quad (28-38)$$

of the illumination, will usually be <1 so that the field will not be completely polarized but will be mostly polarized with a small unpolarized component ($g \leq 1$). At the other end of the spectrum we have unpolarized illumination which consists of a large unpolarized component and a small polarized one ($g \geq 0$). The vector-imaging model of the preceding section will now be extended to include partially polarized illumination as well as unpolarized illumination.

The key idea for the extension is as follows: If light is unpolarized, every possible polarizer (linear, circular, elliptical) transmits 50% of the incident intensity. This is equivalent to the statement that every polarization state contains half of the overall intensity. Unpolarized light cannot be represented by a single Jones vector because Jones vectors describe completely polarized states, by definition. But it can be represented in the Jones vector calculus by the incoherent superposition of two orthogonal polarization states $\vec{E} = (E_x, E_y)$ and $\vec{E} = (-E_y^*, E_x^*)$ of equal intensity. For partial polarization the intensities of both states differ by $(1 - g)/(1 + g)$:

$$\text{Intensity of partially polarized light} \triangleq \frac{1 + g}{2} \|\vec{E}\|^2 + \frac{1 - g}{2} \|\tilde{E}\|^2. \quad (28-39)$$

For the simulation of partially polarized illumination, instead of one coherent image for a given point in the sigma-region, we have to compute two images. One for the

original illumination polarization state \vec{E}_i and one for its orthogonal complement \vec{E}_i^\perp . Because they are not correlated, the image intensity for partially polarized illumination is the sum of both intensities weighted with prefactors according to eq. (28-39) that depend on the degree g of polarization, i.e.

$$\begin{pmatrix} I_x \\ I_y \\ I_z \end{pmatrix}_{\text{coh}}(\vec{r}, \vec{a}) = \quad (28-40)$$

$$\left| F_\alpha \left\{ \sqrt{\frac{1+g(\vec{p})}{2}} (\vec{p}) \mathbf{J}(\vec{p}) \vec{E}_{\text{Mask}}(\vec{a}, \vec{p}, \vec{E}_i) \right\} \right|^2 + \left| F_\alpha \left\{ \sqrt{\frac{1-g(\vec{p})}{2}} (\vec{p}) \mathbf{J}(\vec{p}) \vec{E}_{\text{Mask}}(\vec{a}, \vec{p}, \vec{E}_i^\perp) \right\} \right|^2.$$

28.7

Vector Point-spread Function

The vector point-spread function VPSF, is an extension of the scalar point-spread function (PSF) to polarized illumination.

28.7.1

VPSF for Complete Polarization

The focus of a polarized pupil was computed in chapter 27. In the present section we assume that the Jones vector field in the exit pupil stems from the transmission of an incident homogeneous polarization \vec{E}_i by a Jones matrix pupil $\mathbf{J}(p_x, p_y)$.

$$\vec{E}(p_x, p_y) = \mathbf{J}(p_x, p_y) \vec{E}_i. \quad (28-41)$$

The VPSF for a completely polarized illumination is the Fourier transform of the individual vector components after they are transformed by ψ into 3D-space and a possible pupil-apodization factor ($\cos^{-1/2}\vartheta' \cos^{1/2}\vartheta \dots$ for systems that satisfy the sine condition) was then applied

$$\vec{h}(x, y) = F \left\{ \psi(p_x, p_y) \mathbf{J}(p_x, p_y) \vec{E}_i \right\}. \quad (28-42)$$

where $F\{\}$ denotes the component-wise Fourier transform.

We can define the “vector intensity” of the VPSF as the intensity of the individual components of the VPSF

$$I(x, y) = |\vec{h}(x, y)|^2 = |F \{ \psi(p_x, p_y) \mathbf{J}(p_x, p_y) \vec{E}_i \}|^2. \quad (28-43)$$

The scalar intensity is the norm of the VPSF

$$I(x, y) = \left\| \vec{h}(x, y) \right\|^2 = \left\| F \left\{ \mathbf{T}(p_x, p_y) \mathbf{J}(p_x, p_y) \vec{E}_i \right\} \right\|^2. \quad (28-44)$$

Foci for various input polarizations have been shown in chapter 27. An example of the x -, y -, and z -component and overall intensity of a linearly polarized focus is shown in figure 28-17.

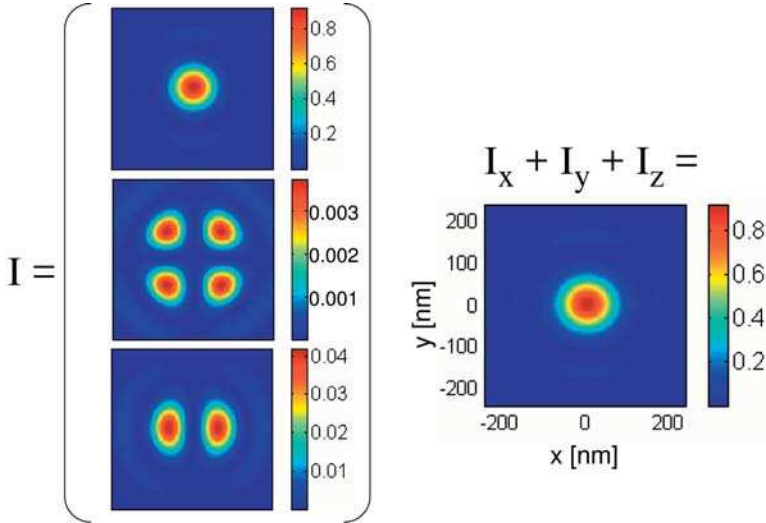


Figure 28-17: Linearly polarized focus.

28.7.2

VPSF for Unpolarized Illumination

For unpolarized illumination the intensity of the point-spread functions for two orthogonal incident polarization states has to be formed. If the incident states are just $\vec{E}_x = [1, 0]$ and $\vec{E}_y = [0, 1]$ this reduces to a superposition of the intensities of the columns of the Jones matrix pupil

$$I_{\text{unpol}}(x, y) = \left\| F \left\{ \psi(p_x, p_y) \mathbf{J}(p_x, p_y) \vec{E}_x \right\} \right\|^2 + \left\| F \left\{ \psi(p_x, p_y) \mathbf{J}(p_x, p_y) \vec{E}_y \right\} \right\|^2. \quad (28-45)$$

In the paraxial case (small NA) ψ has a vanishing effect and the equation simplifies further to

$$\begin{aligned} I_{\text{unpol}}(x, y) &= \left\| F \left\{ \mathbf{J}(p_x, p_y) \vec{E}_x \right\} \right\|^2 + \left\| F \left\{ \mathbf{J}(p_x, p_y) \vec{E}_y \right\} \right\|^2 \\ &= |F\{J_{xx}(p_x, p_y)\}|^2 + |F\{J_{xy}(p_x, p_y)\}|^2 + |F\{J_{yx}(p_x, p_y)\}|^2 + |F\{J_{yy}(p_x, p_y)\}|^2. \end{aligned} \quad (28-46)$$

28.8

Polarized Optical Transfer Function

28.8.1

Polarized Illumination

For completely polarized *coherent* illumination, the optical transfer function is just the Jones matrix pupil.

$$PCTF = \boldsymbol{\psi}(p_x, p_y) \mathbf{J}(p_x, p_y). \quad (28-47)$$

For completely polarized, *incoherent* illumination, we define the optical transfer function as the inverse Fourier transform of the point-spread function intensity and obtain

$$POTF = \left(\mathbf{T}(p, q) \mathbf{J}(p, q) \vec{\mathbf{E}}_i \right) \otimes \left(\mathbf{T}(p, q) \mathbf{J}(p, q) \vec{\mathbf{E}}_i \right)^*, \quad (28-48)$$

where we denote with the autocorrelation of a vector the autocorrelation of its individual components.

28.8.2

Unpolarized Illumination

For *unpolarized* illumination in the paraxial case, we obtain the optical transfer function as the autocorrelation of the single elements of the Jones matrix pupil

$$UOTF = J_{xx} \otimes J_{xx}^* + J_{xy} \otimes J_{xy}^* + J_{yx} \otimes J_{yx}^* + J_{yy} \otimes J_{yy}^*, \quad (28-49)$$

where the dependence on the pupil coordinates has been omitted for clarity. If the autocorrelation of a matrix is defined as the sum of the autocorrelation of the matrix elements, then a shorthand notation is

$$UOTF = \mathbf{J} \otimes \mathbf{J}^*. \quad (28-50)$$

28.9

Jones Matrix Pupil

The Jones matrix pupil, considered as an extension of the scalar pupil to optical systems which change the state of polarization, was introduced, to our knowledge, by W. Urbanczyk [28-10], [28-11]. The concept of polarization aberrations is due to Chipman et al. [28-12], [28-13]. He applied a direct decomposition of the Jones pupil into elementary Jones pupils by means of the Pauli spin matrices.

28.9.1

Definition for Completely Polarized Illumination

A polarization optical system changes the polarization state of the transferred light. To this end the scalar pupil is extended to a Jones-matrix pupil, i.e., to each point of the pupil a Jones matrix is attributed which yields the change in the polarization state between the entrance and exit pupil (figure 28-18)

$$\begin{pmatrix} E_x^{\text{out}}(p_x, p_y) \\ E_y^{\text{out}}(p_x, p_y) \end{pmatrix} = \begin{pmatrix} J_{xx}(p_x, p_y) & J_{xy}(p_x, p_y) \\ J_{yx}(p_x, p_y) & J_{yy}(p_x, p_y) \end{pmatrix} \begin{pmatrix} E_x^{\text{in}}(p_x, p_y) \\ E_y^{\text{in}}(p_x, p_y) \end{pmatrix} \quad (28-51)$$

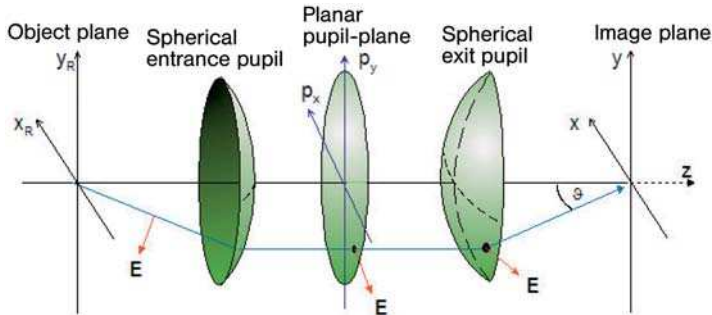


Figure 28-18: Geometry and coordinates for the Jones matrix pupil.

and in a condensed form

$$\vec{E}_{\text{out}}(p_x, p_y) = \mathbf{J}(p_x, p_y) \vec{E}_{\text{in}}(p_x, p_y). \quad (28-52)$$

For a scalar optical system the Jones matrix pupil is equal to a complex scalar factor multiplied by the identity matrix

$$\mathbf{J}(p_x, p_y) = A_0(p_x, p_y) e^{i\psi_0(p_x, p_y)} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \quad (28-53)$$

As an example we consider an artificial Jones pupil constructed from a quadratically increasing radial retardance of max. 20 nm, a corresponding diattenuation of max. 20%, and 50 nm defocus at an NA of 0.9. A retardance with radial symmetry is quite typical for high-NA optical systems and is therefore the subject of research [28-14], [28-15]. To make our example more interesting we have constructed the radial retardance from a superposition of the intrinsic birefringence of two 111 CaF₂ plates that are rotated by 60° with respect to each other. The resulting Jones pupil is depicted in figure 28-19. Compared to real optical systems it shows strongly exaggerated polarization effects.

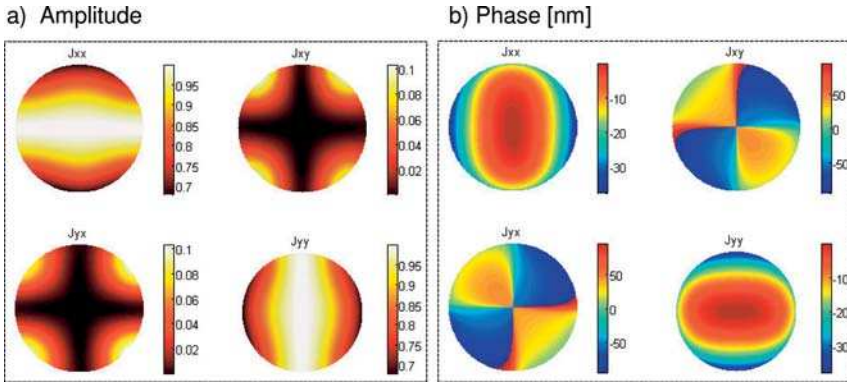


Figure 28-19: Jones pupil amplitude (a) and phase (b) in nm.

The main amplitude is situated on the diagonal elements (figure 28-19a) indicating that the cross-talk between the polarizations, i.e., the spill-over from x -polarization to y -polarization and vice versa is still quite small (the amplitude of 0.1 corresponds to an intensity of 1%). Due to the radial symmetry of the diattenuation and birefringence the cross-talk becomes a maximum under 45° and 135° . The phase distribution (figure 28-19b) of the diagonal elements is a combination of the defocus and birefringence. It is continuous across the pupil. The phase distribution of the non-diagonal elements, however, is discontinuous. It shows a phase jump of π corresponding to 96.5 nm. Consequently, the amplitude maxima of J_{xy} and J_{yx} are π -phase shifted with respect to their neighbors. This does not indicate a discontinuity of the complex Jones matrix element, because the real and imaginary part of the Jones pupil element are still continuous. The position of the phase discontinuity is the x and y -axis which correspond to the roots of the amplitude. There the complex function changes its sign, i.e., the phase experiences a π -phase shift.

28.9.2

Separation of a Scalar Factor

If the scalar aberrations prevail over the polarization aberrations, it is useful to separate a scalar factor from the Jones-matrix pupil, i.e., to write

$$\mathbf{J}(p, q) = A(p, q)e^{i\psi(p, q)}\hat{\mathbf{J}}(p, q) \quad (28-54)$$

and in terms of the Jones vectors

$$\vec{\mathbf{E}}_{\text{out}}(p, q) = A(p, q)e^{i\psi(p, q)}\hat{\mathbf{J}}(p, q)\vec{\mathbf{E}}_{\text{in}}(p, q) . \quad (28-55)$$

However, the solution is not unique. A straightforward solution would be to factor out a matrix element and to refer the remaining elements to it

$$\mathbf{J} = J_{xx} \begin{pmatrix} 1 & J_{xy}J_{xx}^{-1} \\ J_{yx}J_{xx}^{-1} & J_{yy}J_{xx}^{-1} \end{pmatrix} = J_{xx} \hat{\mathbf{J}} \quad (28-56)$$

where the dependence on p, q has been neglected for clarity. However, because this method particularly emphasizes one Jones-matrix element, we do not recommend it.

Alternatively the prefactor can be referred to the mean value of all four matrix elements

$$\mathbf{J} = K \begin{pmatrix} J_{xx}K^{-1} & J_{xy}K^{-1} \\ J_{yx}K^{-1} & J_{yy}K^{-1} \end{pmatrix} \quad (28-57)$$

with

$$K = \frac{1}{4} (J_{xx} + J_{xy} + J_{yx} + J_{yy}), \quad (28-58)$$

or it can be referred to the mean value of the diagonal elements:

$$\mathbf{J} = K_D \begin{pmatrix} J_{xx}K_D^{-1} & J_{xy}K_D^{-1} \\ J_{yx}K_D^{-1} & J_{yy}K_D^{-1} \end{pmatrix} \quad (28-59)$$

with

$$K_D = \frac{1}{2} (J_{xx} + J_{yy}). \quad (28-60)$$

28.9.3

Decomposition into Retardance and Diattenuation

Applying the polar decomposition of a Jones matrix from section 26.3 to the Jones matrix pupil, yields the pupil decomposition

$$\mathbf{J} = (\mathbf{UDU}^+)(\mathbf{UV}^+) = \mathbf{J}_p \mathbf{J}_R.$$

$\mathbf{J}_p = \mathbf{UDU}^+$ denotes a partial polarizer because its eigenvalues are given by the elements of the diagonal matrix \mathbf{D} which are real and positive (they are the singular values).

$\mathbf{J}_R = \mathbf{UV}^+$ denotes a retarder because it is unitary: $\mathbf{UV}^+(\mathbf{UV}^+)^+ = \mathbf{UV}^+\mathbf{VU}^+ = \mathbf{1}$. A unitary matrix can be diagonalized by solution of the eigenequation: $\mathbf{J}_R = \mathbf{W}^{-1}\mathbf{R}\mathbf{W}$ with a unimodular diagonal matrix \mathbf{R} and a transformation matrix \mathbf{W} . Because of its unimodularity ($\det = 1$) the elements of \mathbf{R} are pure retardances.

The Jones pupil can be characterized by the eigenpolarizations and the eigenvalues of the partial polarizer and the retarder, as is summarized below (D_{11} , D_{22} and R_{11} , R_{22} are the diagonal elements of \mathbf{D} and \mathbf{R} , respectively).

Quantity	Formula	Orientation	Cause	Effect
Scalar transmission	$T = \frac{1}{2}(D_{11}^2 + D_{22}^2)$	Scalar	Volume absorption and coating-reflection and absorption	Apodization, Telecentricity error
Diattenuation	$\Delta T_D = \frac{D_{11}^2 - D_{22}^2}{D_{11}^2 + D_{22}^2}$	Axis of higher transmission	Differential transmission of the coatings	Partial polarization, Change of polarization state
Scalar phase	$\varphi = \frac{1}{2}(\arg(R_{11}) + \arg(R_{22}))$	Scalar	Design phase	Aberrations
Retardance	$\Delta\varphi = \arg(R_{22}) - \arg(R_{11})$	Fast axis	Stress-induced birefringence, Intrinsic birefringence, Coating retardance	Change of polarization state

Diattenuation and retardance have a pair of eigenpolarizations and for a complete characterization these must also be provided. The corresponding eigenpolarizations are orthogonal because \mathbf{J}_R is unitary and \mathbf{J}_P is Hermitian. The orientation ϕ of the eigenpolarizations (E_x, E_y) is according to section 26.2.3

$$\tan(2\phi) = \frac{2|E_x||E_y|}{|E_x|^2 - |E_y|^2} \cos \delta,$$

and the ellipticity

$$e = \tan \chi = \pm \frac{b}{a} \quad \text{with} \quad \sin(2\chi) = \frac{2|E_x||E_y|}{|E_x|^2 + |E_y|^2} \sin \delta$$

and $\delta = \arg(E_y) - \arg(E_x)$.

28.9.4

Example

The representation of the artificial Jones pupil of figure 28-19 is shown in figure 28-20.

- The *scalar transmission* (top left) has the shape of an apodization with circular symmetry. Because of the circular symmetry no telecentricity error is to be expected for this example.
- The *diattenuation* (top center) also shows circular symmetry. The higher transmission is in the radial direction. This is in accordance with conventional coatings showing a higher p than s transmission. For unpolarized illumination, the pupil yields 20% degree of polarization at the rim with the p-image exceeding the s-image, and 0% in the centre.

- The *ellipticity* of the eigenpolarizations (top right) is zero, i.e., the eigenpolarizations are linear.
- The *scalar phase* (bottom left) has a maximum value of $(1 - \sqrt{1 - 0.9^2})$ 50 nm = 28 nm at the rim.
- The *retardance* (bottom center) with a maximum value of 20 nm at the rim has its fast axis in the tangential direction, meaning that the higher refractive index (which yields the slow axis) is oriented radially. An unpolarized incident wave would experience focus fading, indicating that both eigenpolarizations are focused to different positions. With a retardance of 20 nm the distance between the s and p-polarized “focus” would be 13 nm. In this context the notion of a focus should be used with caution because s and p-polarized light does not produce a perfect focus in the paraxial region (no interference from pupil coordinates which are 90° apart).
- The three-fold symmetry of the *ellipticity* (bottom right) of the retarder eigenpolarizations stems from the three-fold symmetry of the [111] retardance.

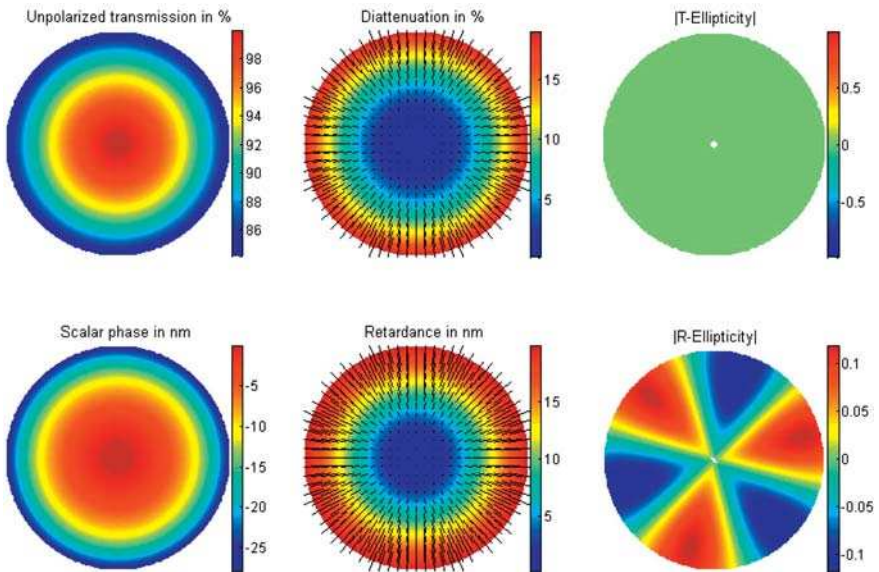


Figure 28-20: SVD evaluation of Jones pupil. Explanation in text.

28.10

Jones Matrix Pupils in the Polarization Matrix Calculus

Generally, electromagnetic waves are neither completely polarized nor completely unpolarized: they are partially polarized. This means that the components of the wave's Jones vector are not completely correlated. The coherency matrix calculus is a convenient way to describe this partial correlation.

If an input field with a polarization matrix \mathbf{P}_{in} traverses two points of the pupil described by two Jones matrices \mathbf{J}_1 and \mathbf{J}_2 , the output polarization matrix is obtained by [28-16]

$$\mathbf{P}_{\text{out}} = (\mathbf{J}_1 + \mathbf{J}_2)\mathbf{P}_{\text{in}}(\mathbf{J}_1 + \mathbf{J}_2)^+ \quad (28-61)$$

with the intensity

$$I = \text{trace}\{(\mathbf{J}_1 + \mathbf{J}_2)\mathbf{P}_{\text{in}}(\mathbf{J}_1 + \mathbf{J}_2)^+\}. \quad (28-62)$$

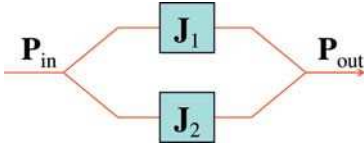


Figure 28-21: Two-beam interference in the polarization matrix calculus.

Resolving this expression yields

$$\begin{aligned} I &= \text{trace}\{\mathbf{J}_1\mathbf{P}_{\text{in}}\mathbf{J}_1^+\} + \text{trace}\{\mathbf{J}_2\mathbf{P}_{\text{in}}\mathbf{J}_2^+\} + 2 \text{Re}\{\text{trace}\{\mathbf{J}_1\mathbf{P}_{\text{in}}\mathbf{J}_2^+\}\} \\ &= I_0 + \text{Re}\{C\} \end{aligned} \quad (28-63)$$

with the mean intensity

$$I_0 = \text{trace}\{\mathbf{J}_1\mathbf{P}_{\text{in}}\mathbf{J}_1^+\} + \text{trace}\{\mathbf{J}_2\mathbf{P}_{\text{in}}\mathbf{J}_2^+\} \quad (28-64)$$

and the interference term

$$C = 2 \text{trace}\{\mathbf{J}_1\mathbf{P}_{\text{in}}\mathbf{J}_2^+\}. \quad (28-65)$$

In order to include vector effects, the polarization matrix has to be extended to three dimensions. To this end we write the polarization matrix formally as the dyadic product of the 3D electric field vector

$$\mathbf{P} = \langle \vec{\mathbf{E}} \circ \vec{\mathbf{E}}^+ \rangle = \begin{pmatrix} \langle E_x E_x^* \rangle & \langle E_x E_y^* \rangle & \langle E_x E_z^* \rangle \\ \langle E_y E_x^* \rangle & \langle E_y E_y^* \rangle & \langle E_y E_z^* \rangle \\ \langle E_z E_x^* \rangle & \langle E_z E_y^* \rangle & \langle E_z E_z^* \rangle \end{pmatrix}. \quad (28-66)$$

Such a 3D polarization matrix is obtained for a partial polarized field traversing two Jones matrices \mathbf{J}_1 and \mathbf{J}_2 , and rotated into the propagation direction corresponding to the pupil coordinates by application of the transformation matrix ψ

$$\mathbf{P}_{\text{out}} = (\psi_1\mathbf{J}_1 + \psi_2\mathbf{J}_2)\mathbf{P}_{\text{in}}(\psi_1\mathbf{J}_1 + \psi_2\mathbf{J}_2)^+. \quad (28-67)$$

The intensity becomes

$$I = \text{trace}\{\mathbf{P}_{\text{out}}\} = \text{trace}\{(\boldsymbol{\psi}_1\mathbf{J}_1 + \boldsymbol{\psi}_2\mathbf{J}_2)\mathbf{P}_{\text{in}}(\boldsymbol{\psi}_1\mathbf{J}_1 + \boldsymbol{\psi}_2\mathbf{J}_2)^+\} \quad (28-68)$$

and resolved

$$\begin{aligned} I &= \text{trace}\{\boldsymbol{\psi}_1^+\mathbf{J}_1\mathbf{P}_{\text{in}}(\boldsymbol{\psi}_1\mathbf{J}_1)^+\} + \text{trace}\{\boldsymbol{\psi}_2^+\mathbf{J}_2\mathbf{P}_{\text{in}}(\boldsymbol{\psi}_2\mathbf{J}_2)^+\} \\ &+ 2\text{Re}\{\text{trace}\{\boldsymbol{\psi}_1^+\mathbf{J}_1\mathbf{P}_{\text{in}}(\boldsymbol{\psi}_2\mathbf{J}_2)^+\}\} = I_0 + \text{Re}\{C\} \end{aligned} \quad (28-69)$$

with the mean intensity

$$I_0 = \text{trace}\{\boldsymbol{\psi}_1^+(\mathbf{J}_1\mathbf{P}_{\text{in}}\mathbf{J}_1^+)\boldsymbol{\psi}_1^+\} + \text{trace}\{\boldsymbol{\psi}_2^+(\mathbf{J}_2\mathbf{P}_{\text{in}}\mathbf{J}_2^+)\boldsymbol{\psi}_2^+\} \quad (28-70)$$

and the interference term

$$C = 2\text{trace}\{\boldsymbol{\psi}_1^+(\mathbf{J}_1\mathbf{P}_{\text{in}}\mathbf{J}_2^+)\boldsymbol{\psi}_2^+\}. \quad (28-71)$$

28.11

Jones-matrix-based System Optimization

A “perfect” pupil function in scalar optics is one without phase effects. Accordingly, a “perfect” Jones matrix pupil is one without polarization and phase effects, i.e., one that is proportional to the identity matrix. Therefore a valid optimization goal for polarization optics is to bring the Jones pupil for all field points as near as possible to the identity matrix. Mathematically we have to minimize the norm of their difference, i.e.,

$$\|\mathbf{J} - \mathbf{I}\| \rightarrow \min. \quad (28-72)$$

From this a simple strategy for the polarization optimization of compound optical systems can be derived: As a starting point the system’s Jones-matrix pupil is written as the product of the component Jones pupils

$$\mathbf{J}_{\text{system}}(p, q) = \prod_{i=1}^N \mathbf{J}_i(p, q). \quad (28-73)$$

The components’ Jones pupils must be combined to subgroups that are individually orthogonal, so that their combined effect cancels and the system’s Jones pupil becomes the identity matrix for each pupil point. As an example, consider a straight-forward compensation strategy for the spatial dispersion induced birefringence of CaF_2 (cf. chapter 26):

- a) Upper line in figure 28-22: The Jones pupil of a single plane plate of 111 CaF_2 has a three-fold symmetry. If we superpose two of these lens elements

- that are rotated with respect to each other about an angle of 60° , the combined Jones matrix has rotational symmetry with tangential birefringence.
- Lower line in figure 28-22: The Jones pupil of a single plane plate of 100 CaF_2 has a four-fold symmetry. If we superpose two of these elements that are rotated with respect to each other about an angle of 45° , the combined Jones matrix has rotational symmetry with radial birefringence.
 - Last column in figure 28-22 If the thickness of the lens elements in the 111 and 100 cut is of ratio $3/2$ then their combined effect cancels almost completely. The remaining retardance stems from the fact that the eigenpolarizations of the combined elements are not linear but are slightly elliptical.

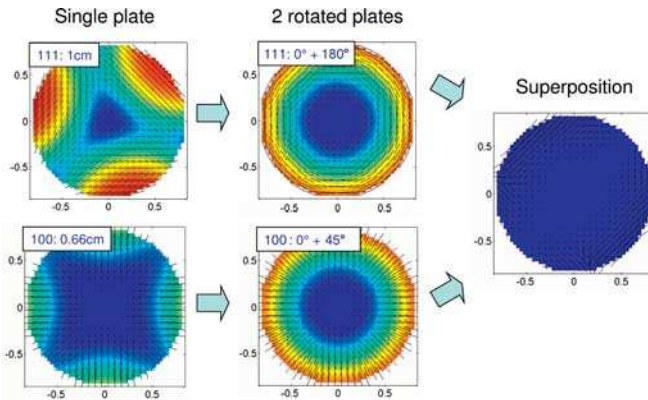


Figure 28-22: Optimization of spatial dispersion induced birefringence by clocking.

28.12 Aberrations of the Transmitted Wavefront

The characterization of transmitted wavefronts is an extremely successful concept for scalar optical systems. The aberrations of the wavefront define the imaging properties in detail. Furthermore, the wavefront data can be used to adjust actuators in order to optimize imaging properties after the assembly of the lens. So a very straightforward way of characterizing a lens for polarized illumination is to define a polarized wavefront and to characterize the wavefront errors and the polarization errors separately. However, because the phase difference of two polarized waves depends on the difference in the polarization states, a unique definition of a scalar phase, and hence a wavefront, is impossible. If we introduce a wavefront at all, we always have to balance the wavefront and polarization errors.

If the transmitted polarization state of the pupil can be considered as constant with small deviations it makes sense to define a scalar phase with respect to this polarization state. With the reference polarization \vec{E}_{ref} state as either the design polarization or the mean polarization

$$\vec{E}_{\text{ref}} = \frac{1}{\pi NA^2} \iint_{\text{pupil}} \vec{E}(p_x, p_y) dp_x dp_y, \quad (28-74)$$

the scalar phase for partially polarized illumination of degree g becomes

$$\varphi_{\text{sc}}(p_x, p_y) = \arg \left\{ \sqrt{\frac{1+g}{2}} \vec{E} \cdot \vec{E}_{\text{ref}}^* + \sqrt{\frac{1-g}{2}} \vec{E} \cdot \vec{E}_{\text{ref}}^* \right\}. \quad (28-75)$$

28.13

Jones–Zernike Wavefront Aberrations

In this section a modal method is derived for characterization of the imaging properties for a given Jones vector distribution in the exit pupil [28-17].

28.13.1

Principle of the Modal Characterization of a Jones Pupil

While scalar optical systems are characterized by an expansion of the scalar phase into a system of orthogonal functions (most frequently the Zernike polynomials) this approach is not sufficient for polarized systems. The idea of the following is to expand not the scalar wavefront alone, but the complex vector function into an appropriate set of elementary functions \vec{F}_i .

$$\vec{E}(p_x, p_y) = \sum_{i=1}^{\infty} Z_i \vec{F}_i(p_x, p_y). \quad (28-76)$$

with expansion coefficients Z_i . Because the complex vector field is expanded into a linear sum of elementary functions this approach is a modal one with the elementary functions being the modes.

The modal approach depends on the illumination polarization. It allows, however, a direct assessment of the transmitted field. Because of this it can be applied at the same time to the paraxial transmitted field as to the high-NA three-dimensional field in the image plane. For a characterization of the complete Jones pupil, the modal decomposition has to be applied for two orthogonal input polarizations.

28.13.2

Jones–Zernike Expansion

Because the expansion of the scalar wavefront in Zernike polynomials was very successful, the Zernike expansion presents itself as a simple and straightforward approach for a modal decomposition of the complex vector. The Jones–Zernike expansion is an expansion of the real and the imaginary part of the complex vector

field (in the pupil the 2D Jones vector) into Zernike polynomials U_j separately. In the paraxial case we get

$$\begin{aligned} \vec{E}(p_x, p_y) &= \begin{pmatrix} E_x^r(p_x, p_y) + iE_x^i(p_x, p_y) \\ E_y^r(p_x, p_y) + iE_y^i(p_x, p_y) \end{pmatrix} \\ &= \sum_{j=1}^{\infty} \left[Z_{jx} \begin{pmatrix} U_j(p_x, p_y) \\ 0 \end{pmatrix} + Z_{jy} \begin{pmatrix} 0 \\ U_j(p_x, p_y) \end{pmatrix} \right] = \sum_{j=1}^{\infty} U_j(p_x, p_y). \end{aligned} \quad (28-77)$$

The high-NA case is the same with \vec{E} having three vector components (after application of ψ). The scalar coefficients of the scalar Zernike expansion become vector coefficients for the vector case.

This decomposition takes place in the real and imaginary part and not in the amplitude and phase because the representation of the phase as a superposition of Zernike polynomials requires a constant amplitude which cannot be assumed for vector fields even for constant overall intensity: Even if $|E_x|^2 + |E_y|^2$ are constant, this is not necessarily the case for $|E_x|$ and $|E_y|$ alone.

The x - and y -component of the first four Jones–Zernike polynomials are sketched in figure 28-23 (numbering according to the fringe convention)

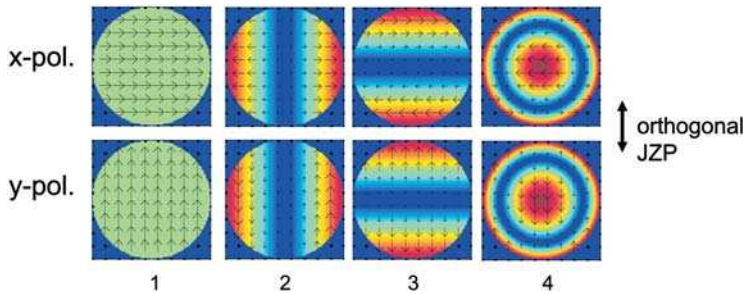


Figure 28-23: x - and y -component of Jones–Zernike polynomials 1–4.

As in the scalar case we can group polynomials with identical radial and orthogonal angular dependence (i.e., $\cos(m\varphi)$, $\sin(m\varphi)$).

28.13.3

Properties of the Jones–Zernike Polynomials

The Jones–Zernike coefficients have two useful basic properties:

1. They reduce to conventional-phase Zernike coefficients for weakly modulated scalar wavefronts. Because of

$$A(p_x, p_y) = A_0 e^{i\varphi(p_x, p_y)} \approx A_0 (1 + i\varphi(p_x, p_y)) \tag{28-78}$$

the complex coefficients become purely imaginary and equal to the phase expansion.

2. Their norm

$$\|\vec{Z}_i\| = \sqrt{\sum_{j=1}^3 |Z_i^j|^2} \tag{28-79}$$

is independent of the basis of the polarization state. The reason is that a transformation of the polarization basis is described by multiplication with a unitary matrix, and this does not change the norm.

3. Example: Jones–Zernike expansion of a radial birefringence with linear polarized illumination.

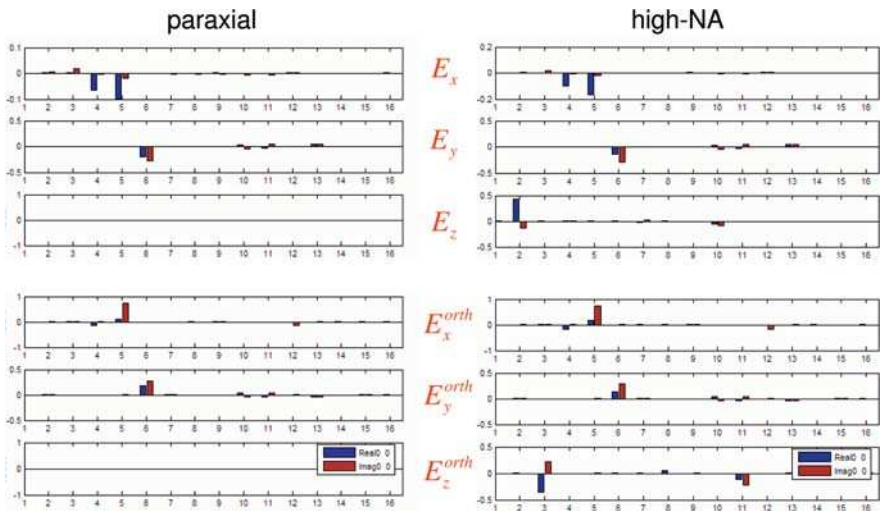


Figure 28-24: Jones–Zernike expansion of the pupil of figure 28-19. Left: paraxial field (no z-component), right: high-NA field (after application of transformation matrix to paraxial field). From top to bottom: design-polarization x, y, and z-component, orthogonal polarization x, y and z-component. Blue (left) real part. Red (right) imaginary part.

Figure 28-24 shows the Jones–Zernike expansion of the pupil of figure 28-19. On the left-hand side the expansion of the paraxial field is shown. The Z_4 -apodization is visible and the Z_5/Z_6 -component caused by the radial retardance. On the right-hand side the same expansion in the high-NA region ($n_{\text{resist}} = 1.8$) is shown. The z -component is excited for the p -field only and yields a Z_2 component for x -polarized incident field and a Z_3 component for y -polarized incidence.

28.14

Literature

- 28-1 O. Wiener, Stehende Lichtwellen und die Schwingungsrichtung polarisirten Lichtes, *Annalen der Physik* **40**, 203 (1890).
- 28-2 P. Drude, W. Nernst. Über die Fluoreszenzwirkungen stehender Lichtwellen. *Wied. Ann.*, **45**, 460–474 (1892).
- 28-3 W. Swindell, Ray-tracing formulas for uniaxial crystals, *J. Opt. Soc. Am.* **52**, 187–192 (1962).
- 28-4 J.D. Trolinger, R.A. Chipman, D.K. Wilson, Polarization ray tracing in birefringent media, *Opt. Eng.* **30**, 461–466 (1991).
- 28-5 C. Gu, P. Yeh, Extended Jones matrix method. II, *J. Opt. Soc. Am. A* **10**, 966–973 (1993).
- 28-6 S. C. McClain, L. W. Hillman, R.A. Chipman, Polarization ray tracing in anisotropic optically active media, *Opt. Soc. Am. A* **10**, 2383–2393 (1993).
- 28-7 I.J. Hodkinson, S. Kassam, Q.H. Wu, Eigen-equations and compact algorithms for bulk and layered anisotropic optical media: Reflection and refraction at a crystal-crystal interface, *J. Comp. Phys.* **133**, 75–83 (1997).
- 28-8 D. G. Flagello, T. Milster, A. E. Rosenbluth, Theory of high-NA imaging in homogeneous thin films, *J. Opt. Soc. Am. A* **13**, 58–64 (1996).
- 28-9 J.W. Goodman, *Introduction to Fourier Optics*, 2nd ed. (McGraw-Hill, New York, 1996).
- 28-10 W. Urbanczyk, Optical imaging in polarized light, *Optik* **63**, 25–35 (1982).
- 28-11 W. Urbanczyk, Optical imaging systems changing the state of light polarization, *Optik* **66**, 301–309 (1984).
- 28-12 R.A. Chipman, Polarization analysis of optical systems, *Opt. Eng.* **28**, 90–98 (1989).
- 28-13 J.P. McGuire, R.A. Chipman, Polarization aberrations. 1. Rotationally symmetric optical systems, *Appl. Opt.* **33**, 5080–5100 (1994).
- 28-14 Y. Unno, Distorted wave front produced by a high-resolution projection optical system having rotationally symmetric birefringence, *Appl. Opt.* **37**, 7241–7247 (1998).
- 28-15 S. Stallinga, Axial birefringence in high-numerical-aperture optical systems and the light distribution close to focus, *J. Opt. Soc. Am. A* **18**, 2846–2858 (2001).
- 28-16 E. Bernabeu, L.L. Sanchez-Soto, Coherency matrix analysis of polarization effects in two-beam interferometry, *Optik* **4**, 134–139 (1986).
- 28-17 M. Totzeck, P. Gräupner, T. Heil, A. Göhnermeier, O. Dittmann, D. Krähmer, V. Kamenov, J. Ruoff, D. Flagello, How to describe polarization influence on imaging, *SPIE Vol. 5754*, in press.

A1

Mathematical Appendix

- A.1 Linear Systems 629
- A.2 Fourier Series and Fourier Integral 631
 - A.2.1 Compilation of Basic Properties of the Fourier Transform 632
 - A.2.2 Special Functions and their Fourier Transforms 634
- A.3 Convolution and Correlation 637
 - A.3.1 Convolution 637
 - A.3.2 Correlation 637
 - A.3.3 Power Spectral Density and RMS Value 638
- A.4 Discrete Signals 639
 - A.4.1 The Sampling Theorem 639
 - A.4.2 Leakage 641
 - A.4.3 Indexing of the Numerical Discrete Fast Fourier Transform 642
- A.5 z-Transform 644
 - A.5.1 Definition 644
 - A.5.2 Numerical Evaluation of the z-transform 646
 - A.5.3 Sinc Interpolation 648
- A.6 Hankel Transform 648
 - A.6.1 Definition 648
 - A.6.2 Numerical Computation 649
- A.7 Practical Calculation of Diffraction Integrals 655
 - A.7.1 The Oscillation Problem 655
 - A.7.2 Spatial and Spectral Resolution 660
 - A.7.3 Periodic Boundary Conditions 662
 - A.7.4 x-z Sampling of the Ewald Sphere 663
 - A.7.5 Equivalent Diffraction Setups 663
 - A.7.6 Optimal Conditioning of the Fresnel Diffraction 666
 - A.7.7 Numerical Algorithms 669
 - A.7.8 Fresnel Integrals 672
- A.8 Orthogonal Polynomials on Rectangular Domains 675
 - A.8.1 Chebyshev Polynomials 675
 - A.8.2 One-dimensional Legendre Polynomials 677

A.8.3	Two-dimensional Chebyshev Polynomials	678
A.8.4	Legendre Polynomials in Two Dimensions	679
A.9	Literature	683

**A.1
Linear Systems**

An image formation is referred to as linear if it possesses the following properties: additivity, distributivity and associativity, i.e., the imaging of the object space O and the image space B onto each other is homomorphic. For such an image formation $F : O \rightarrow B$, the following relationship holds:

$$u = F(\lambda_1 \cdot v + \lambda_2 \cdot w) = \lambda_1 \cdot F(v) + \lambda_2 \cdot F(w) . \tag{A-1}$$

Hence linear image formation satisfies the principle of superposition.

For linear image formation, both object and image space can be represented by orthogonal basis. The image formation can be then traced back to a basis transformation. This will be illustrated below considering as an example the imaging of the object function $G(x)$ into the image function $G'(x')$. Due to the linearity the object function, $G(x)$ can be represented as composed of object points $\delta(x-x_0)$ of amplitude $G(x_0)$ (figure A-1):

$$G(x) = \int G(x_0) \delta(x - x_0) dx_0 . \tag{A-2}$$

The image of an object point is referred to as the impulse response $H(x', x'_0)$:

$$F[\delta(x - x_0)] = H(x', x'_0) . \tag{A-3}$$

Taking into account the principle of superposition one obtains for the image $G'(x')$

$$G'(x') = F[G(x)] = F\left[\int G(x_0)\delta(x - x_0)dx_0\right] = \int G(x'_0)H(x', x'_0)dx'_0 . \tag{A-4}$$

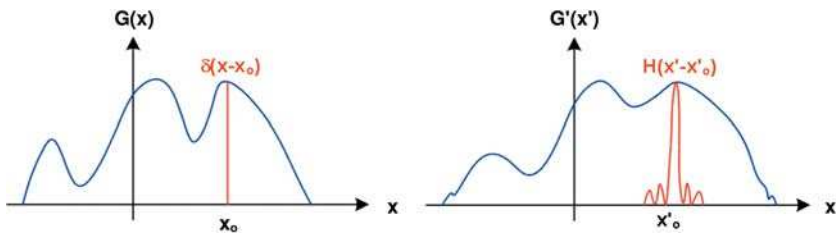


Figure A-1: Linear image formation: object (left) and image (right) after convolving all object points with a shift-invariant impulse response function $H(x)$.

The space invariance requires that the impulse response function is the same over space, hence:

$$H(x', x'_0) = H(x' - x'_0) . \tag{A-5}$$

Such linear systems, called linear shift invariant (LSI) systems, possess therefore the following convolution property (cf. figure A-1):

$$G'(x') = \int G(x'_0) \cdot H(x' - x'_0) dx'_0 = H(x') \otimes G(x') . \tag{A-6}$$

If an orthonormalized basis is chosen for the representation,

$$\begin{aligned} G(x) &= \sum_n g_n \cdot e_n(x) \\ G'(x') &= \sum_m g'_m \cdot e_m(x') \\ H(x') &= \sum_k h_k \cdot e_k(x') \end{aligned} \tag{A-7}$$

the existence of a transfer function can be inferred from the convolution property:

$$\begin{aligned} G'(x') &= \sum_n \sum_m g_n \cdot h_m \int e_n(x'_0) \cdot e_m(x' - x'_0) dx'_0 \\ &= \sum_n \sum_m g_n \cdot h_m \delta_{nm} e_n(x') = \sum_n g_n \cdot h_n \cdot e_n(x') \end{aligned} \tag{A-8}$$

thus

$$g'_n = h_n \cdot g_n . \tag{A-9}$$

A very important property of the linear image formation follows from the associativity of eq. (A-9) and the convolution: the sequence of elementary image formation steps in cascaded or composite image formations is permutable:

$$G'' = F_2[G'] = F_2[F_1[G]] = F_1[F_2[G]] . \tag{A-10}$$

Furthermore, the transfer function of cascaded image formation is obtained by multiplication of the transfer functions of the separate image formation steps:

$$G'' = H \otimes G = (H_2 \otimes H_1) \otimes G = H_2 \otimes (H_1 \otimes G) , \tag{A-11}$$

$$g'' = h \cdot g = h_2 \cdot h_1 \cdot g . \tag{A-12}$$

This property is illustrated in figure A-2 .

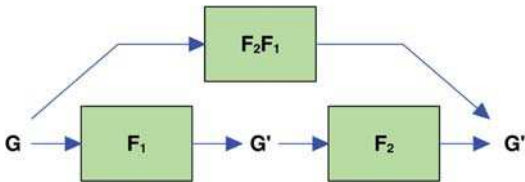


Figure A-2: Cascaded linear system consisting of two subsystems.

For this reason, the transfer function of a linear image formation can generally be decomposed into several separate transfer functions, which is advantageous if, e.g., various factors affecting the quality of the image formation have to be considered. A common example is the effect of the optical imaging system and the detector on the image formation when extracting a photographic and an electronic image, respectively. As well as the blurring of an image point through the point response function H_{optics} , the aerial image is also blurred by the finite granularity of the film used for recording or the fixed pixel size of the camera:

$$G'' = H \otimes G = (H_{detector} \otimes H_{optics}) \otimes G = H_{detector} \otimes (H_{optics} \otimes G). \quad (A-13)$$

Another example is the consideration of various error sources related to the optical system itself such as blurring of the point image as a result of design or alignment errors, or the spurious light resulting from errors in the surface roughness:

$$G'' = H \otimes G = (H_{scatter} \otimes H_{design}) \otimes G. \quad (A-14)$$

A.2

Fourier Series and Fourier Integral

In optics, and in physics generally, one often deals with wave phenomena. Hence the harmonic functions are widely used as orthogonal basis functions. Therefore it is important to present here the basic principles and the essential properties of the Fourier series expansion and the Fourier integral representation [A-1], [A-2].

Periodic functions can be represented by an infinite series expansion in harmonic functions (sine, cosine, or in complex form $e^{ix} = \cos(x) + i \cdot \sin(x)$). This is known as the Fourier series expansion:

$$U(x) = \sum_{m=-\infty}^{\infty} C_m \exp^{i2\pi \frac{mx}{d}}, \quad (A-15)$$

$$C_k = \frac{1}{d} \int_{-d/2}^{d/2} U(x) \cdot \exp^{-i2\pi \frac{kx}{d}} dx. \quad (A-16)$$

For the general case of non-periodic functions, the Fourier integral transform has to be used:

$$U(x) = \int u(v) e^{2\pi i x v} dv, \quad (A-17)$$

$$u(v) = \int U(x) e^{-2\pi i x v} dx. \quad (A-18)$$

In computers or experimental measurements one normally deals with discrete data. This leads to the implementation of the discrete Fourier transform (DFT) which reads in its one-dimensional form as:

$$U(x_k) = \sum_{m=-N/2}^{N/2-1} u(v_m) \cdot e^{i2\pi x_k \cdot v_m}, \quad (\text{A-19})$$

$$u(v_m) = \frac{1}{N} \sum_{k=-N/2}^{N/2-1} U(x_k) \cdot e^{-i2\pi x_k \cdot v_m}. \quad (\text{A-20})$$

A.2.1

Compilation of Basic Properties of the Fourier Transform

Linearity:

The Fourier transform $F[U]$ of a function U is linear.

$$F[a \cdot U(x) + b \cdot V(x)] = a \cdot F[U(x)] + b \cdot F[V(x)]. \quad (\text{A-21})$$

Scaling (or Similarity):

If a function is 'stretched' by a factor a , its Fourier transform is 'squeezed' by the same factor in the frequency domain:

$$U(ax) \Leftrightarrow \frac{1}{|a|} u\left(\frac{v}{a}\right), \quad (\text{A-22})$$

$$\frac{1}{|a|} U\left(\frac{x}{a}\right) \Leftrightarrow u(av). \quad (\text{A-23})$$

Symmetric \Rightarrow Real Spectrum:

Symmetric functions can be expanded in cosine functions only. Therefore their spectrum is real.

$$U(x) = U(-x) \Leftrightarrow \text{Im}[u(v)] = 0. \quad (\text{A-24})$$

Anti-symmetric \Rightarrow Imaginary Spectrum:

On the contrary, anti-symmetric functions exhibit an imaginary spectrum.

$$U(x) = -U(-x) \Leftrightarrow \text{Re}[u(v)] = 0. \quad (\text{A-25})$$

Conjugate:

The Fourier transform of the complex conjugate function, is reflected around the origin in the frequency domain.

$$U^*(x) \Leftrightarrow u^*(-v). \quad (\text{A-26})$$

Area:

The area under a function or its integral, correspond to the constant component or DC-term (value at the origin) of its spectrum.

$$u(0) = \int U(x) dx . \quad (\text{A-27})$$

Shifting:

If a function is shifted, its spectrum is modulated by a phase factor. Modulation of a function by a harmonic function with frequency ν , corresponding to the carrier frequency, shifts the spectrum by an amount ν .

$$U(x - x_0) \Leftrightarrow u(\nu) \cdot e^{-i2\pi\nu x_0} , \quad (\text{A-28})$$

$$U(x) \cdot e^{i2\pi\nu_0 x} \Leftrightarrow u(\nu - \nu_0) . \quad (\text{A-29})$$

Differentiation:

Differentiation of a function corresponds to multiplication of its Fourier transform by a factor proportional to the frequency.

$$\frac{d}{dx} U(x) \Leftrightarrow 2\pi i \nu u(\nu) . \quad (\text{A-30})$$

Integration:

Integration of a function corresponds similarly to dividing its Fourier transform by a factor proportional to the frequency.

$$\int U(x) dx \Leftrightarrow \frac{u(\nu)}{2\pi i \nu} . \quad (\text{A-31})$$

Moment theorem

If $U(x)$ and $u(\nu)$ denote a function and its Fourier transform

$$u(\nu) = \int U(x) \cdot e^{2\pi i \nu x} dx \quad (\text{A-32})$$

the n th derivative with respect to ν is given by

$$\frac{d^n u(\nu)}{d\nu^n} = \int (2\pi \cdot ix)^n U(x) \cdot e^{2\pi i \nu x} dx . \quad (\text{A-33})$$

For $\nu=0$, one obtains the moment theorem which establishes a relationship between the n th moment of a function and the n th derivative of its Fourier transform at the origin $\nu=0$

$$M_n = \int x^n \cdot U(x) dx = \frac{1}{(2\pi \cdot i)^n} \cdot u^{(n)}(0) . \quad (\text{A-34})$$

A.2.2

Special Functions and their Fourier Transforms

The main purpose of this section is to present in analytical form the Fourier transforms of several special functions commonly met in applications [A-1].

Delta function

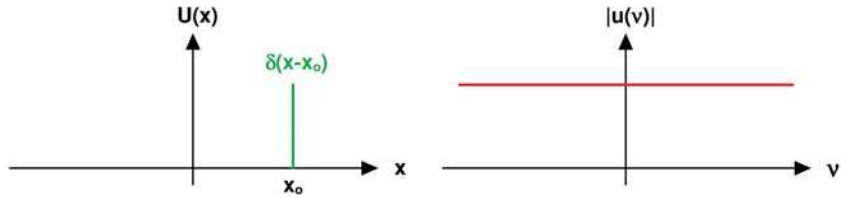


Figure A-3: Fourier transform of the delta function.

$$\delta(x - x_0) = \int e^{2\pi i v(x-x_0)} dv \Leftrightarrow e^{-i2\pi v x_0} . \tag{A-35}$$

The δ -function obeys the so-called sifting property:

$$\int U(x)\delta(x - x_0)dx = U(x_0) . \tag{A-36}$$

Comb function (delta comb)

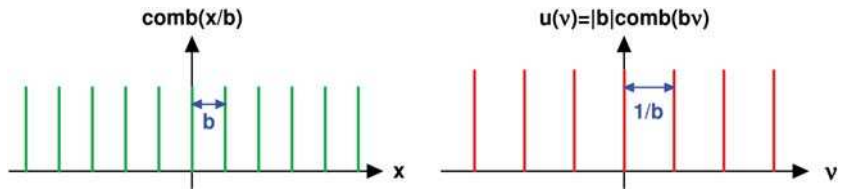


Figure A-4: Fourier transform of the comb function.

$$\text{comb}\left(\frac{x}{b}\right) = |b| \cdot \sum_n \delta(x - nb) = \sum_n e^{i2\pi n \frac{x}{b}} , \tag{A-37}$$

$$\text{comb}\left(\frac{x}{b}\right) \Leftrightarrow |b| \cdot \text{comb}(bv) . \tag{A-38}$$

Rect function and sinc function

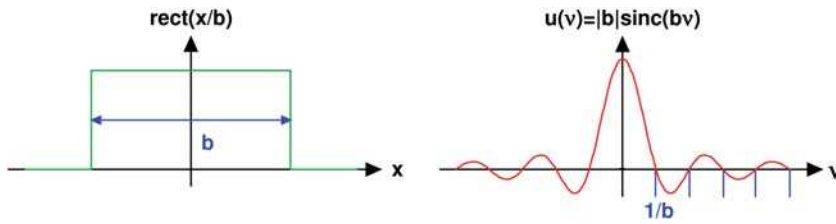


Figure A-5: Fourier transform of the rect function.

$$\text{rect}\left(\frac{x}{b}\right) = \begin{cases} 1 & |x| \leq b/2 \\ 0 & \text{else} \end{cases} \quad (\text{A-39})$$

$$\text{sinc}(x) = \frac{\sin(\pi x)}{\pi x} \quad (\text{A-40})$$

$$\text{rect}\left(\frac{x}{b}\right) \Leftrightarrow |b| \cdot \text{sinc}(bv) . \quad (\text{A-41})$$

Gaussian function



Figure A-6: Fourier transform of the Gaussian function.

$$\text{gauss}\left(\frac{x}{b}\right) = e^{-\frac{\pi x^2}{b^2}} \quad (\text{A-42})$$

$$\text{gauss}\left(\frac{x}{b}\right) \Leftrightarrow |b| \cdot \text{gauss}(bv) . \quad (\text{A-43})$$

In the particular case of a parabolic phase function, the Fourier transform is again a parabolic phase function:

$$e^{i\frac{\pi}{2}x^2} \Leftrightarrow iae^{-i\pi av^2} . \quad (\text{A-44})$$

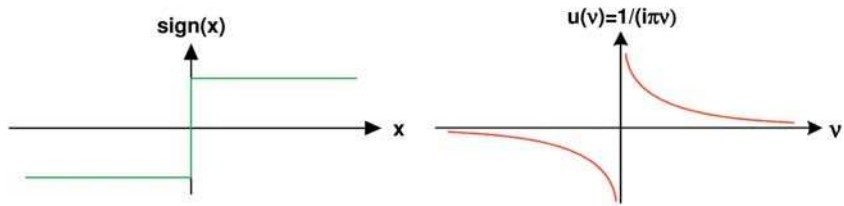
Sign function and $1/\nu$ 

Figure A-7: Fourier transform of the sign function.

$$\text{sign}(x) \Leftrightarrow \frac{1}{\pi i \nu} . \quad (\text{A-45})$$

The spectrum of the sign function $\text{sign}(x)$ extends to infinity. The Fourier transform of the truncated spectrum can be described by the sine integral function $\text{Si}(x)$

$$\text{Si}(x) = \int_0^x \frac{\sin t}{t} dt . \quad (\text{A-46})$$

Circ Bessel function

The Fourier transform of a circular aperture in polar coordinates is given by the first-order Bessel function of the first kind, divided by the frequency ρ :

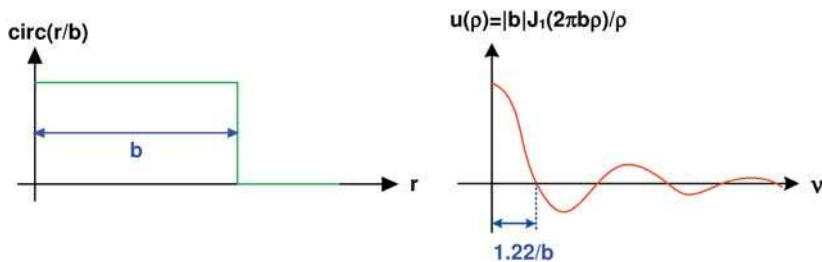


Figure A-8: Fourier transform of the circ function.

$$\text{circ}\left(\frac{r}{b}\right) \Leftrightarrow |b| \frac{J_1(2\pi b \rho)}{\rho} . \quad (\text{A-47})$$

A.3
Convolution and Correlation

A.3.1
Convolution

The product of the spectra of two functions in the frequency domain corresponds to convolution of the functions in the spatial domain [A-3]

$$U(x) \otimes V(x) = \int U(x - x') V(x') dx' \Leftrightarrow u(v) \cdot v(v) . \tag{A-48}$$

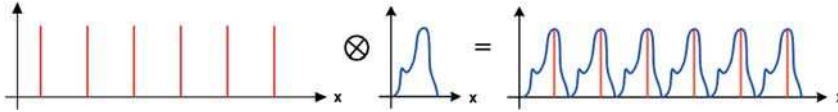


Figure A-9: Convolution of two functions.

As an example for convolution one can take the shifting property:

$$U(x - x_0) = U(x) \otimes \delta(x - x_0) = \int u(v) \cdot e^{-2\pi i v x_0} e^{2\pi i v x} dv \Leftrightarrow u(v) \cdot e^{-2\pi i v x_0} \tag{A-49}$$

and

$$U(x) \cdot e^{2\pi i v_0 x} = \int [u(v) \otimes \delta(v - v_0)] e^{2\pi i v x} dv \Leftrightarrow u(v - v_0) . \tag{A-50}$$

Convolution obeys the commutative, associative and distributive laws:

$$\begin{aligned} U \otimes V &= V \otimes U , \\ U \otimes (V \otimes W) &= (U \otimes V) \otimes W , \\ U \otimes (V + W) &= (U \otimes V) + (U \otimes W) . \end{aligned} \tag{A-51}$$

As already demonstrated, it plays a fundamental role in linear image formation.

A.3.2
Correlation

The correlation of two functions is given by a similar integration

$$U(x) \circ V(x) = \int U(x') V(x' + x) dx' . \tag{A-52}$$

It is not commutative:

$$\begin{aligned} W(x) &= U(x) \circ V(x) , \\ W(-x) &= V(x) \circ U(x) . \end{aligned} \tag{A-53}$$

The correlation can be represented as a convolution:

$$U(x) \circ V(x) = U(x) \otimes V(-x) . \tag{A-54}$$

The correlation function is commonly used for comparative analysis of two signals or images. An important case in particular is the autocorrelation:

$$U(x) \circ U^*(x) = \int U(x') U^*(x' + x) dx' . \tag{A-55}$$

The autocorrelation function is usually used in its normalized form, i.e., its maximum at the origin $x = 0$ is normalized to 1:

$$U(x) \circ U^*(x) = \frac{\int U(x') U^*(x' + x) dx'}{\int |U(x)|^2 dx} . \tag{A-56}$$

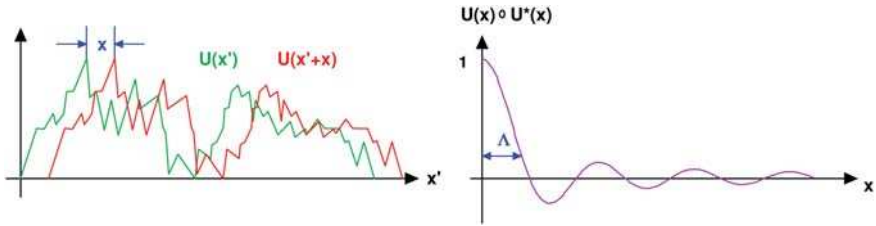


Figure A-10: Building of the autocorrelation of a function $U(x)$.

The correlation length corresponds to the distance λ at which the normalized autocorrelation drops to the value $1/e$. The Fourier transform of the autocorrelation corresponds to the power spectrum

$$U(x) \circ U^*(x) \Leftrightarrow |u(\nu)|^2 . \tag{A-57}$$

A.3.3

Power Spectral Density and RMS Value

The power spectral density (PSD) function of $h(x,y)$ is defined by

$$PSD(\nu_x, \nu_y) = \frac{\left| \iint_A h(x, y) e^{-i2\pi(\nu_x x + \nu_y y)} dx dy \right|^2}{A} = \frac{|h(\nu_x, \nu_y)|^2}{A} \tag{A-58}$$

where A is the surface area. The PSD is frequently used, e.g., to describe surface roughness errors. For statistical surfaces, the PSD has rotational symmetry (the ori-

entation of the different sinusoidal Fourier components on the surface is statistical). The RMS value σ of a frequency band is thus given by

$$\sigma = \sqrt{\int_{\nu_{\min}}^{\nu_{\max}} 2\pi\nu \text{PSD}(\nu) d\nu} \quad (\text{A-59})$$

with

$$\nu = \sqrt{\nu_x^2 + \nu_y^2}. \quad (\text{A-60})$$

A.4

Discrete Signals

A.4.1

The Sampling Theorem

The discrete nature of the data derived from measurements and the computer processing of such data require a detailed consideration of the discrete Fourier transform. The discrete sampling of a continuous function $U(x)$ replaces the integral transform by a discrete summation. The appropriate choice of the sampling interval and the truncation window (number of terms) is essential in order to avoid errors between the computed and the desired transform. The discrete sampling of a function $U(x)$ can be represented as [A-3]:

$$U_d(x) = U(x) \cdot \text{comb}\left(\frac{x}{dx}\right) \quad (\text{A-61})$$

where the sampling interval is introduced by the comb function

$$\text{comb}\left(\frac{x}{dx}\right) = \sum_n \delta(x - n \cdot dx). \quad (\text{A-62})$$

If the sampling interval is dx , the maximum frequency that can be resolved (truncation in the frequency domain) is given by the reciprocal value of dx :

$$\nu_{\max} = \frac{1}{dx}. \quad (\text{A-63})$$

The spectrum of the sampled function given by eq. (A-62) will be convolved in the frequency domain with a comb function:

$$u_d(\nu) = u(\nu) \otimes dx \cdot \text{comb}(dx \cdot \nu). \quad (\text{A-64})$$

Hence the sampling in space with an interval dx results in periodic repetition of the spectrum in the frequency domain with a period $\nu_{\max} = 1/dx$ (cf. Brillouin zones in solid state physics). The sampling interval should be small enough that the periodically repeating spectrum corresponding to the continuous function does not overlap in the frequency domain (aliasing effect):

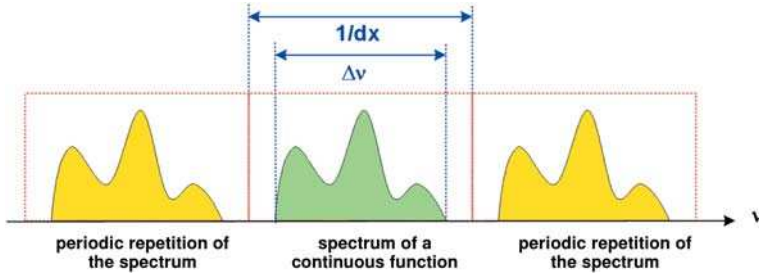


Figure A-11: Appropriate sampling interval.

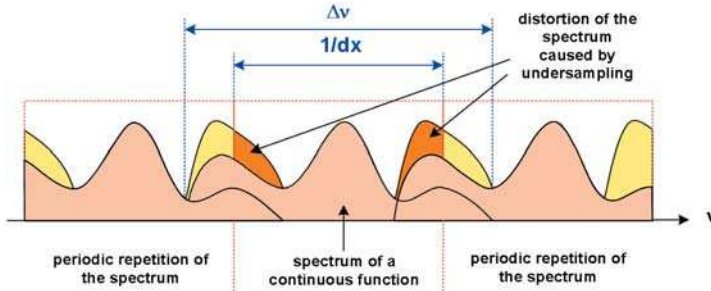


Figure A-12: Occurrence of aliasing for a sampling interval which is not fine enough.

In general this is expressed in distortion of the spectrum in the frequency domain, since due to the periodic repetition of the spectrum, high frequencies can interfere with low frequencies. Hence, the discrete sampling is applicable only to bandwidth-limited signals. In this case it is sufficient to select the sampling interval to be small enough in order to avoid such a cross-talk effect:

$$dx < \frac{1}{\Delta \nu} . \quad (\text{A-65})$$

$\Delta \nu$ in eq. (A-65) corresponds to the spectral bandwidth of the continuous function (cf. figure A-11). When a function exhibits higher frequencies the above-mentioned errors will occur (cf. figure A-12).

Let us take as an example a function $U(x)$ which is sampled over a window Δx with a finite number of points N . The use of the series expansion and hence the discrete Fourier transform implies periodicity of the function. Hence the actual signal function $S(x)$ is obtained by truncation with a rect function and convolution with a comb function:

$$S(x) = \left[U(x) \cdot \text{rect}\left(\frac{x}{\Delta x}\right) \cdot \text{comb}\left(\frac{x}{dx}\right) \right] \otimes \text{comb}\left(\frac{x}{\Delta x}\right). \quad (\text{A-66})$$

The period (window) Δx is chosen as N times the sampling interval, $\Delta x = N \cdot dx$. For the corresponding spectrum one obtains in the frequency domain:

$$s(\nu) = \{ [u(\nu) \otimes \text{sinc}(\Delta x \cdot \nu)] \otimes \text{comb}(dx \cdot \nu) \} \cdot \text{comb}(\Delta x \cdot \nu). \quad (\text{A-67})$$

The convolution with the first delta comb results in a periodic repetition of the spectrum. The multiplication with a rect function reduces the spectrum to the first Brillouin zone. This corresponds to a convolution of the sampled function in the spatial domain with a sinc function ($\text{sinc}(x/dx) = \text{sinc}(\Delta \nu \cdot x)$). Therefore the spectrum of a bandwidth-limited signal can be expressed as

$$s(\nu) = [u(\nu) \otimes \text{sinc}(\Delta x \cdot \nu)] \cdot \text{comb}(\Delta x \cdot \nu) \quad (\text{A-68})$$

where

$$dx = \frac{\Delta x}{N} = \frac{1}{\Delta \nu} \quad (\text{A-69})$$

and

$$d\nu = \frac{\Delta \nu}{N} = \frac{1}{\Delta x} \quad (\text{A-70})$$

hold. This coupling between the spacings in the space and the frequency domain given by eq. (A-69) and (A-70) respectively is another formulation of the sampling theorem.

A.4.2

Leakage

Consider the sampling of a sine function at N points with a window Δx , a multiple of the period $d = \Delta x/4$. Since $d\nu = 1/\Delta x$ one would expect the frequencies of the sine signal to occur at $\pm 1/d = \pm 4/\Delta x = \pm 4 d\nu$, i.e., at $m = \pm 4$.

It follows from (A-69) that:

$$s(\nu) = [\delta(\pm 4d\nu) \otimes \text{sinc}(\Delta x \cdot \nu)] \cdot \text{comb}(\Delta x \cdot \nu) \quad (\text{A-71})$$

i.e., when the sampling window is matched the discrete Fourier transform of a sine function yields only two frequency components in the spectrum (see figure A-13).

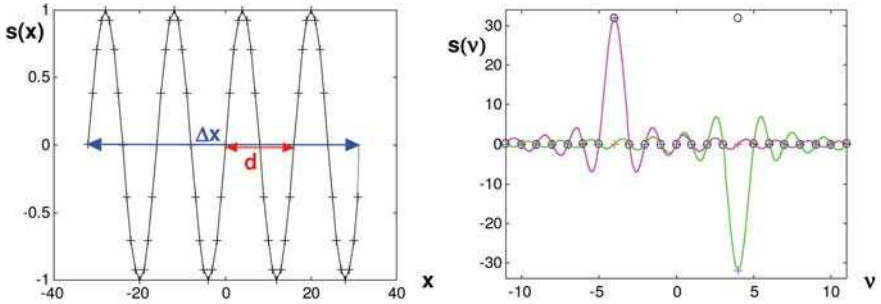


Figure A-13: Sampling interval equal to an integer multiple of the signal period.

Consider now the case when the same sine signal is sampled with a window Δx which is not an integer multiple of the period but equals, e.g., 3.8 periods. From the imaginary part of the spectrum shown in figure A-14, it can be seen that the mismatch between the sampling window and the period results in noticeable oscillations accompanying the peaks.

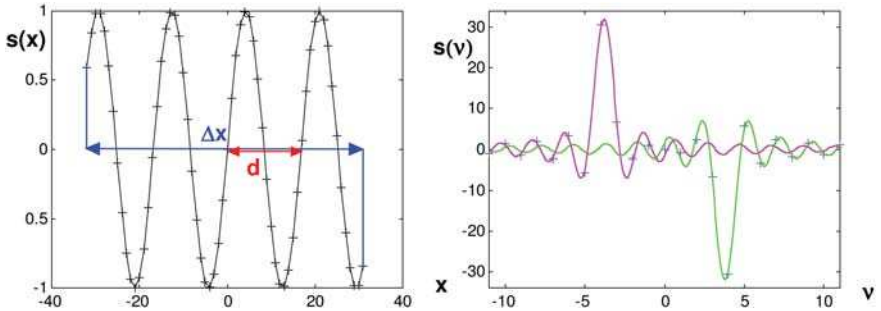


Figure A-14: Sampling interval not equal to an integer of the signal period.

$$s(\nu) = [\delta(\pm 3.8d\nu) \otimes \text{sinc}(\Delta x \cdot \nu)] \cdot \text{comb}(\Delta x \cdot \nu) . \quad (\text{A-72})$$

The exact frequencies must be evaluated by fitting with a sinc function. This phenomenon is known as leakage. Leakage has to be taken into account, e.g., in the phase-shift interferometry where the wave phase is measured. In this case either a large number of phase measurements or a precisely aligned phase step is required.

A.4.3

Indexing of the Numerical Discrete Fast Fourier Transform

According to the Cooley–Tuckey algorithm, the discrete sampling in the fast Fourier transform (FFT) is asymmetric since it operates with an even number of points, one of which corresponds to the center of the sampled range [A-4], [A-5]. Hence, if the total number of points is N , the center is normally set at $j = N/2 + 1$, and the points

$j = 2$ and $j = N$ are then symmetric. An additional index $j = 1$ is allocated on the left side. This indexing is illustrated in figure A-15.

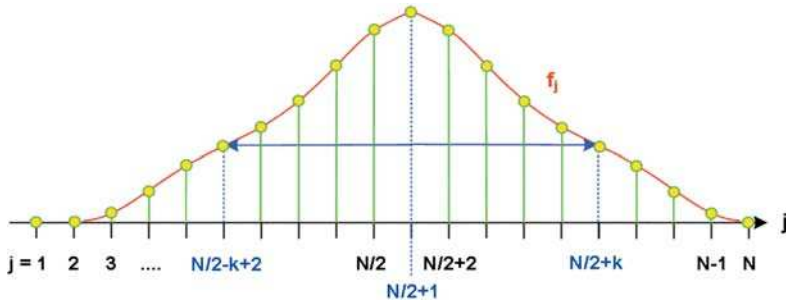


Figure A-15: Asymmetric discrete sampling in the numerical fast Fourier transform in the spatial domain.

In the frequency domain the off-set is found at frequency $\nu = 0$, i.e., index $j = 1$, the point $j = 2$ corresponds again to $j = N$, the maximum (positive) frequency is at $j = N/2$ and $j = N/2 + 2$ corresponds to the minimum (negative) frequency. This indexing in the spectral domain directly after the fast transform is illustrated in figure A-16.

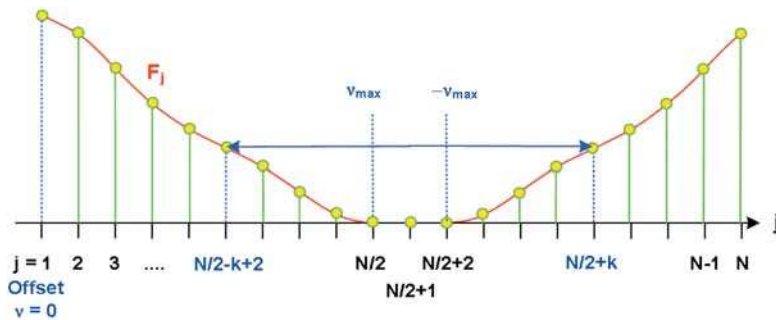


Figure A-16: Asymmetric discrete sampling in the numerical fast Fourier transform in the spectral domain. To get the right order of the spectral components, a shifting operation has to be done.

The spectrum can be presented on a continuous frequency scale by transposition of the negative frequency points.

In the two-dimensional case the corresponding frequency components are lying along the edges of the quadrant. Figure A-17 illustrates the sequence of the frequencies.

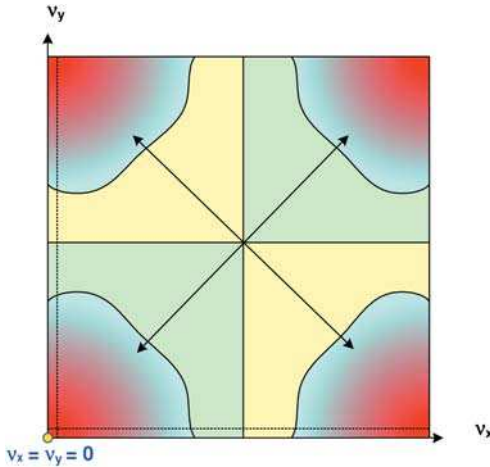


Figure A-17: Asymmetric discrete sampling in the numerical fast Fourier transform presented here in the two-dimensional frequency domain.

A.5 z-Transform

A.5.1

Definition

The problematic fixed coupling of the spatial and spectral resolution in the numerical discrete Fourier transform, dictated by the sampling theorem (see Vol. 1, section 8.4.2), can be avoided by applying the z -transform [A-6].

The general discrete z -transform of the function f , for given discrete data $f(x_j) = f_j$ with $j = 0, 1, 2, \dots, N-1$ and a likewise discrete sequence of numbers z_k with $k = 0, 1, 2, \dots, M$, reads

$$F_k = F(z_k) = \sum_{n=0}^{N-1} f_n z_k^{-n}. \quad (\text{A-73})$$

In accordance with its application in signal processing, this algorithm is named the chirp z -transform (CZT).

Special cases of the general transform are:

1. In the particular case when z is a real exponential function, the z -transform is equivalent to the well-known Laplace transform.

$$z_k = e^{s^T}. \quad (\text{A-74})$$

2. For a set of points (function z) equally spaced around the unit circle

$$z_k = e^{\frac{2\pi i k}{N}} \quad (\text{A-75})$$

where $N = M$, one obtains the discrete Fourier transform.

For a more general form of the z -function

$$z_k = A W^{-k} \quad (\text{A-76})$$

and $|W| \neq 1$ a spiral contour in the complex z -plane has to be evaluated. In this case the CZT can be efficiently computed by a convolution using an algorithm similar to that of the fast Fourier transform. Consider

$$A = A_0 \cdot e^{2\pi i \theta_0} \quad (\text{A-77})$$

and

$$W = W_0 \cdot e^{2\pi i \phi_0} \quad (\text{A-78})$$

where $2\pi\phi_0$ is the angular spacing on the contour with respect to the origin. Insertion of the z -series into the expression for the transform yields

$$F_k = \sum_{n=0}^{N-1} f_n A^{-n} W^{nk} . \quad (\text{A-79})$$

The classical DFT implies that the function is periodic. In the complex z -plane this corresponds to sampling at equidistantly spaced points around the unit circle with a frequency separation of $2\pi/N$ (see figure A-18).

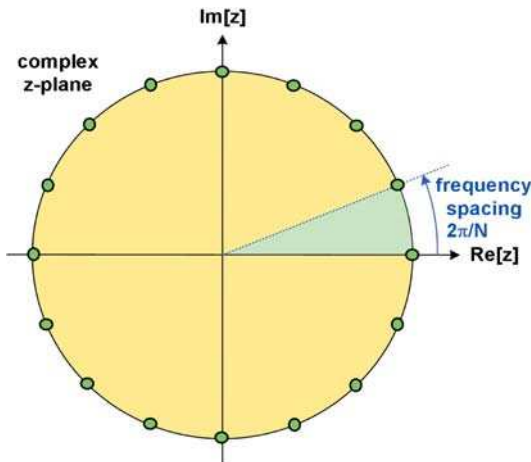


Figure A-18: Sampling points around the unit circle for the classical Fourier transform.

In the case of the chirp z-transform, this sampling is confined to a narrow frequency interval. In the general case a spiral contour occurs. If the frequency sampling is applied without filtering, a frequency zoom appears, as the sampling diagram shown in figure A-19.

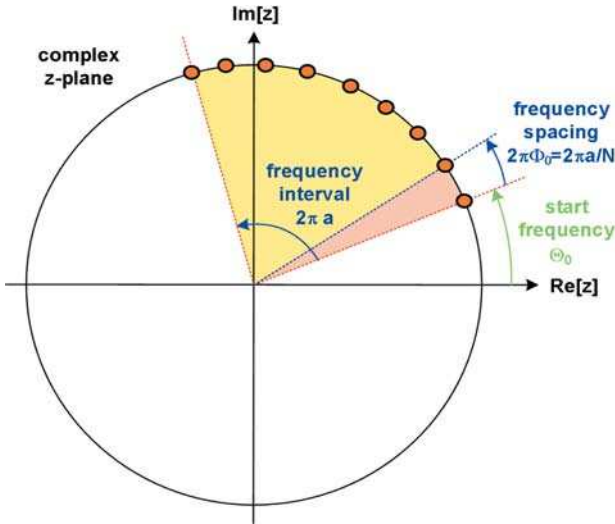


Figure A-19: Sampling points around the unit circle in the case of the chirp z-transform, the zooming factor a is explained in eq. (A-85)

A.5.2

Numerical Evaluation of the z-transform

With the help of the modified binomial formula [A-7]

$$nk = \frac{1}{2} [n^2 + k^2 - (k - n)^2] \tag{A-80}$$

the evaluation of the z-transform is reduced to an algorithm due to Bluestein

$$F_k = W^{\frac{k^2}{2}} \sum_{n=0}^{N-1} [f_n A^{-n} W^{\frac{n^2}{2}}] W^{-\frac{(n-k)^2}{2}} = W^{\frac{k^2}{2}} [f_n A^{-n} W^{\frac{n^2}{2}}] \otimes W^{-\frac{n^2}{2}}. \tag{A-81}$$

The computation of the above expression requires the following steps:

1. Weighting or multiplication of the given series f_n by $A^{-n} W^{\frac{n^2}{2}}$.
2. Fourier transform of the function obtained.
3. Forming a function $W^{-\frac{n^2}{2}}$.
4. Fourier transform of this function.
5. Multiplication of the Fourier series in order to compute the convolution.
6. Inverse transform of the product and multiplication by $W^{\frac{k^2}{2}}$.

The computation of the z-transform is hence reduced to a convolution, which can be evaluated by a fast Fourier transform algorithm. The additional computational steps required result in increased complexity in comparison to the simple Fourier transform. The computational time is increased about three times but the exact factor depends on the choice of the number of points N , M and the selected FFT algorithm. The advantages of the z-transform over the Fourier transform are related basically to two properties:

1. The desired number of points M does not have to coincide with the given number of points N . Hence one can save computer time if it is necessary to calculate only a few points in the frequency domain.
2. Since W itself is a complex number, it can be represented in general in the form

$$W = W_0 e^{2\pi i \Phi_0} . \quad (\text{A-82})$$

In the case of the Fourier transform

$$\Phi_0 = \frac{1}{N} \quad (\text{A-83})$$

and the points occupy exactly one angular circuit on the unit circle. As a consequence of this property the frequency values ν of the result of the transform are fixed by the sampling theorem in the form

$$\Delta x \cdot \Delta \nu = 1 / N . \quad (\text{A-84})$$

If the phase of W in the CZT is reduced by multiplication with a factor $a < 1$

$$\Phi_0 = a / N \quad (\text{A-85})$$

this results in sampling in the frequency domain with a step

$$\Delta \nu = \frac{a}{\Delta x \cdot N} . \quad (\text{A-86})$$

Therefore, the choice of the resolution in the frequency domain is not restricted. In particular one can compute the spectrum on a specified grid and avoid any interpolation. This is of special importance for optical transfer computations using Fourier algorithms where only fields within the free apertures need to be considered [A-8], [A-9]. In addition, the advantages of the z-transform are significant when sequential transforms have to be performed. This is the case, e.g., in optical transfer problems related to cavity calculations or beam propagation in inhomogeneous media.

A.5.3

Sinc Interpolation

The change in the number of points of a sampled signal or their spacing can be carried out by interpolation. According to the sampling theorem the best method for interpolating a continuous signal is by the sinc function. However, this method can be applied only to equidistantly sampled functions.

For a bandwidth-limited signal with a spectral width $1 / 2\Delta x$ the sinc interpolation is exact in the sense of information theory. The sinc interpolation exhibits similar advantages for a finite number of sampling points. Such a discrete sinc interpolation obeys the formula

$$a_{\text{resample}}(x) = \sum_{k=0}^{N-1} a_k \cdot \frac{\sin \left[\pi \cdot \left(\frac{x}{\Delta x} - k \right) \right]}{N \cdot \sin \left[\frac{\pi}{N} \cdot \left(\frac{x}{\Delta x} - k \right) \right]}. \quad (\text{A-87})$$

In practice this means that the given function is subjected to a Fourier transform, and the result, after zero-padding, is inverse-transformed back to the spatial domain. The limitation of this method is that the zero-padding can be carried out only by integer values so that the rescaling occurs in discrete steps only.

That is why the application of the chirp-z transform is more favourable, in which an arbitrary scale transformation can be achieved by performing the operation with different rational chirp factors a_1 and a_2 , respectively.

A.6**Hankel Transform**

A.6.1

Definition

For problems with rotational symmetry, the formulation of the Fourier transform in cartesian coordinates is not adequate. The so-called Hankel transform arises naturally in this case and is computationally advantageous since it utilizes the given symmetry. If the two-dimensional Fourier transform is presented in polar coordinates [A-3]

$$\begin{aligned} x &= r \cdot \cos \theta, & y &= r \cdot \sin \theta, \\ v_x &= u \cdot \cos \varphi, & v_y &= u \cdot \sin \varphi, \end{aligned} \quad (\text{A-88})$$

taking into account the definition of the Bessel function in the form

$$J_n(z) = \frac{1}{2\pi} \int_0^{2\pi} e^{i(n\beta - z \sin \beta)} d\beta \quad (\text{A-89})$$

one arrives at the Hankel transform in a representation which corresponds to the Fourier transform in cylindrical coordinates

$$F(u, \varphi) = \sum_{n=-\infty}^{\infty} (-i)^n e^{in\varphi} \int_0^{\infty} \int_0^{2\pi} f(r, \theta) e^{-in\theta} J_n(2\pi r u) r dr d\theta . \quad (\text{A-90})$$

If there is no dependence on the azimuthal angle, one obtains from it the simplified form of the Hankel transform for axisymmetric cases

$$F(u) = 2\pi \int_0^{\infty} f(r) J_0(2\pi r u) r dr . \quad (\text{A-91})$$

In the discrete form of the Hankel transform the radius is sampled with a radial interval

$$\Delta r = \frac{R}{N-1} \quad (\text{A-92})$$

and discrete points

$$r_n = n \Delta r, \quad n = 0, 1, 2, \dots, N-1, \quad (\text{A-93})$$

where R is the maximum radius and N is the total number of points. Similarly, in the frequency domain

$$\Delta u = \frac{N-1}{2RN}, \quad (\text{A-94})$$

$$u_n = n \Delta u, \quad n = 0, 1, 2, \dots, N-1, \quad (\text{A-95})$$

holds and the sampling theorem reads

$$\Delta r \cdot \Delta u = \frac{1}{2N} .$$

A.6.2

Numerical Computation

Different approaches for the numerical computation of the Hankel transforms can be found in the corresponding literature. The basic ideas and algorithms are given below:

1. Gardner transform

Using an exponential sampling of the coordinates, the so-called Gardner transform, the integral of the axisymmetric Hankel transform can be converted into a convolution integral [A-10] and hence numerically computed by the FFT methods. If the spatial and frequency coordinates are exponentially scaled, the product in the argument of the Bessel function can be written as a sum in the exponent and the integral takes the form of a correlation integral. One disadvantage of this method is the ex-

tremely non-equidistant sampling which does not allow a point at $r=0$: one obtains a non-equidistant grid with increasing discretization errors in the vicinity of the origin. If the argument and the coordinate in the axisymmetric Hankel transform

$$F(u) = 2\pi \int_0^{\infty} f(r) J_0(2\pi r u) r dr \quad (\text{A-96})$$

are substituted by

$$r = r_o \cdot e^{\alpha x}, \quad u = u_o \cdot e^{\alpha y} \quad (\text{A-97})$$

using the auxiliary functions defined as

$$\bar{F}(y) = u \cdot F(u), \quad (\text{A-98})$$

$$\bar{f}(y) = r \cdot f(r), \quad (\text{A-99})$$

$$\bar{J}(z) = 2\pi \cdot \alpha \cdot r_o u_o \cdot e^{\alpha z} \cdot J_0(2\pi \cdot r_o u_o \cdot e^{\alpha z}), \quad (\text{A-100})$$

one obtains the Hankel transform in the form of a correlation integral

$$\bar{F}(y) = \int_{-\infty}^{\infty} \bar{f}(x) \cdot \bar{J}(x+y) dx \quad (\text{A-101})$$

which can be numerically evaluated with high computational efficiency by a fast Fourier transform. This is possible if only x and y are equidistantly sampled which automatically means that this is not the case for r and u . Since the point $r=0$ in particular does not appear in the sampling, the integration is actually carried out starting from some minimum value of the radius r . In order to correct the resulting errors, in a first approximation, a correction term which takes into account the integration from $r=0$ to $r=r_o$ has to be provided.

The parameters r_o , u_o and α used in the formulation can be arbitrarily chosen and can be optimized in such a way that the sampling both in the spatial and in the frequency domain make it meaningful.

A specific problem of this numerical computation of the Hankel transform is the exponentially non-equidistant sampling, which leads to a decreasing number of points for increasing spatial and frequency coordinates.

2. Dual algorithm

For small values of the argument the Bessel function can be either numerically integrated or expanded in a Taylor series, while for large values of the argument one can use asymptotic approximations [A-11]. One limitation of this approach is related to problems at the boundary between these two ranges. The boundary has to be automatically found by an adaptive procedure. In practice one obtains small jump discontinuities and the solution in the asymptotic range is superposed by long-wave ripples.

3. Projection algorithm

The function $f(r)$, which has to be transformed, is extended antisymmetrically for $r < 0$. With the following representation of the Bessel function

$$J_0(z) = \frac{1}{\pi} \int_{-1}^1 \frac{e^{izt}}{\sqrt{1-t^2}} dt \tag{A-102}$$

a projection integral is obtained as a result of the transform, which can then be computed with high efficiency by a FFT [A-12]. One limitation of this method is the singularity at $r=0$ which always leads to increasing inaccuracy at small radii. This range, however, carries the most essential information in the case of optical beam propagation and is characterized by the highest intensity.

4. Direct integration with linear accuracy

When high precision is required for the application of the Hankel transform in optical propagation algorithms it is more reasonable to abstain from fast procedures possessing similar computational efficiency to the FFT which, according to the above mentioned algorithms 1 and 3, is determined by $N \log N$ and to rely on the classical approach, the computational time of which is proportional to N^2 , maintaining higher precision. The deviation of a twice-transformed function from its original can be used as a criterion for sufficient precision of the numerical computation of the transform. Such a test is very sensitive to the above-mentioned jump discontinuities, the inaccuracy in the vicinity of the origin and the long-wave ripples. In this sense the precision criteria in the form of absolute errors, normally used in the corresponding literature when dealing with single transforms, are not adequate for evaluation of optical algorithms employing sequential transforms.

The procedure given below describes an efficient and precise numerical approach for the computation of the Hankel transform.

1. Using the equidistant discrete radial grid, the integral of the Hankel transform is represented as a sum of the contributions from all subintervals:

$$F(u) = \sum_{n=0}^{N-1} 2\pi \int_{r_n}^{r_{n+1}} f(r) J_0(2\pi r u) r dr \tag{A-103}$$

2. Assuming that for sufficiently fine sampling $f(r)$ changes only weakly within each interval, it can be approximated by an average value

$$\bar{f}(r_n) = \frac{1}{2} [f(r_{n+1}) + f(r_n)] \tag{A-104}$$

giving a constant in front of the integral:

$$F(u) = \sum_{n=0}^{N-1} 2\pi \bar{f}(r_n) \int_{r_n}^{r_{n+1}} J_0(2\pi r u) r dr \tag{A-105}$$

- The remaining integral over the Bessel function can then be evaluated using the formulae

$$\begin{aligned} \int J_0(x) x dx &= x J_1(x), \\ \int J_0(rx) r dr &= \frac{r}{x} J_1(rx). \end{aligned} \quad (\text{A-106})$$

5. Direct integration with quadratic precision

For even higher requirements to the precision a parabolic approximation for the function $f(r)$ in each sub-interval can be used in a second step. The resulting integrals over the Bessel functions are then more complex but can be still reduced to well-known functions. The practical realization of the parabolic approximation follows the steps given below:

- The integral of the Hankel transform is decomposed into contributions from all sub-intervals. The parabolic approximation of the function $f(r)$ is applied to each three points $r_n \dots r_{n+1} \dots r_{n+2}$

$$f(r) = Ar^2 + Br + C \quad (\text{A-107})$$

where the parameters of the parabola are given by

$$A = \frac{f_{n+2} + f_n - 2f_{n+1}}{2 \Delta r^2}, \quad (\text{A-108})$$

$$B = \frac{f_{n+2} - f_n}{2 \Delta r} - 2Ar_{n+1}, \quad (\text{A-109})$$

$$C = f_{n+1} - Ar_{n+1}^2 - Br_{n+1}. \quad (\text{A-110})$$

The three points at the two boundaries have to be chosen asymmetrically in each case.

- Substitution of this parabolic representation in the interval $r_n \dots r_{n+1}$, together with the integral relationships

$$I_0(x) = \int J_0(x) dx = x J_0(x) + \frac{\pi x}{2} [J_1(x) H_0(x) - J_0(x) H_1(x)], \quad (\text{A-111})$$

$$\int x^2 J_0(x) dx = x^2 J_1(x) + x J_0(x) - I_0(x), \quad (\text{A-112})$$

$$\int x^3 J_0(x) dx = x^3 J_1(x) + 2x^2 J_0(x) - 4x J_1(x) \quad (\text{A-113})$$

yields

$$\begin{aligned}
 F(u) = & \frac{1}{u} \cdot \sum_{n=0}^{N-1} \left[r_{n+1} J_1(2\pi u r_{n+1}) \cdot \left(C - \frac{4A}{(2\pi u)^2} + Br_{n+1} + Ar_{n+1}^2 \right) \right. \\
 & + r_{n+1} J_0(2\pi u r_{n+1}) \cdot \left(\frac{B}{2\pi u} + r_{n+1} \frac{2A}{2\pi u} \right) + I_0(2\pi u r_{n+1}) \cdot \frac{-B}{(2\pi u)^2} \\
 & - r_n J_1(2\pi u r_n) \cdot \left(C - \frac{4A}{(2\pi u)^2} + Br_n + Ar_n^2 \right) \\
 & \left. - r_n J_0(2\pi u r_n) \cdot \left(\frac{B}{2\pi u} + r_n \frac{2A}{2\pi u} \right) + I_0(2\pi u r_n) \cdot \frac{-B}{(2\pi u)^2} \right]. \quad (\text{A-114})
 \end{aligned}$$

In the above equation H_m denotes the Struve functions which are defined as

$$H_m(x) = \frac{2 \left(\frac{x}{2}\right)^m}{\sqrt{\pi} \Gamma(m + 1/2)} \int_0^{\frac{\pi}{2}} \sin(x \cos t) \sin^{2m} t \, dt. \quad (\text{A-115})$$

High-accuracy Padé approximations, which can be fast-evaluated, exist both for the Bessel and the Struve functions.

3. The above formula is not applicable in the particular case $u = 0$. Since $J_0(0)=1$ one obtains, for the special case of zero frequency

$$\begin{aligned}
 F(0) = & 2\pi \sum_{n=0}^{N-1} \left(\frac{1}{4} A r_{n+1}^4 + \frac{1}{3} B r_{n+1}^3 + \frac{1}{2} C r_{n+1}^2 \right. \\
 & \left. - \frac{1}{4} A r_n^4 - \frac{1}{3} B r_n^3 - \frac{1}{2} C r_n^2 \right). \quad (\text{A-116})
 \end{aligned}$$

The larger number of terms in the sums, which have to be evaluated for this more accurate form of the algorithm, makes the computation somewhat slower. On the contrary, the higher precision of the evaluation requires less sampling points for the integration while providing the same accuracy of the final result.

Figure A-20 shows numerical calculations of the accuracy of the Hankel transforms in terms of the residuum, using the different algorithms described above, and as a function of the sampling point number N for two different functions: a smooth Gaussian function and a top-hat function with an abrupt transition. It can be seen that the slow, direct numerical integration with parabolic approximation is the method with the highest accuracy. It is interesting to note that the fast method based on the Gardner transform is quite precise for a Gaussian function but is definitely the least precise method for a top-hat profile.

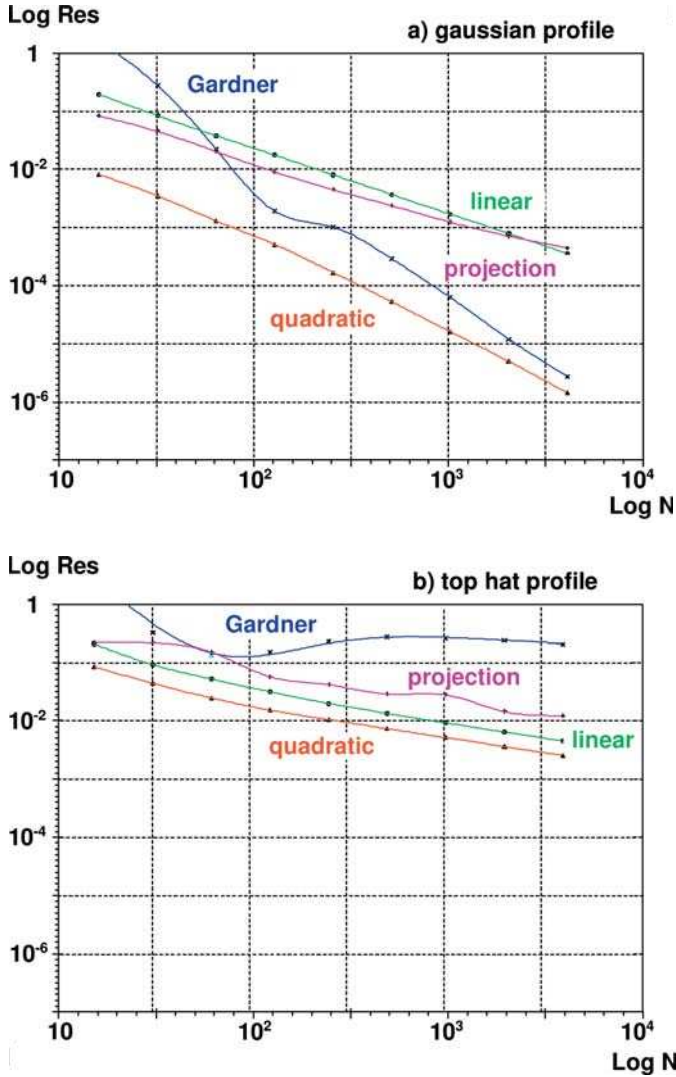


Figure A-20: Logarithm of the residuum for the numerical Hankel transform using different approaches for a Gaussian (a) and a top-hat (b) profile.

As well as the methods outlined in this section, a number of other approaches for the computation of the Hankel transform can be found in the corresponding literature. Depending on the problem to be solved and the requirements for accuracy and speed, some of them may prove to be more suitable in particular cases. In principle, however, because of the singularity problem at $r = 0$, it always makes sense to check whether the choice of the two-dimensional fast Fourier transform is preferable.

A.7
Practical Calculation of Diffraction Integrals

A.7.1
The Oscillation Problem

The practical calculation of diffraction integrals is a cumbersome problem because of the numerical effort required to get accurate results. In one dimension, the main structure of the diffraction integral is of the form

$$U(x_2) = C \cdot \int_{-a}^a U(x_1) \cdot \frac{e^{-ikr_{12}}}{r_{12}} dx_1 \tag{A-117}$$

where r_{12} is the distance between the starting point $P(x_1)$ and the observation point $P(x_2)$. The main problem in computing this integral is the highly oscillating exponent, which requires a very large number of grid points to fulfil the sampling theorem. The major contribution from the integral comes from the points of stationary phase, as the summation over the oscillating parts results in a vanishing amplitude.

Both factors in the above integral contribute to the phase, in general the starting amplitude $U(x_1)$ contains a defocusing part, with a spherical distribution and additional variations caused by aberrations. The second factor characterizes the propagation part of the Huygens spherical wavelets and is critical, if the Fresnel number of the propagation geometry is high. Figure A-21 shows the geometry of the corresponding setup.

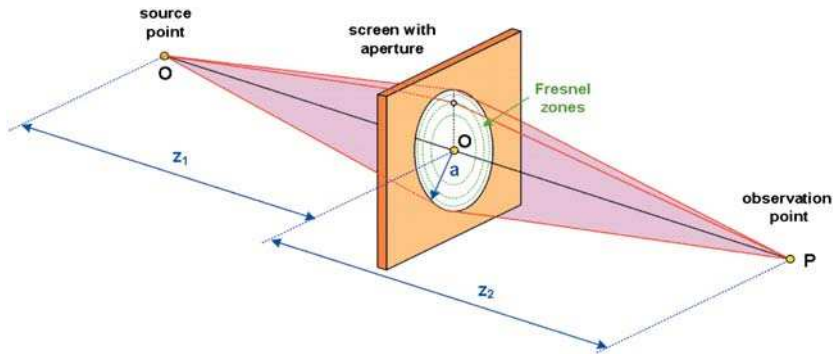


Figure A-21: Geometry of the diffraction calculation. The field of the source is diffracted by an aperture in a screen. The number of Fresnel zones seen by the observation point determines the calculation effort required.

For the mathematical calculation, only that part of the phase function, which is wrapped to the interval 2π , is relevant. If the diffraction integral is written in the Fresnel approximation

$$U(x_2) = C \cdot \int_{-a}^a |U(x_1)| \cdot e^{2\pi \cdot i \cdot \Phi(x_1)} \cdot e^{-\frac{i\pi}{\lambda z} (x_2 - x_1)^2} dx_1 \tag{A-118}$$

where Φ is the phase of the incoming field, figure A-22 illustrates this behavior for an observation point lying at infinity and a defocused incoming wave.

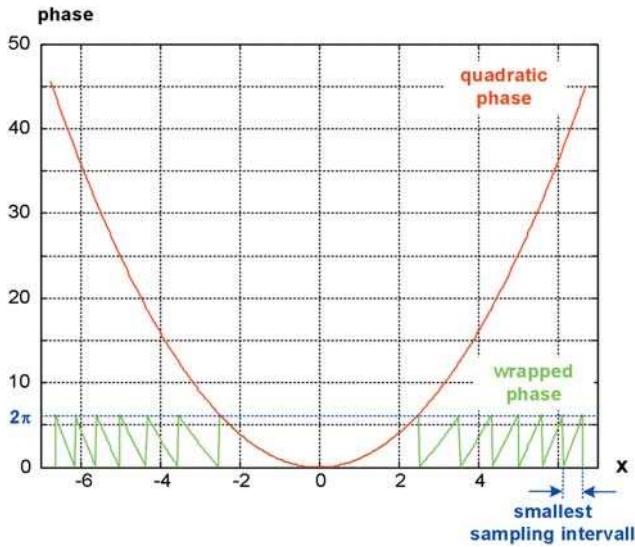


Figure A-22: Parabolic phase function and wrapped version to the interval 2π .

The wrapped phase shows the oscillating behavior in form of a zig-zag curve. To describe this distribution sufficiently accurately, the density of the sampling points is determined at the boundary for the smallest feature size. It can be seen from this figure, that the largest derivation of the phase function defines this limit and therefore the number of grid points which are necessary.

The oscillating behavior can also be demonstrated by the real part of the integrand. Since this a smooth function, a mathematical discussion using this term is much easier.

For the special case of a field with spherical aberration of third order, figure A-23 shows the phase function for an observation point at infinity, a spherical aberration of 3λ corresponding to $c_9 = 3$ and different values of the defocusing. The corresponding phase function may be given by the equation

$$\Phi(x) = 3 \cdot (6x^4 - 6a \cdot x^2) . \quad (\text{A-119})$$

The three cases illustrated in the figure correspond to the values of the defocusing parameter $a = 0, 1$ and 1.5 . The first picture a) shows the fastest oscillation, this highest frequency occurs at the boundary $x = 1$. In the case of a defocusing compensation according to the Zernike definition of the spherical aberration with $a = 1$, the highest spatial frequency is reduced by a factor of 2. The optimal choice of the defocusing in picture c) with $a = 1.5$ reduces this frequency again by a factor of two. The

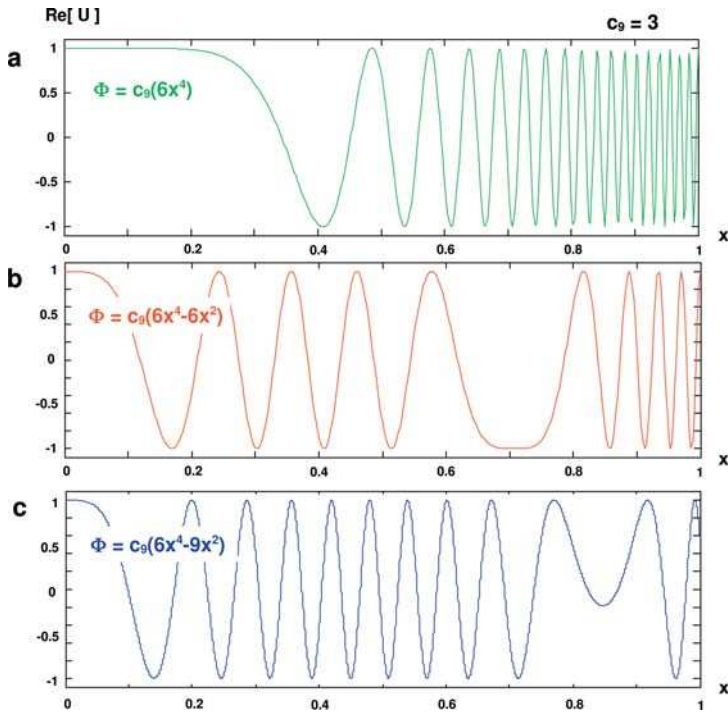


Figure A-23: Phase function for a third-order spherical aberration of 3λ :
 a) without any defocusing contribution; b) with defocusing according to the Zernike polynomial; c) with optimal defocusing for sampling requirements.

third case therefore allows the calculation of the diffraction with one quarter of the sampling points.

The wavefront and the gradient of the wavefront are shown for these three cases in figure A-24. The Zernike definition exhibits the smallest absolute value of the phase, but the optimal choice of a has the smallest gradient. The greatest value of the gradient occurs at the middle of the aperture radius at $x = 0.5$ and at the boundary at $x = 1$ with different signs. This is shown in figure A-25. Figure A-26 illustrates the maximal phase gradient as a function of the defocusing parameter a as a result of numerical calculation. The discrete number of oscillations causes the vibration of the curve for $a < 1.5$.

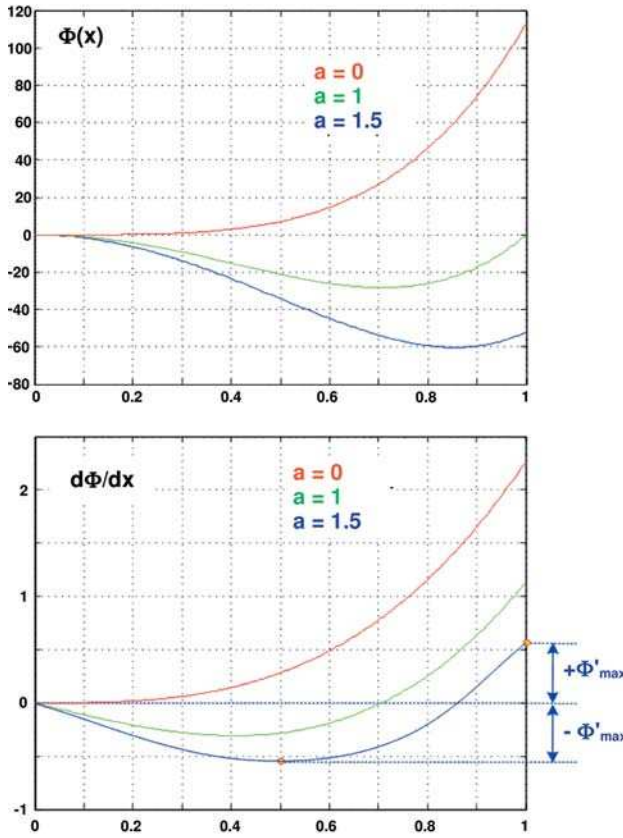


Figure A-24: Phase function for a third-order spherical aberration of 3λ and the three values of defocusing according to figure A-23, and the corresponding gradients of the phase function.

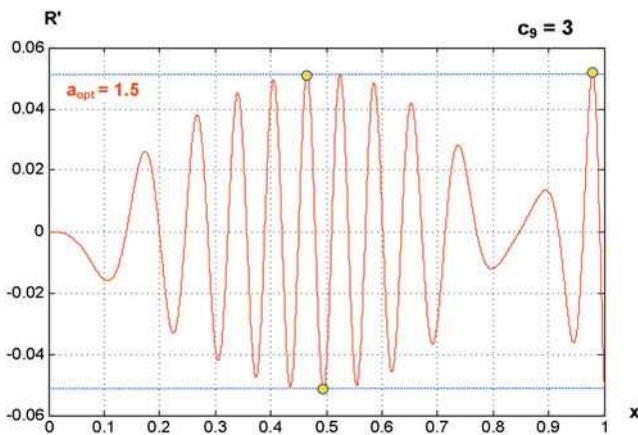


Figure A-25: Gradient of the phase function for a third-order spherical aberration of 3λ and optimal defocusing with $a_{\text{opt}} = 1.5$.

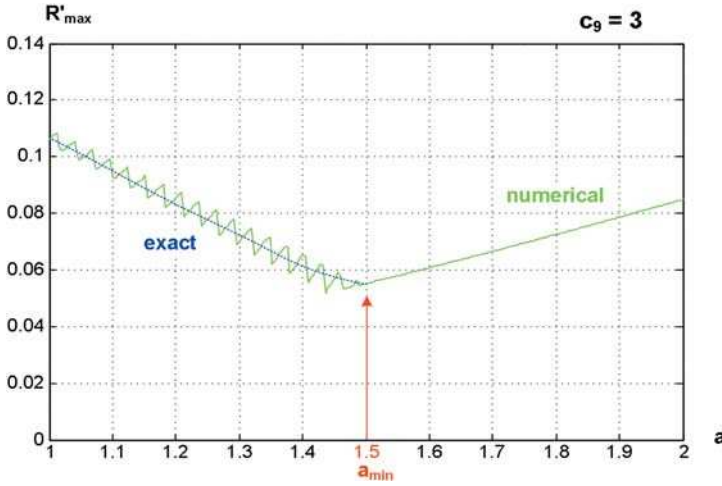


Figure A-26: Maximal gradient of the phase function as a function of the defocusing parameter as a result of numerical calculation.

As a result of this concrete illustration, it can be summarized that the optimal sampling conditions occur, if the gradient of the phase surface has its smallest value across the whole integration range. This causes severe problems in the calculation, if there are large wave aberrations with high orders or short propagation distances with a large variation of the second factor in eq. (A-118).

If only the effects of the propagation distances and defocusing are relevant, the Fresnel number of the diffraction setup is the important term for dealing with sampling requirements. In the principal setup of figure A-21, the Fresnel number is defined by the equation

$$N_F = \frac{a^2}{\lambda} \cdot \left(\frac{1}{z_1} + \frac{1}{z_2} \right). \quad (\text{A-120})$$

The value of the Fresnel number gives the number of Fresnel diffraction zones in the aperture plane, over which the integration has to be performed. According to the sampling theorem, it is necessary to have four grid points over one period of the oscillating phase. But in reality, a better choice is to have six to ten points per period to cancel out the vanishing contributions .

The amplitude distribution in the diffraction integral in eq. (A-118) has not been mentioned until now. In principle, variations of the amplitude of the incoming field with corresponding high spatial frequencies also influence the necessary number of sampling points. But, in practice, the effects of the phase are nearly always dominant and more critical.

A.7.2

Spatial and Spectral Resolution

Propagating electromagnetic fields are bandwidth-limited having a maximum frequency of $1/\lambda$. Hence no aliasing effects occur if the sampling rate is chosen to be higher than $2/\lambda$. This is possible, however, only for small-scale wave-optical computations. In general, one can restrict the simulation to propagation angles limited by the numerical aperture of the imaging optics. Using

$$N = \frac{L}{dx} = L \cdot \Delta\nu = L \cdot 2 \cdot \frac{n \sin \alpha}{\lambda} = 2 \cdot \frac{L}{\lambda} \cdot NA \quad (\text{A-121})$$

one can determine the minimum number of sampling points required, N , dependent on the linear field size L , the numerical aperture NA and the wavelength λ . In order to apply the FFT, however, it is preferable to have N equal to powers of 2.

At the same time one should take into account that a quadratic array consisting of 1024×1024 sampling points, for a typical precision of 4 bytes per float, already requires a storage space of 8.4 MB. This might not seem too much, having in mind the increasing efficiency of modern computers. Nevertheless, the 1024 sampling points are sufficient for the description of only 512 full wavelengths, i.e., for $\lambda = 633$ nm, some $324 \mu\text{m}$ of the linear field extent. The amount of data necessary for wave-optical simulation of a high-aperture microscope objective with a lens diameter of ~ 8 mm is of the order of 8.5 GB.

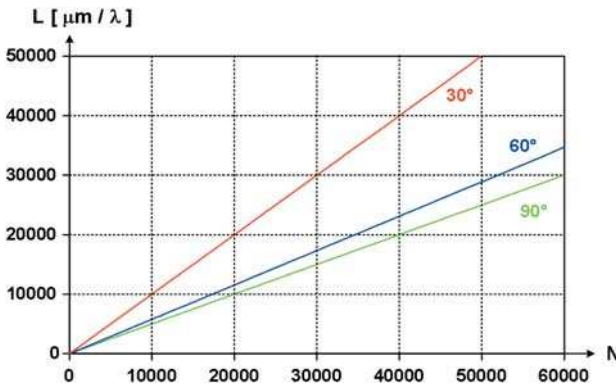


Figure A-27: Linear dimension of the wave-optical simulation field with sampling up to propagation angles of 30°, 60° and 90° and increasing number of sampling points N .

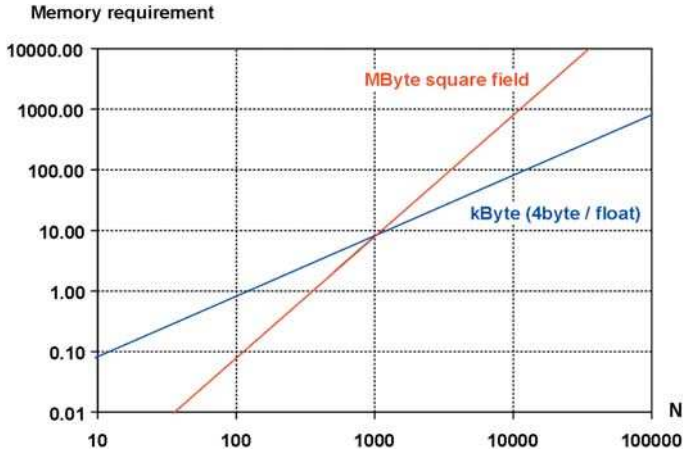


Figure A-28: Memory requirements for a wave-optical simulation dependent on the number of sampling points N (linear and 2D square field; a typical 32 bit-compiler with 4bytes per float is assumed).

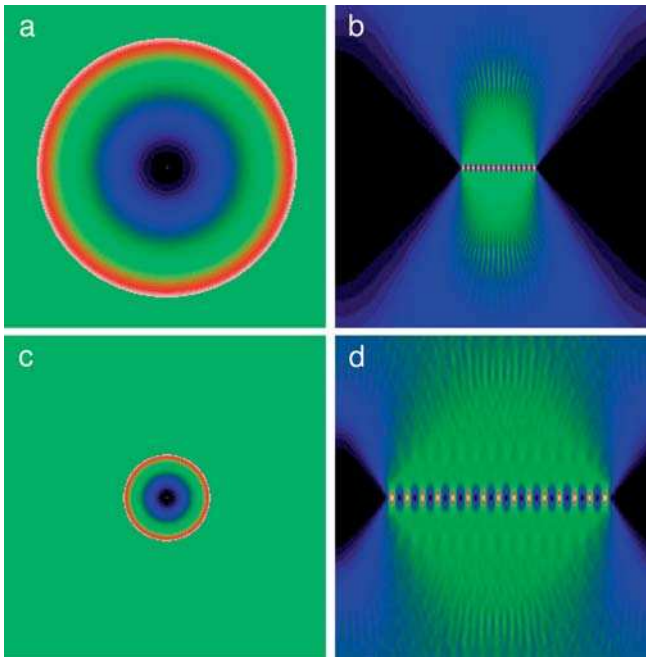


Figure A-29: a) Sampling in the frequency domain in accordance with the sampling theorem up to $\nu_{\max} = \pm NA/\lambda = 1/\lambda$ (with an aperture at $NA = 0.8$); b) calculation field sampled with $\lambda/2$; c) zero padding; maximum frequency $\nu_{\max} = \pm 3/\lambda$; d) calculation field sampled with $\lambda/6$.

It is, however, often desirable to ensure a substantially better spatial resolution in the wave-optical computation as may generally be necessary. For the computation of the point image, e.g., much higher sampling rates are required. Also in the simulation of the partially coherent image formation it is desirable to retrieve the influence of details which are much finer than the resolution limit of $\lambda/2$ and to evaluate the effect of different aspect ratios or separations from the objects (so-called proximity effects or the influence of assist features such as serifs). Sampling further beyond the sampling theorem leads to strong under-filling of the spectrum, the so-called zero padding (figure A-29c). As a consequence of the zero padding, for the same number of sampling points, the frequency domain is to an increasing degree under-sampled. Obviously a trade-off has to be found, having in mind that in general the spatial resolution deteriorates with increasing frequency of wave front aberrations.

A.7.3

Periodic Boundary Conditions

In the discrete representation, both the spectrum and the object field should be thought of as being periodically repeating. Thus reflections at the boundaries of the calculation field may occur, e.g., in the simulation of the free space propagation. In order to avoid errors in the simulation results for optical image formation and in particular for the propagation of wave fields, caused by periodicity effects, one should also allow for sufficient space in the boundary regions of the spatial domain.

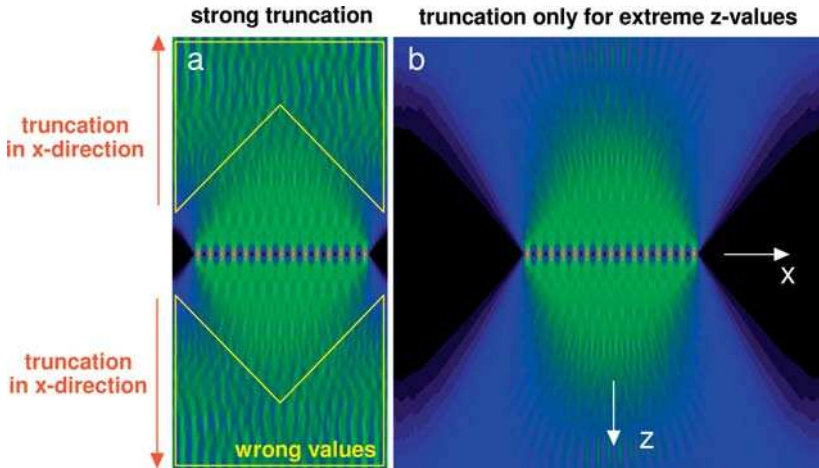


Figure A-30: Perturbations in the boundary region caused by the periodic continuation of the calculation field in the x direction.

Furthermore, for stepwise simulations of the free space propagation (e.g., using the wave propagation methods [A-13]), it is reasonable to apply a weak absorber ($\sim e^{-\kappa dx}$ where dx is the distance from the boundary and κ is an absorption coefficient; reliable values are $\kappa < 0.0005/\lambda$) in the field boundary region covering a few dozens of pixels in order to suppress the periodic boundary conditions.

A.7.4
x-z Sampling of the Ewald Sphere

For three-dimensional image formation, the image is computed from the spectrum on the Ewald sphere. In order to utilize the high speed of the FFT, cartesian coordinates have to be chosen for the representation. The sampling points in this case do not correspond to the coordinates on the Ewald sphere. Some authors have provided extensive considerations of a sampling theorem in the ν_z direction. The spectrum can be accurately occupied using the sinc interpolation or the following trick can be simply applied instead: the 2D Fourier transform of a field corresponds to the 1D Fourier transform of all rows (e.g., in the ν_x direction) followed by a transform of all columns (e.g., in the ν_z direction) of the 2D field and vice versa. Since the spectrum is discrete in the z direction and consists only of a delta function, the column Fourier transform yields a harmonic function in each column. Thus the 2D Fourier transform of the 2D spectrum proceeds as an analytical Fourier transform in the z direction and subsequent FFT over all rows (in the x direction) and one obtains the 2D field distribution in accordance with the standard sampling theorem. At the same time the x and z scaling factors can be arbitrarily chosen and may be different; e.g., it is often necessary to select a substantially smaller number of image planes to be computed in the z direction, with larger separations dz , in comparison with the sampling in the x direction with an interval dx .

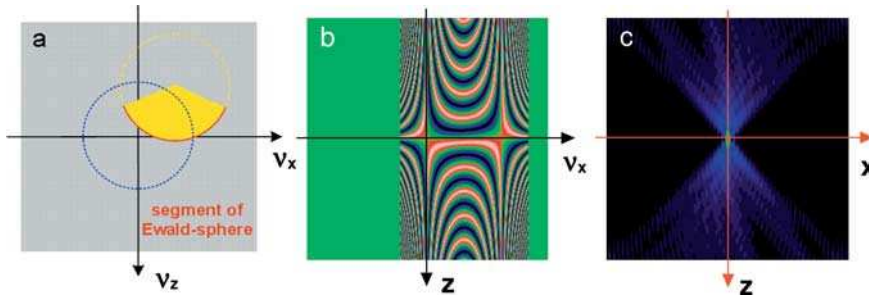


Figure A-31: (a) Computation of an x-z intensity scan from the Ewald-sphere; (b) by analytical Fourier transformation in the ν_z direction; and (c) final application of the FFT in ν_x for all lines z .

A.7.5
Equivalent Diffraction Setups

The diffraction integral in the Fresnel approximation reads according to eq. (A-118)

$$U(x_2) = C \cdot \int_{-a}^a |U(x_1)| \cdot e^{2\pi i \Phi(x_1)} \cdot e^{-\frac{i\pi}{\lambda z} (x_2 - x_1)^2} dx_1 . \tag{A-122}$$

It is possible to transform this integral in an equivalent form, but with a rather better conditioning of the sampling problem. The following equation with an arbitrary number M delivers identical results [A-14]

$$U(x_2, z) = C \cdot e^{-\frac{i\pi(M-1)x_2^2}{\lambda \cdot M \cdot z}} \cdot \int \left[U(x_1, 0) \cdot e^{-\frac{i\pi(1-M)x_1^2}{\lambda \cdot z}} \right] \cdot e^{-\frac{i\pi \cdot M}{\lambda \cdot z} \left(x_1 - \frac{x_2}{M}\right)^2} dx_1 \quad (\text{A-123})$$

where x_1 is the coordinate in the source or pupil plane, over which the integration is calculated. The additional factor in the square brackets describes a defocusing in paraxial approximation. The physical interpretation of this equation is that, for every diffraction setup, there are other geometries, for which the scaling of the observation plane, a final paraboloidal phase correction and the propagation distance, have to be chosen appropriately, but the field distribution is identical. This is illustrated in figure A-32. It should be noted, that this equivalence of the equations is only valid in paraxial approximation.

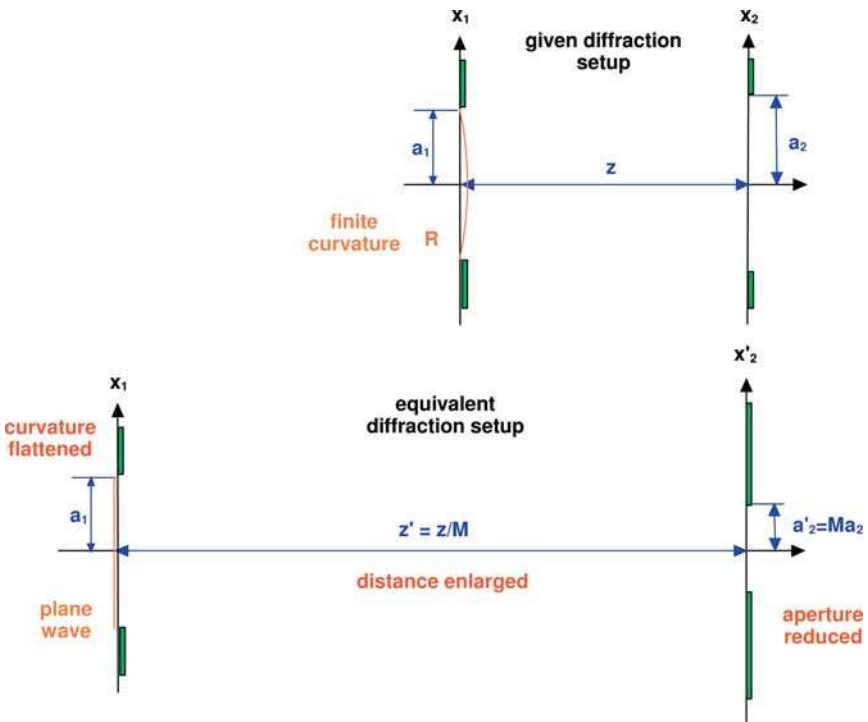


Figure A-32: Equivalent diffraction setup with scaling of the aperture and the distance, and using a different phase correction for $M = 0.5$.

The choice of the magnification factor M is arbitrary and can be used to optimize the sampling conditions for the calculation of the integral. If the incoming field has no phase distortions of higher order and is therefore of paraboloidal form, the opti-

mal defocusing corresponds to a flattening of the phase surface. As a result, the phase front becomes plane after the equivalence transformation. If the radius of curvature is R and the propagation distance z , the best choice of the magnification factor M is

$$M = 1 + \frac{z}{R} . \tag{A-124}$$

The correct propagation distance is then given by

$$z' = \frac{z}{M} \tag{A-125}$$

and the scaling of the coordinates in the observation plane results from

$$x_2 = M \cdot x_1 , \quad y_2 = M \cdot y_1 . \tag{A-126}$$

Finally, the field has to be corrected by a parabolic phase factor with the radius of curvature

$$R' = R - z . \tag{A-127}$$

It should be noted, that the above formulas are also valid in the case of negative values of M , this happens for distances z greater than R . In this case, an internal focal point is located at the propagation distance. If propagation in a focal plane is required, the value of M is nearly or exactly zero. In this special case, the equivalence transformation is not suitable, it is only advantageous for real Fresnel transitions.

There is another geometrical interpretation of the equivalence transform. The flattening of the field corresponds to a transition into spheroidal coordinates with a center in the focal point. Figure A-33 shows this interpretation. In particular, this picture illustrates the occurrence of an internal focal point with a change in the sign of the coordinate in the observation plane.

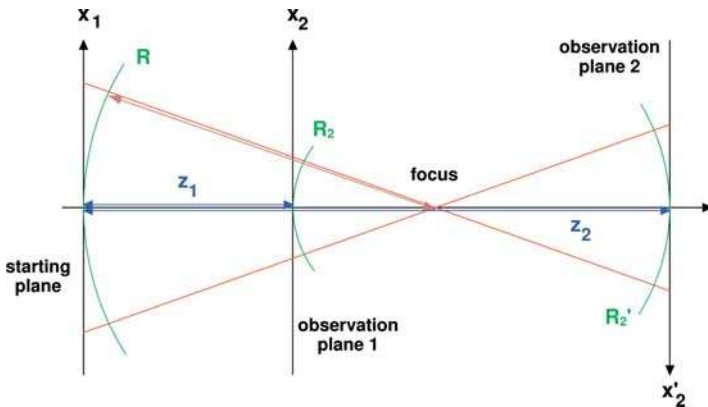


Figure A-33: Equivalent diffraction setup with scaling of the aperture and the distance, and using a different phase correction.

It should be mentioned that this representation of the equivalence transform can also be interpreted as a fractional Fourier transform, which does not correspond to a full 90° rotation in the phase space as in the Fraunhofer case [A-15], [A-16]. There are some additional attempts in the literature to use the above idea to more special configurations. For example, in [A-17] the transform is used for oblique off-axis beam propagation.

A.7.6

Optimal Conditioning of the Fresnel Diffraction

If a diffraction calculation has to be performed, first of all, the optimal conditioning of the given setup has to be analyzed. In most cases, the distinction between a Fresnel setup with a near-field transition and a Fraunhofer approximation with far-field conditions has to be made. There are four different scenarios depending on the distance from the source, the stop and the observation plane, Figure A-34 shows these different geometries [A-18]. A lens brings a plane at a finite distance to infinity, in the optical sense.

If a diffraction calculation in the Fraunhofer approximation has to be made, only a simple Fourier transform is necessary or a simple integration, if the computation of the integral is realized directly.

On the other hand, in the Fresnel case, the spectral distribution first has to be calculated, and after propagating this, a transform back to the spatial domain has to be made.

If a light field is separated in the focal region and the outer domain, there are four cases for a diffraction calculation [A-19], [A-20] depending on the location of the origin and the observation plane. These four possibilities are shown in figure A-35. In the outer ranges outside the focal region, the geometrical curvature is of a considerable size, so an equivalence transformation, according to the previous section, is advantageous. Inside the focal region in the range of the focal depth, the curvature is rather small, there is no real problem in the sampling of the complex field. The separation between the four cases is not rigorous and gives equal results in the overlapping ranges. But a violation of the criteria of optimal calculating ranges, results in an increasing computational effort. The four cases are:

1. Transition $O \rightarrow O$ between outer regions of the focal plane. The Fresnel calculation uses the equivalence transform, with the Fourier transform \hat{F} it follows that

$$U_{O-O}(x_2) = \frac{e^{-ik\frac{z}{M}}}{\sqrt{|M|}} e^{-\frac{i\pi(M-1)}{\lambda Mz}x_2^2} \hat{F}^{-1} \left\{ e^{\frac{i\lambda z}{4\pi M}k_0^2} \hat{F} \left[U(x_1) \cdot e^{-\frac{i\pi(1-M)}{\lambda z}x_1^2} \right] \right\}. \quad (\text{A-128})$$

The scaling factor has to be optimized according to the criteria described above, it lies in the range

$$M = 1 + \frac{z}{R} \quad (\text{A-129})$$

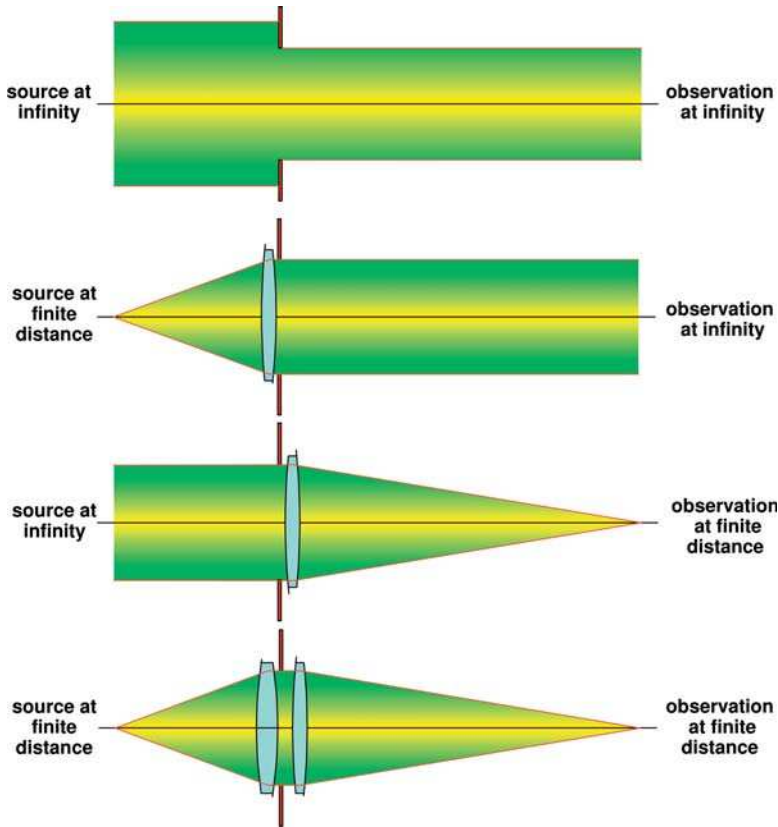


Figure A-34: Equivalent diffraction setup with scaling of the aperture and the distance, and using a different phase correction.

with R the mean radius of curvature of the field in the starting plane. If the lateral size of the aperture opening in the starting plane is a_1 , the diffraction calculation has an effective Fresnel number

$$N_F^{\text{eff}} = \frac{M a_1^2}{\lambda z} . \tag{A-130}$$

2. Transition $I \rightarrow I$ between inner regions of the focal plane. This is the easiest case and the calculation procedure follows the formula

$$U_{I \rightarrow I}(x_2) = \hat{F}^{-1} \left\{ e^{j\pi\lambda z v^2} \cdot \hat{F}[U(x_1)] \right\} . \tag{A-131}$$

The phase front is nearly plane and an additional flattening transform is not necessary.

3. Transition I → O from the inner to the outer region of the focal range. This corresponds to a simple Fraunhofer transition into the far field, the calculation can be written as

$$U_{I-O}(x_2) = \frac{1}{\sqrt{\lambda z}} \cdot e^{-\frac{i\pi}{\lambda z} x_2^2} \cdot \hat{F}^{-1} \left\{ U(x_1) \cdot e^{-\frac{i\pi}{\lambda z} x_1^2} \right\}. \tag{A-132}$$

Since the relations of the Fourier transform are valid, the scaling of the coordinates in the observation plane can be calculated according to the equation

$$x_2 = \lambda \cdot z \cdot \nu \tag{A-133}$$

with the spatial frequency ν .

4. Transition O → I from the outer to the inner region of the focal range. This also corresponds to a Fraunhofer transition, but first the flattening phase factor has to be applied. Formally, the equation is identical to (A-132), but the quadratic factor now has a considerable effect:

$$U_{O-I}(x_2) = \frac{1}{\sqrt{\lambda z}} \cdot e^{-\frac{i\pi}{\lambda z} x_2^2} \cdot \hat{F}^{-1} \left\{ U(x_1) \cdot e^{-\frac{i\pi}{\lambda z} x_1^2} \right\}. \tag{A-134}$$

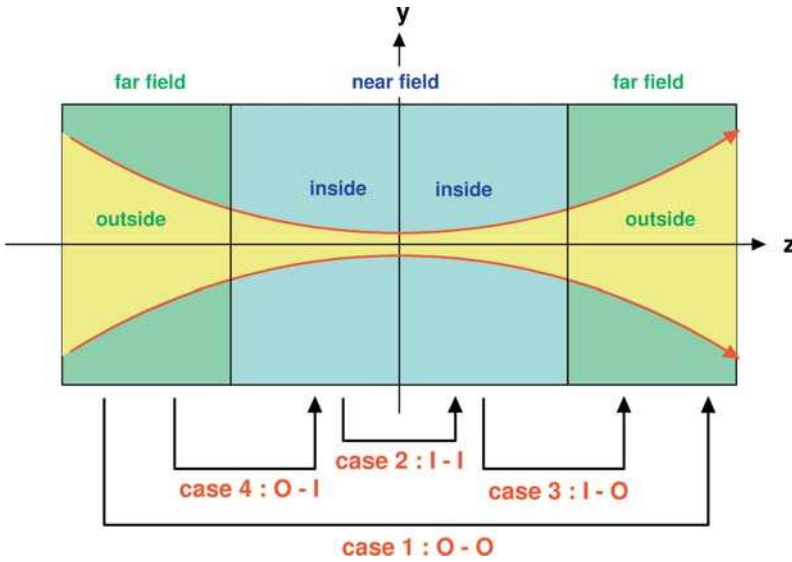


Figure A-35: Equivalent diffraction setup with scaling of the aperture and the distance, and using a different phase correction.

If in the more general case, the incoming phase front is not paraxial, but has some higher order perturbations with steep phase gradients, the optimal choice of the factor M is determined in the sense of an optimal sampling scenario. The condition for the best value of defocusing with a paraxial radius of curvature R_0 reads, in one dimension

$$\left| \frac{d\Phi}{dx} - \frac{x}{R_0} \right|_{\max, \text{all } x} = \min. \quad (\text{A-135})$$

As a result, the gradient of the remaining phase after the equivalence transform has the smallest value which is possible.

The above explanation uses free-space propagation and is limited to only one transverse dimension. But the idea of the equivalence transform can be extended without problems to two dimensions and setups with paraxial system parts [A-21].

A.7.7

Numerical Algorithms

For the calculation of a diffraction problem, there are several possible numerical realizations. The optimal choice of the algorithm depends on the computer memory resources, the computing time, the required accuracy and the use of an intelligent pre-calculation, which is able to select the best available tool. There are a lot of calculation schemes proposed in the literature. In this section, only some of the important algorithms will be briefly described. Concrete aspects of the implementation can be found in the references.

1. Fraunhofer case with the Fourier method

When the Fraunhofer approximation can be applied, the use of the fast Fourier transform is the most usual way of doing this. The calculation can then be performed very quickly. Some disadvantages and problematic aspects are:

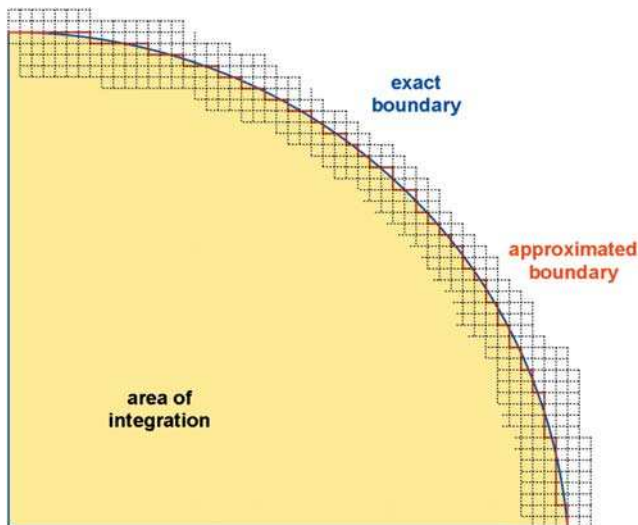


Figure A-36: Discretization of the boundary for the two dimensional integration.

- 1.1 The grid is assumed to be equidistant. If this is not the case in the starting plane, a cumbersome interpolation has to be carried out with the complex field.
- 1.2 According to 1.1, the description of the boundary is only very rough and has a corrugated form. This causes some errors in the computation, depending on the geometry of the integration area. Figure A-36 illustrates this problem. The accuracy can be increased, if a suitable weighting of the mesh points near the boundary is introduced [A-22].
- 1.3 If the usual Fourier transform with the fast Cooley–Tuckey algorithm is used, the spatial grid size and the number of points defines the spacing in the frequency domain. If a special frequency resolution is required, it is necessary to use a zero padding or to apply the famous chirp-z-transform, which is described in detail in section A-5.

2. Fresnel case with the Fourier method

In the range of the Fresnel approximation, the most important point is to use the equivalence transform, described in section A-3. Since all the integrals have the form of a convolution, the numerical computation can be done by means of the fast Fourier algorithms as in the previous case. All the problems discussed there are also present in the Fresnel case. One additional aspect should be mentioned here. In the calculation, a transition to the frequency range is first carried out and after propagating the plane waves the distribution is transformed back to the spatial domain, therefore a violation of the sampling theorem in the frequency domain cannot always be recognized from the result. For this reason it is recommended to control the convergence of the frequency spectrum in the intermediate calculation step.

It should be noticed, that in the Fresnel regime, there is a mixture between the resolution and therefore between the sampling problems in the spatial and frequency domain. This mixture changes with the propagation distance z and there is not one algorithm, which fulfils all the sampling requirements [A-23].

3. Direct evaluation of the integral

If the diffraction integrals in the various formulations are evaluated directly, the computational effort increases with the linear number of sampling points N as N^4 . This causes a long computation time, if only 'brute force' methods are applied. The advantage of this procedure is that, in contrast to the Fourier methods, there are no problems with sampling in the frequency domain or with non-equidistant grids in the source plane. In practice, the accuracy of the result is not sufficient, if only a rectangular integration rule is applied to the evaluation of the sum.

There are several clever attempts which have been reported in the literature which deal with this problem.

One method uses a finite element calculation algorithms to evaluate the integrals with a reasonable small number of sampling intervals [A-24], [A-25]. The main advantage of this method is a very accurate modelling of the boundary of the source

area. Especially complicated aperture distributions with smart obscuration structures can be described well.

One of the oldest methods of dealing with the highly oscillating integrand in the diffraction integral is to approximate the amplitude and phase function linearly inside the grid cells [A-26], [A-27]. Figure A-37 shows this idea for a single discrete sampling interval. The contribution to the one-dimensional diffraction integral can be written as

$$I_j = \int_{x_j}^{x_{j+1}} A(x) \cdot e^{2\pi i \Phi(x)} dx . \quad (\text{A-136})$$

If $A(x)$ and $\Phi(x)$ are linearly approximated, the integral can be solved analytically inside this single sampling interval.

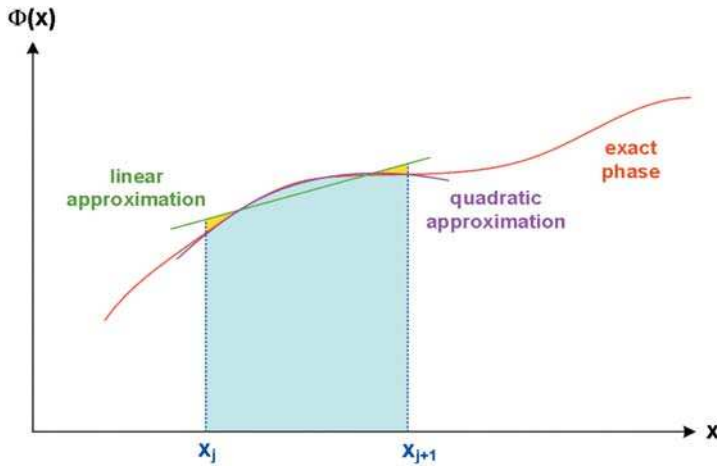


Figure A-37 Linear or quadratic approximation of the phase function $\phi(x)$ inside a grid interval for the numerical computation of the diffraction integral.

It is always useful to have large sampling intervals, to keep the computational time small. But if the curvature of the functions $A(x)$ and $\Phi(x)$ are not negligible in the interval Δx , an error occurs in the calculation. The next best approximation therefore uses a quadratic approximation of the two functions [A-28], [A-29]. In this case, the remaining integrals inside one interval are not elementary. But in the cited references, in the so-called SSP-method, a clever selection and approximation of the contributions delivers an accurate and fast integration scheme. This quadratic approximation of the amplitude and phase distribution, also indicated in figure A-37, improves the accuracy significantly.

4. Propagation with high numerical aperture

The near-field diffraction calculation scheme as described in section A.7.6 in eq. (A-131) is, in principle, an angular propagation. If the given field distribution is Fourier transformed, this corresponds to an expansion into plane waves, every Fourier component represents a plane wave, the value of the coefficient gives the

relative content of this inclined plane wave. The propagation of plane waves in homogeneous media can be calculated by a simple phase factor, depending on the inclination angle. For the spatial frequency ν , the propagator reads

$$T(z) = e^{\frac{2\pi i}{\lambda} z \cdot \sqrt{1 - \lambda^2 \nu^2}} . \quad (\text{A-137})$$

In paraxial approximation, this term can be written as

$$T_{\text{parax}}(z) = e^{\frac{2\pi i}{\lambda} z} \cdot e^{-\pi i \lambda \cdot z \cdot \nu^2} \quad (\text{A-138})$$

and this is the form of the factor as it is used in eq. (A-131).

Recently, a more exact formula has been reported, where the paraxial approximation is expanded to the high-angle range in the framework of a scalar model [A-30]. In the extension of the paraxial model, where the spatial frequencies are given by

$$x_2 = \lambda \cdot z \cdot \nu_x \quad (\text{A-139})$$

in the range of high numerical aperture, the coordinates in the observation plane are described by

$$x_2 = \lambda \cdot r \cdot \nu_x . \quad (\text{A-140})$$

The distance r between the observation point and the source point can be calculated by the equation

$$r = \sqrt{z^2 + (x_2 - x_1)^2 + (y_2 - y_1)^2} . \quad (\text{A-141})$$

The coordinate x_2 in the observation plane then depends in a nonlinear manner on the frequency and there occurs a coupling between the two coordinate directions x and y

$$x_2 = \frac{\lambda \cdot z \cdot \nu_x}{\sqrt{1 - \lambda^2 \cdot \sqrt{\nu_x^2 + \nu_y^2}}} . \quad (\text{A-142})$$

As a result of this more general formulation, the output spatial coordinates are not equidistant and have to be re-interpolated to a desired equidistant grid.

A.7.8

Fresnel Integrals

The Fresnel integral in its complex formulation is defined as

$$F(x) = \int_0^x e^{\frac{1}{2} i \pi t^2} dt . \quad (\text{A-143})$$

It is related to the Gaussian error function $\text{erf}(z)$ by [A-31]

$$F(x) = C(x) + i \cdot S(x) = \frac{1+i}{2} \cdot \text{erf} \left[\frac{\sqrt{\pi}}{2} \cdot (1-i) \cdot x \right] . \quad (\text{A-144})$$

The following asymptotic approximations of the Fresnel integral are valid in the limit of very small and large values of the argument, respectively:

1. Small values of the argument $x \ll 1$

$$F(x) = x \cdot e^{\frac{1}{2}i\pi x^2} . \tag{A-145}$$

2. Large values of the argument $x \gg 1$

$$F(x) = \frac{i + 1}{2} - \frac{i}{\pi x} \cdot e^{\frac{1}{2}i\pi x^2} . \tag{A-146}$$

If the real and imaginary parts of the integral are separated

$$F(x) = C(x) + i \cdot S(x) \tag{A-147}$$

one obtains

$$C(x) = \int_0^x \cos\left(\frac{1}{2}\pi t^2\right) dt , \tag{A-148}$$

$$S(x) = \int_0^x \sin\left(\frac{1}{2}\pi t^2\right) dt . \tag{A-149}$$

The behavior of the real and imaginary parts as functions of the real argument x is illustrated in figure A-38.

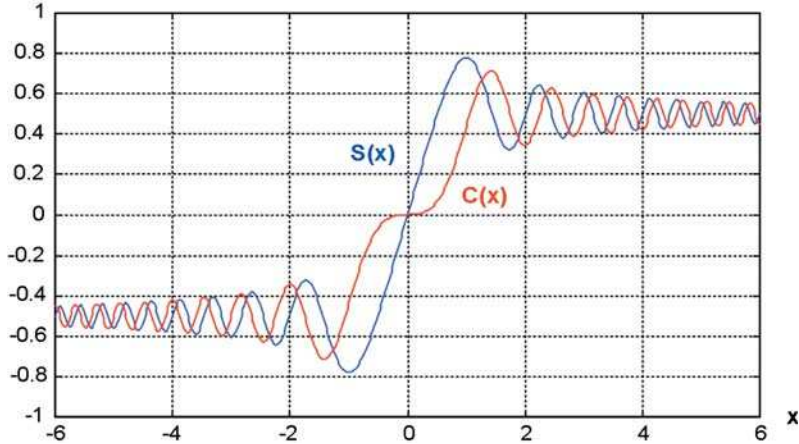


Figure A-38: Real $C(x)$ and imaginary $S(x)$ parts of the Fresnel integral.

For coherent Fresnel diffraction of an infinite plane wave incident on a long edge, the intensity can be presented in terms of Fresnel integrals as

$$I(t) = \frac{1}{2} \cdot \left[\left(\frac{1}{2} - C(t) \right)^2 + \left(\frac{1}{2} - S(t) \right)^2 \right] \tag{A-150}$$

where t is a scalable argument related to the Fresnel number N_F :

$$t = \sqrt{\frac{k}{z \cdot \pi}} \cdot x = \sqrt{\frac{2}{\lambda \cdot z}} \cdot x = \sqrt{2N_F} . \quad (\text{A-151})$$

The Fresnel diffraction at an edge can be illustrated by means of the so-called Cornu spiral [A-3] shown in Figure A-39. The amplitude of the diffracted field is decomposed into real and imaginary parts which equal the Fresnel integrals $C(t)$ and $S(t)$, respectively. The Fresnel integrals are plotted on the horizontal and vertical axes of a diagram. The resulting curve has a spiral shape with asymptotic points at $(1/2, 1/2)$ and $(-1/2, -1/2)$ for very large and small values of the argument t . These correspond to large separations from the geometrical shadow boundary in the light and dark regions, respectively. The amplitude of the field is given by the distance from a point on the curve to the starting point $(-1/2, -1/2)$. The convergence towards the asymptotic points reflects the decay of the diffraction ripples for large separations from the edge.

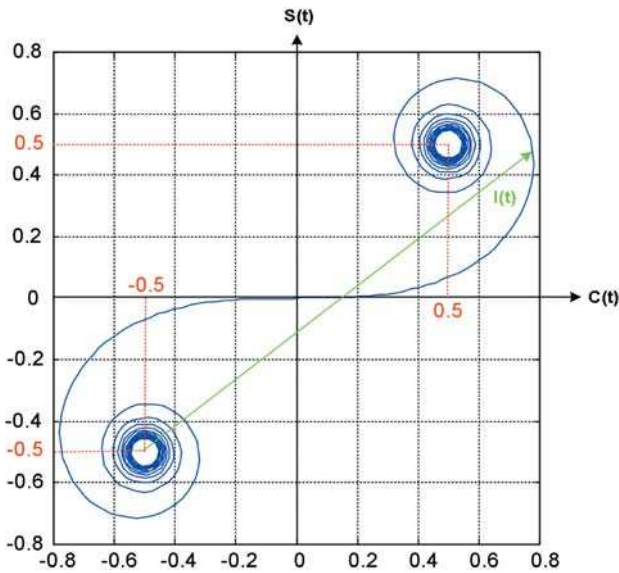


Figure A-39: The Cornu spiral as a visualization of the Fresnel integrals.

The intensity at the boundary of the geometrical shadow amounts to $I(0) = 1/4$. It continuously fades away to zero in the shadow region while in the region of geometrically full illumination one observes an oscillating behavior with convergence to the value 1 corresponding to undisturbed illumination. This is shown in figure A-40.

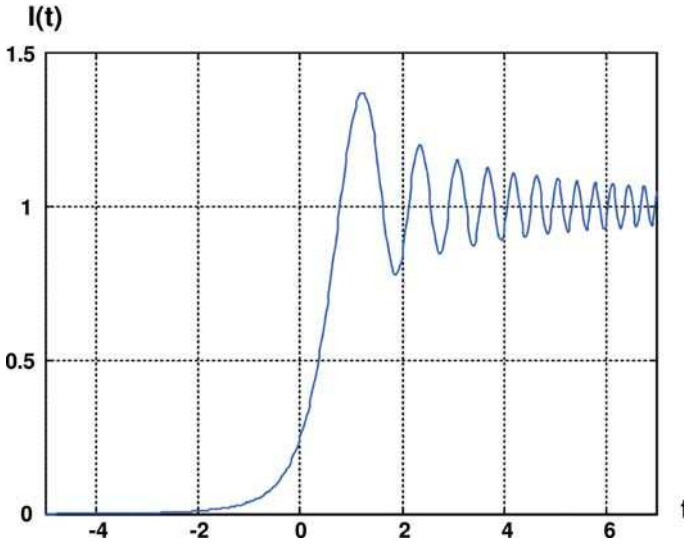


Figure A-40: Diffraction pattern of a straight edge in the Fresnel approximation

**A.8
Orthogonal Polynomials on Rectangular Domains**

**A.8.1
Chebyshev Polynomials**

The one-dimensional Chebyshev polynomials of the first kind $T_n(x)$ are defined on the unit interval $-1 \dots x \dots 1$ with a weight function corresponding to the reciprocal value of the height of the unit circle [A-31]

$$\int_{-1}^{+1} T_n(x) \cdot T_m(x) \cdot \frac{1}{\sqrt{1-x^2}} dx = \begin{cases} 0 & \text{if } n \neq m \\ \pi/2 & \text{if } n = m > 0 \\ \pi & \text{if } n = m = 0. \end{cases} \quad (\text{A-152})$$

The explicit form of the polynomials can be easily written in terms of circular functions

$$T_n(x) = \cos [n \cdot \arccos(x)] . \quad (\text{A-153})$$

The lowest order Chebyshev polynomials of the first kind read

$$\begin{aligned}
T_0(x) &= 1 \\
T_1(x) &= x \\
T_2(x) &= 2x^2 - 1 \\
T_3(x) &= 4x^3 - 3x \\
T_4(x) &= 8x^4 - 8x^2 + 1 \\
T_5(x) &= 16x^5 - 20x^3 + 5x \\
T_6(x) &= 32x^6 - 48x^4 + 18x^2 - 1 \\
T_7(x) &= 64x^7 - 112x^5 + 56x^3 - 7x \\
T_8(x) &= 128x^8 - 256x^6 + 160x^4 - 32x^2 + 1 \\
T_9(x) &= 256x^9 - 576x^7 + 432x^5 - 120x^3 + 9x \\
T_{10}(x) &= 512x^{10} - 1280x^8 + 1120x^6 - 400x^4 + 50x^2 - 1 \\
T_{11}(x) &= 1024x^{11} - 2816x^9 + 2816x^7 - 1232x^5 + 220x^3 - 11x \\
T_{12}(x) &= 2048x^{12} - 6144x^{10} + 6912x^8 - 3584x^6 + 840x^4 - 72x^2 + 1.
\end{aligned} \tag{A-154}$$

The characteristic behavior of these functions can be seen in figure A-41.

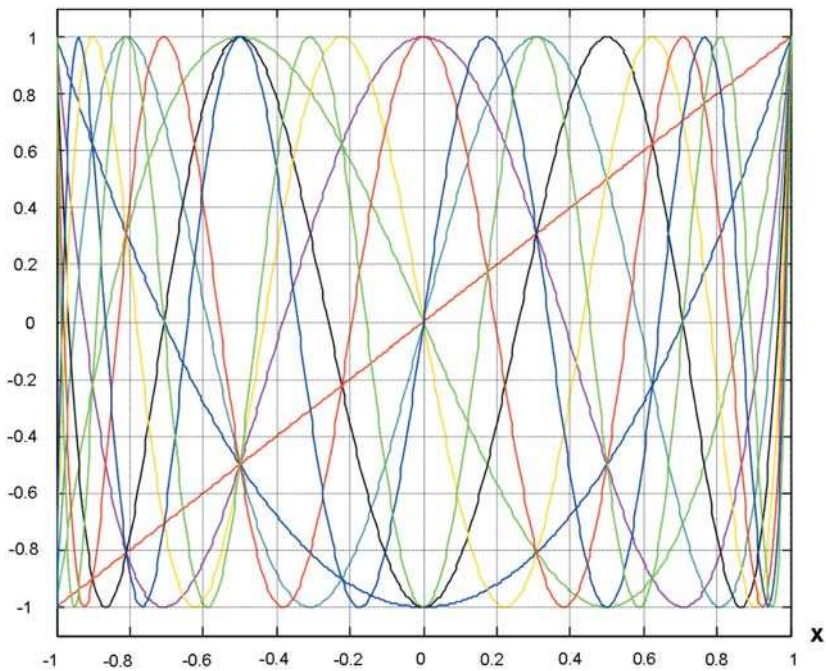


Figure A-41: Lowest order Chebyshev polynomials of the first kind in one dimension.

A Chebyshev polynomial at one point can be expressed by neighboring Chebyshev polynomials at the same point (recurrence relation)

$$T_{n+1}(x) = 2x \cdot T_n(x) - T_{n-1}(x) . \quad (\text{A-155})$$

A.8.2

One-dimensional Legendre Polynomials

The orthogonality relation of the one-dimensional Legendre polynomials with constant weight function reads [A-31]

$$\int_{-1}^{+1} P_n(x) \cdot P_m(x) dx = \begin{cases} 0 & \text{if } n \neq m \\ \frac{2}{2n+1} & \text{if } n = m . \end{cases} \quad (\text{A-156})$$

The recurrence relation is given by

$$(n+1) \cdot P_{n+1}(x) = (2n+1) \cdot x \cdot P_n(x) - n \cdot P_{n-1}(x) . \quad (\text{A-157})$$

The explicit form of the first polynomials is

$$\begin{aligned} P_0(x) &= 1 \\ P_1(x) &= x \\ P_2(x) &= (3x^2 - 1)/2 \\ P_3(x) &= (5x^3 - 3x)/2 \\ P_4(x) &= (35x^4 - 30x^2 + 3)/8 \\ P_5(x) &= (63x^5 - 70x^3 + 15x)/8 \\ P_6(x) &= (231x^6 - 315x^4 + 105x^2 - 5)/16 \\ P_7(x) &= (429x^7 - 693x^5 + 315x^3 - 35x)/16 \\ P_8(x) &= (6435x^8 - 12012x^6 + 6930x^4 - 1260x^2 + 35)/128 \\ P_9(x) &= (12155x^9 - 25740x^7 + 18018x^5 - 4620x^3 + 315x)/128 \\ P_{10}(x) &= (46189x^{10} - 109395x^8 + 90090x^6 - 30030x^4 + 3465x^2 - 63)/256 \\ P_{11}(x) &= (88179x^{11} - 230945x^9 + 218790x^7 - 90090x^5 + 15015x^3 - 693x)/256 \\ P_{12}(x) &= (676039x^{12} - 1939938x^{10} + 2078505x^8 - 1021020x^6 + 225225x^4 \\ &\quad - 18018x^2 + 231)/1024 . \end{aligned} \quad (\text{A-158})$$

The lowest order Legendre polynomials are plotted in figure A-42.

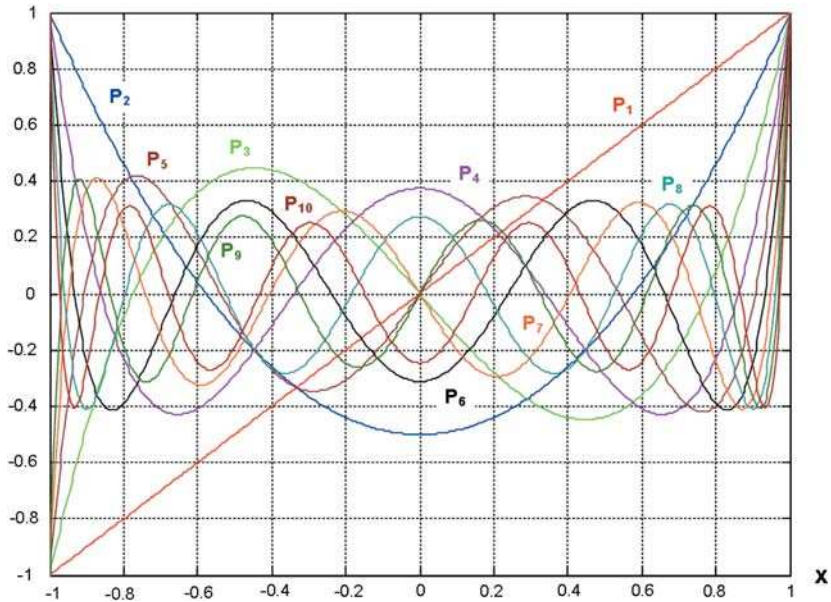


Figure A-42: Lowest order Legendre polynomials in one dimension.

A.8.3

Two-dimensional Chebyshev Polynomials

In two dimensions, a function can be constructed in the form of a product of the one-dimensional Chebyshev polynomials as follows

$$W(x, y) = \sum_n \sum_m D_{nm} \cdot T_n(x) \cdot T_m(y) . \quad (\text{A-159})$$

Figure A-43 shows the characteristic surface shapes of this class of functions. The very regular behavior and the fact that the weight function makes the outer region of the pupil area count more, are again drawbacks, which have to be mentioned here.

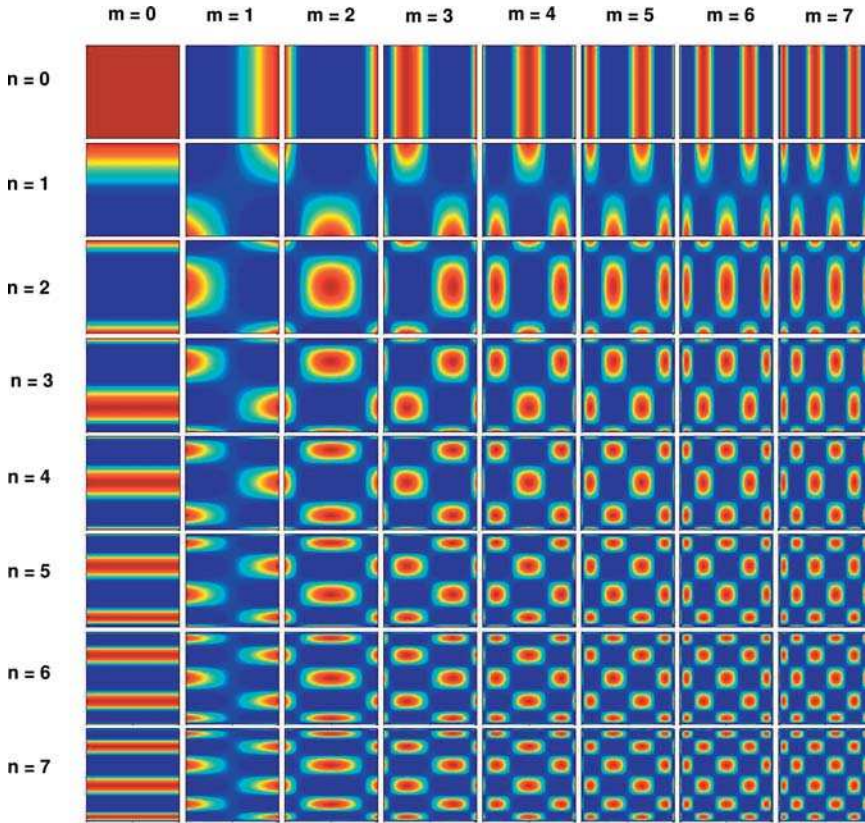


Figure A-43: Representation of the products of one-dimensional Chebyshev polynomials on a square-shaped area.

A.8.4 Legendre Polynomials in Two Dimensions

A factorized form based on the one-dimensional Legendre polynomials is possible in two dimensions [A-32], [A-33], as given below

$$W(x, y) = \sum_n \sum_m A_{nm} P_n(x) \cdot P_m(y) . \tag{A-160}$$

This form, however, is not orthogonal on the rectangular area of definition. Specifically, the zero points of one of the factors in the above-defined form, give zero for all orders of the other coordinate direction. Figure A-44 illustrates this two-dimensional use of the Legendre polynomials on a square-shaped aperture.

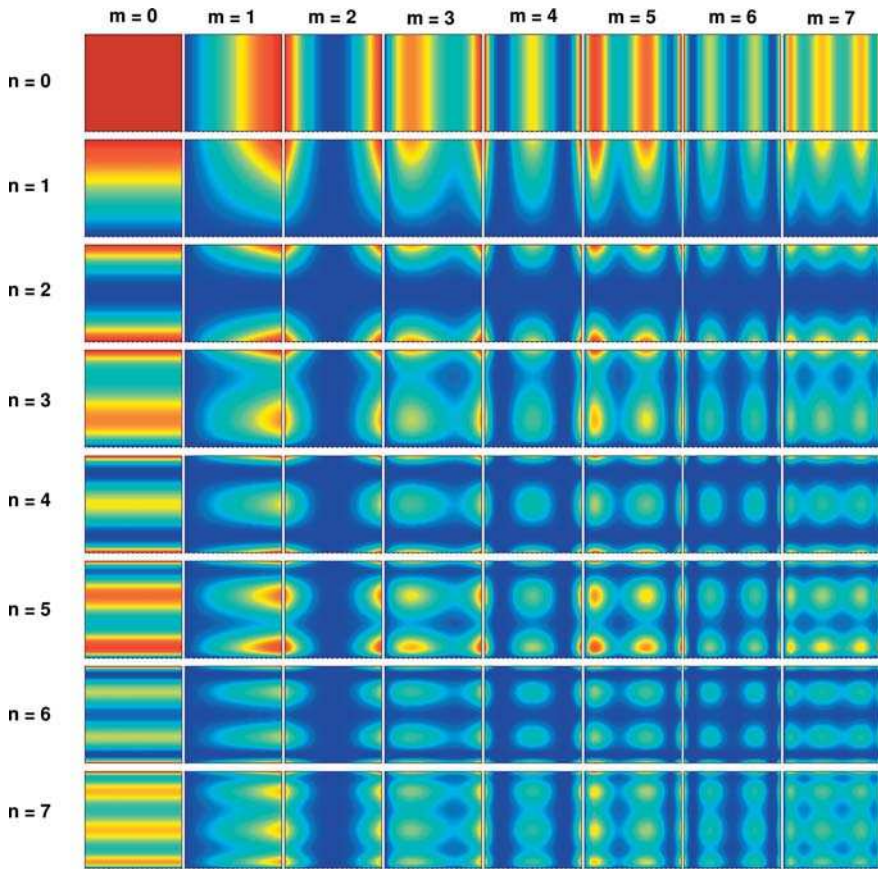


Figure A-44: Representation of the products of one-dimensional Legendre polynomials on a square-shaped area.

Another problem associated with such an attempt at two-dimensional generalization is the behavior of the polynomials near the boundary of the definition area. As can be seen in figure A-42, the functions exhibit their largest values at the ends of the interval. In the product representation for the two-dimensional case, this effect is amplified. Figure A-45 shows the extreme peak amplitudes of the two-dimensional functions in the corners of the square-shaped area which can be seen more clearly than in figure A-44. Such a surface shape is not typical for wavefronts in standard optical systems and therefore this type of surface description is not really adequate for this type of problem.

The best approach for the description of the behavior of two-dimensional wavefronts was developed only very recently [A-34]. It is based on the use of Zernike polynomials. If the Gram-Schmidt orthogonalization procedure is applied to the classical Zernike polynomials, for a rectangular shape of the definition area, one obtains the generalized two-dimensional Legendre polynomials. The lowest orders of these functions resemble the Zernike polynomials and therefore can be used for the usual

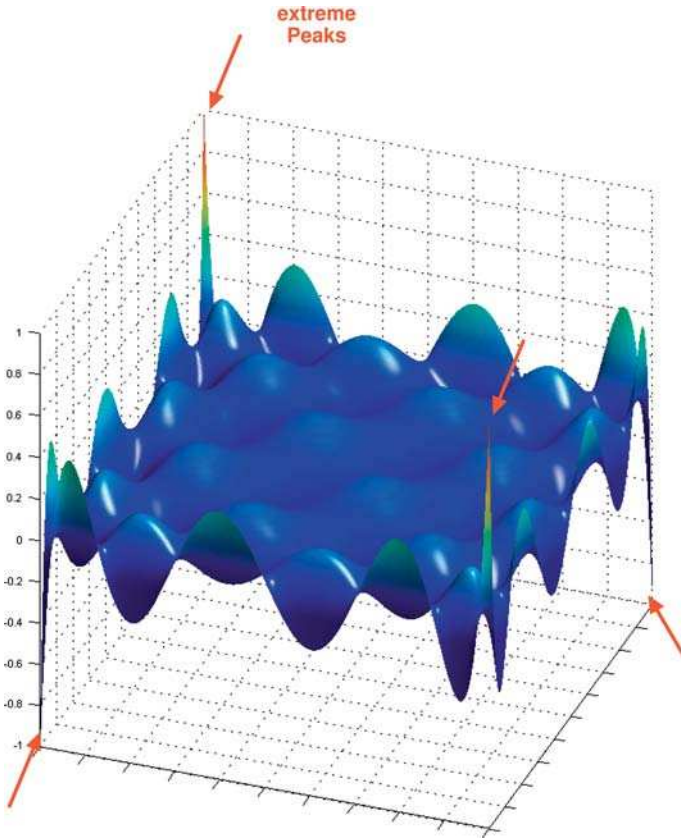


Figure A-45: Peak values of the product of one-dimensional Legendre polynomials of 7th order (P_7) defined on a square-shaped area.

interpretation of various aberration types like spherical aberration, astigmatism, coma, etc. The orthogonality ensures favorable mathematical properties. Figure A-46 shows the first 36 generalized two-dimensional Legendre polynomials.

If the new functions are expanded in Zernike polynomials in the form

$$Q_m = \sum_{j=1}^m c_j \cdot Z_j \tag{A-161}$$

with the orthogonality condition

$$\int_{-1}^1 \int_{-1}^1 Q_k Q_j dx dy = \delta_{kj} \tag{A-162}$$

then the orthogonalization algorithm yields a transform matrix for the determination of the new coefficients c_j . The lowest coefficients c_j are listed in table A-1. It should be noted that only the lower triangular region of the transfer matrix is filled.

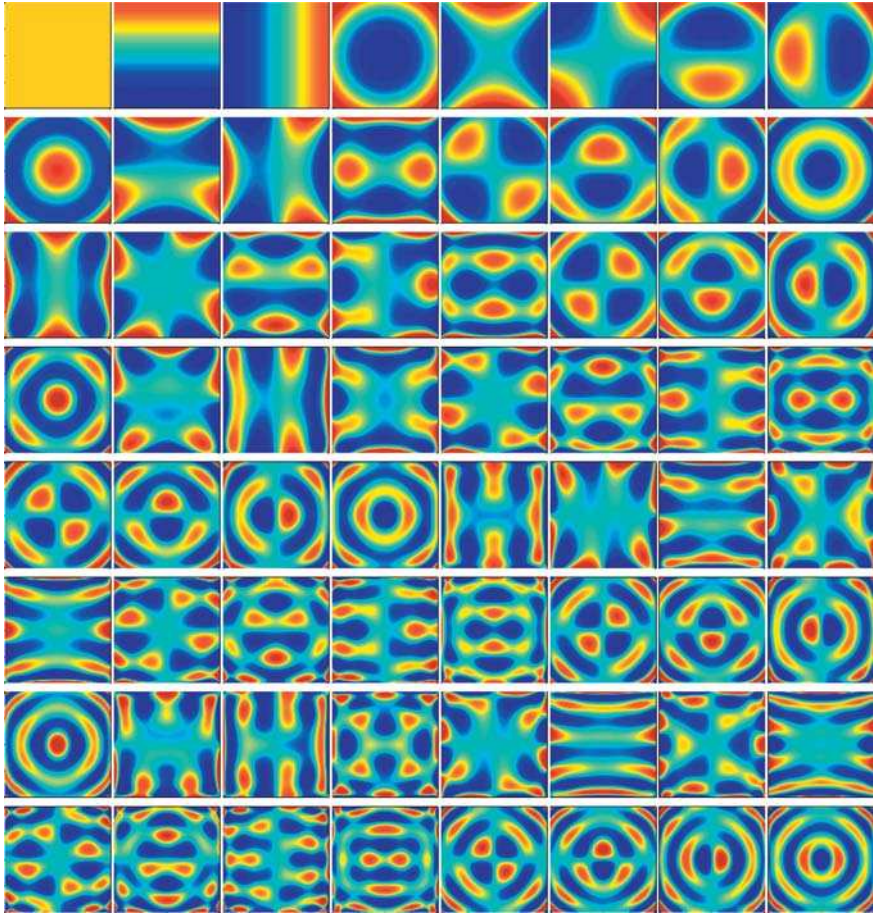


Figure A-46: Representation of the generalized two-dimensional Legendre polynomials. The indices run continuously from left to right and from top to bottom.

Table A-1: Coefficients of the generalized Legendre polynomials in two dimensions using expansion in Zernike polynomials.

Order m	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
1	0.5								
2		0.8660							
3			0.8660						
4	-0.1976			0.5929					
5					1.1859				
6						0.7500			
7		-0.5820					0.7275		
8			-0.5820					0.7275	
9	-0.0840			-0.6299					0.4009
10		0.0710					0.4615		
11			-0.0710					-0.4615	
12	0.1826			0.3025	-0.4571				0.2600
13						-0.9114			
14		-0.0370					-1.0053		
15			-0.0370					-1.0053	
16	0.0826			0.1362	-0.0338				-0.8132
17	-0.0513			-0.0068	0.1925				0.2248

A.9 Literature

- A-1** D. C. Champeney, *Fourier Transforms and their Physical Applications* (Academic Press, London, New York, 1973).
- A-2** W. T. Cathey, *Optical Information Processing and Holography* (John Wiley and Sons, New York, 1974).
- A-3** R. N. Bracewell, *Two-dimensional Imaging* (Prentice Hall, Englewood Cliffs, 1995).
- A-4** E. Oran Brigham, *The fast Fourier Transform and its Applications*, (Prentice Hall, Englewood Cliffs, 1988).
- A-5** K. R. Castleman, *Digital Image Processing* (Prentice Hall, Upper Saddle River, 1996).
- A-6** L. R. Rabiner, R. W. Schafer and C. M. Rader, *IEEE Trans. AU* **17**, 86 (1969), The chirp z-transform algorithm.
- A-7** DSP Committee (Ed.), *Programs for Digital Signal Processing* (IEEE Press, New York, 1979).
- A-8** S. Roose, B. Brichau and E. W. Stijns, *Opt. Comm.* **97**, 312 (1993), An efficient interpolation algorithm for Fourier and diffractive optics.
- A-9** J. L. Bakx, *Appl. Opt.* **41**, 4897 (2002), Efficient computation of optical disk readout by use of the chirp z transform.
- A-10** A. E. Siegman, *Opt. Lett.* **1**, 13 (1977), Quasi fast Hankel transform.
- A-11** S. M. Candel, *IEEE Trans. ASSP* **29**, 963 (1981), Dual algorithms for fast calculation of the Fourier-Bessel transform.
- A-12** A. V. Oppenheim, G. V. Frisk and D. R. Martinez, *J. Acoust. Soc. Am.* **68**, 523 (1980), Computation of the Hankel transform using projections.
- A-13** K.-H. Brenner and W. Singer, *Appl. Opt.* **32**, 4984 (1993), Light propagation through microlenses: a new simulation method.
- A-14** E. Sziklas and A. E. Siegman, *Appl. Opt.* **14**, 1874 (1975), Mode calculations in unstable resonators with flowing saturable gain 2: Fast Fourier transform method.
- A-15** D. Mas, J. Perez, C. Hernandez, C. Vazquez, J. J. Miret and C. Illueca, *Opt. Commun.* **227**, 245 (2003), Fast numerical calculation of Fresnel patterns in convergent systems.

- A-16** D. Mas, J. Garcia, C. Ferreira, L. M. Bernardo and F. Marinho, *Opt. Commun.* **164**, 233 (1999), Fast algorithms for free-space diffraction patterns calculation.
- A-17** J.-L. Kaiser, E. Quertemont and R. Chevallier, *Opt. Commun.* **233**, 261 (2004), Light propagation in the pseudo-paraxial Fresnel approximation.
- A-18** F. Medina, J. Garcia-Sucerquia, R. Castaneda and G. Matteucci, *Optik* **111**, 547 (2004), Angular criterion to distinguish between Fraunhofer and Fresnel diffraction.
- A-19** G. N. Lawrence and S.H. Hwang, *Appl. Opt.* **31**, 5201 (1992), Beam propagation in gradient refractive-index media.
- A-20** R. R. Shanon and J. C. Wyant, *Applied Optics and Optical Engineering*, Vol XI (Academic Press, San Diego, 1992), Ch. 3: G. N. Lawrence, *Optical Modeling*.
- A-21** A. J. Lambert and D. Fraser, *Appl. Opt.* **37**, 7933 (1998), Linear systems approach to simulation of optical diffraction.
- A-22** P. Luchini, *Comput. Phys. Commun.* **31**, 303 (1984), Two-dimensional Numerical Integration using a square mesh.
- A-23** A. VanderLugt, *Appl. Opt.* **29**, 3352 (1990), Optimum sampling of Fresnel transforms.
- A-24** H. G. Kraus, *Opt. Eng.* **30**, 1191 (1991), Finite element Fourier and Abbe transform methods for generalization of aperture function and geometry in Fraunhofer diffraction theory.
- A-25** H. G. Kraus, *Opt. Eng.* **32**, 368 (1993), Finite element area and line integral transforms for generalization of aperture function and geometry in Kirchhoff scalar diffraction theory.
- A-26** A. C. Ludwig, *IEEE Trans AP-16*, 767 (1968), Computation of radiation patterns involving numerical double data integration.
- A-27** T. Gravelaeter and J. J. Stamnes, *Appl. Opt.* **21**, 3644 (1982), Diffraction by circular apertures. 1.: Method of linear phase and amplitude approximation.
- A-28** J. J. Stamnes, B. Spjelkavik and H. M. Pedersen, *Optica Acta* **30**, 207 (1983), Evaluation of diffraction integrals using local phase and amplitude approximations.
- A-29** J. J. Stamnes, *Waves in Focal Regions* (A.Hilger, Bristol, 1986).
- A-30** Y. M. Engelberg and S. Ruschkin, *J. Opt. Soc. Am. A* **21**, 2135 (2004), Fast method for physical optics propagation of high-numerical-aperture beams.
- A-31** M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions* (Dover Publications, New York, 1972).
- A-32** P. R. Riera, G. S. Pankretz, D. M. Topa, *Proc. SPIE* **4769**, 130 (2002), Efficient computation with special functions like the circle polynomials of Zernike.
- A-33** J. L. Rayces, *Appl. Opt.* **31**, 2223 (1992), Least-squares fitting of orthogonal polynomials to the wave-aberration function.
- A-34** M. Bray, *Proc. SPIE* **5252**, 314 (2004), Orthogonal polynomials: A set for square areas.

Index

a

Abbe resolution 247
 Abbe sine condition 268
 Abbe theory of image formation 241 ff
 aberrations 212 ff, 309 ff, 325, 351, 384, 621
 absorbing boundary 570
 Airy disc 206, 276, 325
 aliasing 640
 ambiguity function 123, 429
 amplitude edge 366
 angular spectrum 17 ff, 54 ff
 anisotropic media 494 ff, 511, 598
 annular pupil 224 ff, 410 ff
 aplanatic correction 261
 apodization 416 ff
 approximation
 – Born approximation 81, 327
 – finite differences 569
 – Fraunhofer 22, 69
 – Fresnel approximation 21
 – single path approximation 600 ff
 array homogenizer 172 ff
 astigmatism 217
 autocorrelation function 163, 638
 axial coherence length 130 ff
 axial resolution 425 ff

b

Babinet principle 82 ff
 Babinet-Soleil compensator 516
 bandwidth-limited signals 128
 beam splitting in anisotropic media 511
 beam waist 32, 150
 beam twist 157
 biaxial media 499, 510
 birefringent media 494, 497, 599
 Bluestein algorithm 646
 Born approximation 81, 327

boundary condition 23
 – absorbing boundary 570
 – periodic boundary 662
 – Sommerfeld radiation condition 13, 47 ff
 – transparent boundary 571
 boundary diffraction wave 89 ff
 Brewster plate 519

c

calculation of diffraction integrals 61, 655 ff
 canonical pupil 196
 canonical variables 190
 central obscuration 411
 characteristic function 191 ff
 Chebychev polynomials 675
 chirped ring pupil mask 437
 chirp z-transform 644 ff
 circular polarization 470
 circular polarized pupil 535
 circular retarder 486
 coherence function 113 ff, 296 ff
 – circular source 140
 – of a Gaussian Schell beam 154
 – mutual coherence function 114, 265
 coherence length
 – axial coherence length 130 ff
 – transverse coherence length 150, 179 ff
 coherence time 130
 coherence transfer function 296 ff
 coherent image formation 269 ff
 coherent resolution 252
 coherent transfer function 264
 Collins diffraction integral 64 ff
 coma 220
 conjugate planes 65
 conservation of energy 192, 204
 contrast 112, 254, 594
 convolution 637

correlation 637
 coupled dipole method 553
 critical dimension 249
 cross spectral density 116, 128
 cubic phase plate 442 ff

d

Dändlikers representation 330
 dark field illumination 451
 defect filtering 455
 degree of coherence 115, 176, 182
 degree of polarization 473 ff
 depth of focus 335, 430
 depth resolution 335 ff
 diattenuation 616
 dielectric tensor 495
 diffraction at a circular cylinder 542 ff
 diffraction at a rectangular aperture 59
 diffraction integral
 – calculation 61, 655 ff
 – Collins formulation 64 ff
 – Fraunhofer approximation 22, 69
 – Fresnel approximation 21
 – Kirchhoff integral 44 ff
 – Luneburg integral 200
 – Rayleigh-Sommerfeld 48
 – Richards-Wolf integral 526 ff
 – Stratton-Chu integral 538
 – vector Kirchhoff integral 538
 diffractive optical elements 199
 dipole illumination 257
 dipole wave 11
 discrete fast Fourier transform 642
 dispersion relation 6
 distortion 219
 double refraction 514
 double slit 103, 143, 360 ff
 dual algorithm 650

e

edge image 365 ff
 eigenpolarization 483, 502, 506 ff
 eigenwaves 502
 eikonal equation 19, 28, 190
 ellipse of polarization 470
 energy conservation 192, 204
 equivalent diffraction setup 663 ff
 evanescent waves 18, 26
 Ewald sphere 18, 321, 328, 663
 extended depth of focus 426, 442
 extended Zernike polynomials 227
 extraordinary index of refraction 504

f

Faraday rotator 518
 fidelity 272, 287
 finite difference method 568 ff
 Fourier integral 631
 Fourier method for diffraction computation 669
 Fourier modal method 563 ff
 Fourier theory of image formation 263 ff
 Fourier transform 632 ff
 – discrete fast 632, 642
 Fraunhofer diffraction 22, 69
 frequency doubling 453
 Fresnel approximation 21
 Fresnel diffraction 60 ff, 666
 Fresnel ellipsoid 496
 Fresnel equations 540
 Fresnel integral 22, 62, 91, 664, 672
 Fresnel number 69

g

Gardner transform 649
 Gaussian beams 31 ff
 Gaussian Schell beams 149 ff
 geometrical theory of diffraction 90 ff
 Gibbs phenomenon 72
 grating
 – diffraction 71 ff
 – image formation 293, 381 ff
 – sinusoidal phase grating 76
 Greens function 44, 85, 139

h

Hamiltonian 190
 Hankel function 51
 Hankel transform 648
 Helmholtz equation 6
 Helmholtz-Lagrange invariant 233
 Herschel condition 341
 Hilbert space angle 427
 homocentric pupil 198
 Hopkins formalism for periodic objects 288 ff
 Hopkins integral 139
 Hopkins transmission cross coefficient 284 ff, 370, 384
 Huygens principle 24, 46, 52

i

illumination
 – dark field illumination 451
 – dipole illumination 257
 – oblique illumination 247, 364

- partially coherent illumination 277, 286, 370, 405
- polarized illumination 613
- Siedentopf illumination principle 255 ff
- sigma setting 279, 608
- structured illumination 256, 447
- image formation
 - coherent image formation 269 ff
 - image formation of a line 376 ff
 - image of a circular object 400 ff
 - image of an edge 365 ff
 - image of a grating 381 ff
 - incoherent image formation 272 ff
 - Koehler illumination 252 ff
 - polarized image formation 583, 606 ff
 - theory of Abbe 241 ff
- image multiplexing 430
- implicit difference method 572
- incoherent image formation 272 ff
- incoherent optical transfer function 273
- index ellipsoid 497
- index of refraction 8
 - extraordinary index of refraction 504
 - ordinary index of refraction 504
- interference contrast 113
- intrinsic birefringence 500
- irradiance 14 ff
- isolator, optical 518
- isotropic media 497

j

- Jones matrix 479 ff, 489, 507, 613, 618 ff
- Jones vector 468 ff
- Jones vector propagation 480
- Jones-Zernike expansion 622 ff

k

- Kirchhoff diffraction integral 44 ff
- Koehler illumination 252 ff

l

- Laue construction 87
- Laue equation 328
- leakage 641
- Legendre polynomials 677
- light source
 - circular 140
 - coherence domains 137
 - correlated 280
 - point source 110
 - spectral distribution 127
 - thermal source 133
 - uncorrelated 262

- linear systems 309, 629 ff
- line image 376
- line of sight 221 ff
- line spread function 375
- linear polarized pupil 533
- line image 376 ff
- logarithmic phase mask 435
- Luneburg integral 200

m

- magnification 340
- Marechal criterion 212
- Maxwell's equation 2, 502
- McCutchen equation 323
- Michelson interferometer 134
- Mie scattering 547 ff
- mixed characteristic 192
- mode expansion method 156
- moment method 555
- moment theorem of the Fourier theory 633
- moments of the Wigner distribution function 120
- Müller-Jones matrix 493
- Müller matrix 491
- multiplexing of an image 430
- mutual coherence function 114, 265

o

- oblique illumination 247, 364
- optical activity 499
- optical isolator 518
- optical transfer function
 - coherent 264
 - incoherent 273
 - polarized 613
 - three dimensional 332, 343
- ordinary index of refraction 504
- orthogonal polynomials 675 ff

p

- paraxial wave equation 29, 572
- partially coherent imaging 277, 286, 370, 405
- partial polarization 473, 610
- Pauli spin matrices 489
- periodic boundary conditions 662
- phase contrast filtering 450
- phase mask, logarithmic 435
- phase space 231, 300 ff
- phase space analyser 68
- pinhole image 399 ff
- plane wave 9, 512, 531
- Poincare sphere 478
- point characteristic 189

- point spread function 217, 275, 322 ff, 416 ff, 423
 - Airy disc 206, 276, 325
 - apodized pupil 416
 - ideal point spread function 426
 - polychromatic point-spread function 208
 - quasi point source 402 ff
 - three dimensional 423
 - vector point-spread function 611
- polarization
 - beam splitting in anisotropic media 511
 - biaxial media 499, 510
 - birefringent media 494, 497, 599
 - circular polarization 470
 - circular polarized pupil 535
 - dielectric tensor 495
 - double refraction 514
 - eigenpolarization 483, 502, 506 ff
 - ellipse 470
 - extraordinary index of refraction 504
 - Faraday rotator 518
 - illumination 613
 - intrinsic birefringence 500
 - isotropic media 497
 - Jones matrix 479 ff, 489, 507, 613, 618 ff
 - Jones vector 468 ff
 - Jones-Zernike expansion 622 ff
 - linear polarized pupil 533
 - Müller-Jones matrix 493
 - Müller matrix 491
 - optical activity 499
 - optical isolator 518
 - ordinary index of refraction 504
 - partial polarization 473, 610
 - Poincare sphere 478
 - polarization coherence function 183
 - polarization imaging 583, 606 ff
 - polarization matrix 473
 - polarization retarder 482, 484 ff, 492, 516
 - polarization rotator 482, 518
 - polarization state 467
 - polarization ray trace 597 ff
 - polarization transfer function 613
 - quarter wave retarder 516
 - radially polarized pupil 536
 - ray splitting 598
 - retarder 482, 484 ff, 492, 516
 - Stokes vector 475 ff
 - tangentially polarized pupil 537
 - vector interference 593
 - vector Kirchhoff diffraction integral 538
 - vector point-spread function 611
- polychromatic point-spread function 208
- polynomials
 - Chebychev polynomials 675
 - extended Zernike polynomials 227
 - Jones-Zernike expansion 622 ff
 - Legendre polynomials 677
 - orthogonal polynomials 675 ff
 - Tatian polynomials 224
 - Zernike polynomials 212 ff, 439, 622 ff
- power spectral density 128, 638
- Poynting vector 7, 14
- principal refractive indices 495
- projection algorithm 651
- propagation
 - coherence function 139
 - coupled dipole method 553
 - finite difference method 568 ff
 - Fourier modal method 563 ff
 - free space 20, 303
 - Jones vector propagation 480
 - mode expansion method 156
 - rigorous coupled wave analysis 563 ff
 - split-step beam propagation method 576
 - transport of intensity 30
 - Wigner distribution function 146 ff, 303 ff
- pupil
 - pupil function 202, 613
 - pupil mask filter 411, 423 ff, 433 ff
 - annular pupil 224 ff, 410 ff
 - apodization 416 ff
 - canonical pupil 196
 - chirped ring pupil mask 437
 - circular polarized pupil 535
 - filling of an array homogenizer 175
 - homocentric pupil 198
 - linear polarized pupil 533
 - radially polarized pupil 536
 - tangentially polarized pupil 537
 - telecentric pupil 198
- q**
 - quarter wave retarder 516
 - quasi point source 402 ff
- r**
 - radially polarized pupil 536
 - radiant flux 204
 - ray equivalent of Gaussian beams 36
 - Rayleigh length 31, 34
 - Rayleigh-Sommerfeld diffraction integral 48, 53
 - Rayleigh unit 323
 - ray splitting 598
 - refractive index 4, 8

- extraordinary index of refraction 504
- ordinary index of refraction 504
- principal refractive indices 495
- resolution 210, 234, 252, 276
 - Abbe resolution 247
 - axial resolution 425 ff
 - critical dimension 249
 - depth resolution 335 ff
 - gain 425
 - spatial resolution 660
 - spectral resolution 660
 - transverse resolution 425 ff
- retarder 482, 484 ff, 492, 516
- Richards-Wolf diffraction integral 526 ff
- rigorous coupled wave analysis 563 ff
- Ronchi grating 71
- Ronchi test 456

S

- sampling theorem 639
- scattering 556 ff
 - Mie scattering 547 ff
 - scalar scattering 85 ff
 - scattering plate 170
 - stray light 314 ff
 - surface roughness 163
- Schwarzschild angular eikonal 193
- secondary source 176
- Siedentopf illumination principle 255 ff
- sigma setting 279, 608
- sinc interpolation 648
- single-path approximation 600 ff
- sinusoidal phase grating 76
- slit
 - double slit 103, 143, 360 ff
 - slit diffraction 57
 - slit imaging 377 ff
- slowly varying envelope 30
- Sommerfeld condition 13
- space-bandwidth product 233 ff
- spatial coherence 135 ff
- spatial frequency 17, 245
- spatial resolution 660
- speckle 159 ff
- speckle reduction 169 ff
- spectral resolution 660
- spherical aberration 217
- spherical wave 11
- split-step beam propagation method 576
- Stokes vector 475 ff
- Stratton-Chu diffraction integral 538
- stray light 314 ff
- Strehl ratio 211
- structured illumination 256, 447
- super Gaussian profile 417
- surface roughness 163

T

- Talbot effect 385, 394
- tangentially polarized pupil 537
- Tatian polynomials 224
- tayloring 346
- telecentric imaging 197
- telecentric pupil 198
- temporal coherence 126 ff
- thermal light source 133
- three dimensional transfer function 332, 343
- Toraldo pupil filter 434
- transfer function
 - coherence transfer function 296 ff
 - coherent transfer function 264
 - defocused transfer function 429
 - incoherent optical transfer function 273
 - polarized transfer function 613
 - three dimensional transfer function 332, 343
- transform
 - chirp z-transform 644 ff
 - discrete fast Fourier transform 642
 - Fourier transform 632 ff
 - Gardner transform 649
 - Hankel transform 648
 - z-transform 644 ff
- transmission cross coefficient 284 ff, 370, 384
- transparent boundary 571
- transport of intensity equation 30
- transverse coherence length 150, 179 ff
- transverse resolution 425 ff
- truncated Gaussian profile 418
- twisted Gaussian Schell beam 157
- two-point resolution 356 ff

U

- uncertainty relation 122, 234
- uniaxial media 498, 509

V

- van Cittert-Zernike theorem 140
- vector interference 593
- vector Kirchhoff diffraction integral 538
- vector point-spread function 611
- visibility 112, 136, 252

w

- waist of a beam 32, 150
- wave equation 4
 - Helmholtz wave equation 6
 - paraxial wave equation 29, 572
- Wiener experiment 592
- Wiener-Khinchin theorem 129
- Wigner distribution function 116 ff, 300 ff
 - coherent fields 122
 - Gaussian Schell beams 154

- grating 394 ff
- moments 120
- propagation 146 ff, 303 ff

y

- Youngs experiment 143 ff

z

- Zernike polynomials 212 ff, 439, 622 ff
- z-transform 644 ff